Trade-off Management in Early Mechanical Design

Identifying, Understanding, and Mitigating the Causes of Trade-offs Between Design Objectives

Sigurdarson, Nökkvi Steinn Reinholdt

Publication date:
2021

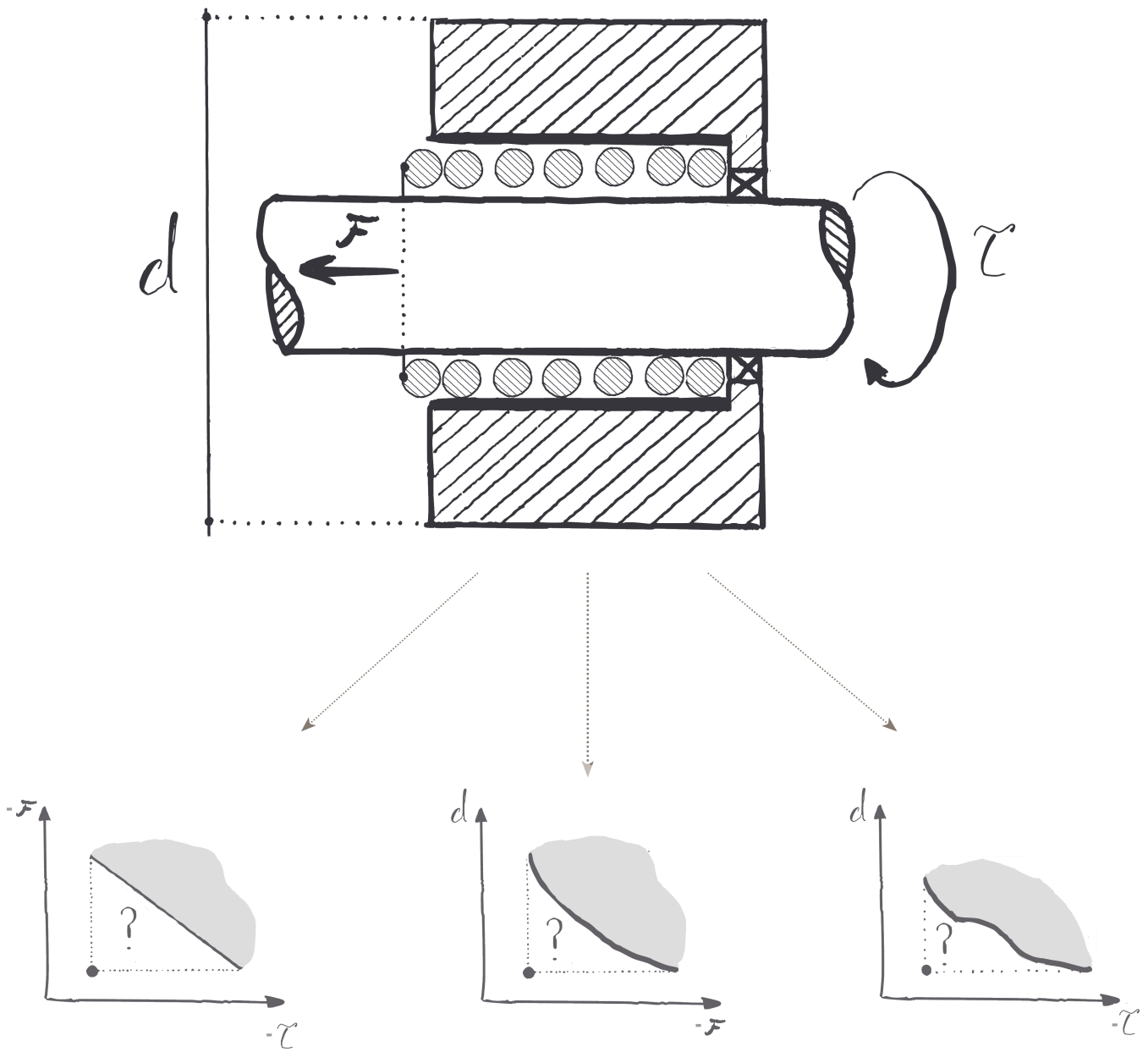
Document Version
Publisher's PDF, also known as Version of record

DTU Mechanical Engineering
Section of Engineering Design and Product Development

Trade-off Management in Early Mechanical Design

Identifying, Understanding, and Mitigating the Causes of Trade-offs Between Design Objectives

PhD Thesis



Trade-off Management in Early Mechanical Design
Identifying, Understanding, and Mitigating the Causes of Trade-offs Between Design Objectives

PhD Thesis
October 30th, 2021

By
Nökkvi Steinn Reinholdt Sigurdarson

Cover photo: Sketch, Nökkvi S. R. Sigurdarson, 2021
Published by: DTU, Section of Engineering Design and Product Development, Nils Koppels Allé, Building 404, 2800 Kgs. Lyngby Denmark
www.mek.dtu.dk

Abstract

Trade-off situations are inescapable in product development. No product can be infinitely durable, sustainable, inexpensive, efficient, user friendly, and so on. Many of these trade-offs occur due to challenges within the domain of mechanical engineering. Make a structure lighter, and it usually loses stiffness. Make a mechanism more accurate, and it usually becomes less mechanically efficient. These situations often arise due to inherent dependencies between the goals a product is designed towards. Yet, they can also emerge due to the constraints that follow the manufacturing processes and materials used to manufacture these products.

This challenge will only grow over time. Competitive pressures drive companies to strive to improve product performance and integrate more features, all the while keeping costs low. In turn, this drives designers to attempt to realise as much functionality with as few components as possible, with more trade-offs arising as a result. This tendency will only increase in the future with societal needs and technological developments introducing more and more objectives that the design engineer needs to take into account. An example of this is the drastically increasing need to develop more sustainable products.

While compromise may at times be inevitable, the lack of up-front awareness, understanding, and mitigation of trade-offs during the initial stages of mechanical design can have substantial consequences for the performance of the end product. This can also delay product development projects and result in unforeseen quality and cost issues in production. In other words, trade-offs can delay technological progress in general.

This PhD thesis describes the development of methods for the management of trade-off situations during the early phases of mechanical design. This includes *Pareto set Dependency Analysis*, a quantitatively founded approach that that builds upon existing monotonicity analysis and design optimisation methods to identify trade-offs and their underlying root causes. This opportunistic yet rigorous approach led to the development redesign and synthesis methods that allow the identification of design changes that result in an improved Pareto set. This implies that the trade-offs have been mitigated or reduced and that an overall improvement in product performance has been achieved. This provides a systematic foundation allowing designers to continually identify design changes that result in improved performance, as the design of the product is gradually refined, even if the need for additional features, functionalities, and requirements arises.

This research was conducted in an industrial-academic collaboration between DTU Mechanical Engineering and Novo Nordisk. Cases from ongoing product development projects were used in the research, one of which is included in this thesis. The SOMA device, an ingestible medical device for the oral delivery of pharmaceutical compounds such as insulin, is used to demonstrate the application of these analysis and redesign methods. Using the novel analysis methods and multiobjective optimisation, several drivers of trade-offs in the SOMA device were identified. Many of these were successfully mitigated using the redesign methodology, resulting in a set of redesigns that exhibit improved Pareto sets, confirming the practical value of the results of the research.

Danish Summary

Kompromissituationer er uundgåelige i produktudvikling. Intet produkt kan være uendeligt holdbart, bæredygtigt, billigt, energieffektivt, brugervenligt, og så videre. Mange af disse kompromissituationer opstår grundet udfordringer der hører til maskinteknikken. Gør man en bærende konstruktion lettere, bliver den typisk mindre stiv. Gør man en mekanisme mere nøjagtig, bliver den typisk mindre energieffektiv. Disse situationer opstår typisk grundet iboende afhængigheder mellem de mål produktet bliver designet hen imod. De kan dog også opstå på grund af de begrænsninger der følger valget af produktionsprocesser og materialer.

Denne udfordring bliver kun større med tiden. Moderne virksomheder konkurrerer ofte ved at lancere nye produkter med bedre ydelse, mere funktionalitet, og en lavere. Omvendt driver dette designingeniører til at forsøge at realisere mere funktionalitet med så få komponenter som muligt. Flere kompromissituationer opstår som resultat. Med til stadighed flere samfundsbehov - såsom behovet for mere bæredygtige produkter - presses designingeniører af flere og flere designmål.

Selv om kompromiser ofte er uundgåelige, kan en mangelfuld identifikation, forståelse, og nedsættelse af bidragsydere til kompromis i den tidlige designprocess have store konsekvenser for det endelige produkt. Dette kan også forsinke produktudviklingsprojekter, og resultere i uforudsete kvalitets og omkostningsproblemer i produktion. Med andre ord nedsætter kompromissituationer i mekanisk produktudvikling samfundets teknologiske fremgang.

Denne PhD afhandling beskriver udviklingen af metoder til at håndtere kompromissituationer aktivt i tidlig mekanisk design. Dette inkludere en analysemetode til at finde de afhængigheder der findes i Pareto sættet, som bygger videre på eksisterende optimeringsmetoder. Denne matematisk funderede tilgang tillod efterfølgende udviklingen af en redesignmetode der resulterer i elimineringen eller formindskelsen af de kompromiser der er involverede i dimensioneringen af et mekanisk system. Dette giver designingeniører et systematisk fundament til at forbedre et produkts ydelse kontinuerligt igennem hele designprocessen.

Dette forskningsprojekt blev gennemført i et industrielt-akademisk samarbejde mellem DTU Mekanik og Novo Nordisk. Casestudier fra igangværende udviklingsprojekter i Novo Nordisk blev gennemført som en del af forskningsprojektet, og et af disse er inkluderet i denne afhandling. Det såkaldte SOMA device; en oral injektionsmekanisme til leveringen af behandlingsmidler såsom insulin blev brugt til at demonstrere de nyudviklede metoder. Resultatet af dette var en succesfuld analyse og redesign process, der tillod udviklingen af adskillige redesigns af SOMA devicet. Mange af disse viste sig at have et forbedret Pareto sæt, hvormed den praktiske værdi af forskningsresultaterne er blevet bekræftet.

Preface

You can't always get what you want,
but if you try sometimes,
you'll find you get what you need.

Keith Richards & Mick Jagger, 1969

This thesis concludes a three and a half year industrial PhD project on the management of trade-offs in early mechanical design. Its focus arose out of the frustration I experienced as a young mechanical design engineer. As opposed to my more experienced colleagues, I was rarely able to avoid creating trade-off situations when designing new products. To put the above quote into context, I simply could not achieve the product performance I wanted. I, therefore, sought to understand *why* these trade-offs occur, *when* they can be avoided, and *how* this could be achieved so that I might actually achieve what I wanted in the synthesis of new mechanical systems.

The project was conducted in the Devices and Delivery Solutions organisation in Novo Nordisk A/S, and at the Department of Mechanical Engineering, at the Technical University of Denmark (DTU), in collaboration with the Optimal Design Laboratory at the University of Michigan. The project was financed by the Industrial Researcher program run by Innovation Fund Denmark (grant no. 7038-00221B) and the Novo Nordisk STAR program. I would like to thank all the organisations involved for making this project possible and for the support I have received throughout.

Based on my conversations with fellow PhD students and young engineers, I have come to believe that good mentors are rare. Yet, I have been so lucky to encounter three truly great ones who have all had an immeasurable influence on my research and the person I am today. Like most things of note in my career as an engineer, this PhD project would not have happened were it not for my industrial supervisor Martin Ebro, who gave me a chance to write my masters thesis in Novo Nordisk back in 2015. As I transitioned into positions as a mechanical designer and then PhD fellow, Martin always had my back and opened doors that would have remained locked. Thank you for your patience and understanding, for believing in my ideas, and your uncanny ability to help improve them and inspire new ones.

My deepest thanks to Tobias Eifler, the one academic supervisor who remained a part of the project throughout. Thank you for always challenging my assumptions, countless stimulating discussions, and your constant willingness to help, even when I decided to embark on the study of topics and fields that were remote to both our backgrounds. Lastly, the contents of this thesis would also not have been possible without the guidance of Panos Y. Papalambros. Thank you so much for your patience, impeccable insight, and welcoming manner during my research stay at the Optimal Design Laboratory. My visit was by far the most challenging yet stimulating experience of my PhD (probably of my entire education). I am so thankful for your willingness to continue our collaboration beyond my research stay and the unofficial supervisor-like role you have taken.

Furthermore, I have been supervised and mentored many more, in both an official and unofficial capacity. Many thanks to Niels-Henrik Mortensen, Thomas J. Howard, Chris Mcmahon of DTU, who have all influenced my work in their own ways. I would also like to thank past and present members of the Robust Design group, especially Tim Brix Nerenst and Herle Kjemstrup Juul-Nyholm, for tagging along and thereby sparing me from the loneliness that would otherwise have followed in continuing as the only PhD student in the group.

I have also had the privilege to conduct research in an industrial setting, and in that regard,

I owe thanks to countless colleagues from Novo Nordisk - especially my colleagues in the Modelling and Simulation team, and to Klaus Bendix for inspiring much of the focus of this research. I would also like to thank Peter Herskind, Anders Maarstrand, Morten Revsgaard Frederiksen, Jeppe Vejgaard-Nielsen and the rest of the SOMA team for sharing data and helping to make the publication of my case studies possible! I know I have asked a lot of you, but your willingness to help and openness to sharing data has been pivotal to my research.

Finally, I owe my greatest thanks to my family and friends, especially my wife Ditte and our daughter Saga. Thank you both for being there, for all the ups and downs that a PhD entails. My wife and I started our PhD's at the same time, and I cannot express my gratitude for the support and understanding one could only get from a partner that is going through the exact same experience. To me, the journey we have been on together since the beginning of this project is a far greater achievement and infinitely more valuable than the contents of this thesis ever could be. Thank you!

Nomenclature

$\mathcal{A}$	Attainable set
$\mathcal{C}$	Pareto Set
$\mathbf{c}(\mathbf{x};\epsilon)$	Vector of bound objectives in the upper bound problem
D_s	Indices of the constraint functions that depend on a shared variable x_i
f	Primary objective function in the upper bound problem
f	The vector of design objectives in negative-null form
$f(x^+)$	A function increasing monotonically w.r.t. *x*
$f(x^-)$	A function decreasing monotonically w.r.t. *x*
F*	A [*k,j*]-matrix of Pareto optimal results
F^0	The utopia point
E	A [*k-1,j*] dimensional matrix of sampled values of ϵ
$\mathbf{g}(\mathbf{x})$	Vector of inequality constraints of the design problem
G*	Matrix of $\mathbf{g}(\mathbf{x}^*)$ values stored for every run
$\mathbf{h}(\mathbf{x})$	Vector of equality constraints of the design problem
H*	Matrix of $\mathbf{h}(\mathbf{x}^*)$ values stored for every run
j	Number of computational iterations ϵ is sampled over
k	Number of objectives
$\mathcal{L}$	The Lagrangian function
n	Number of design variables
P	Vector of design parameters
p	Number of redesign iterations
U	The symbolic cost function used to study the boundaries of the Pareto set after the elimination of the bound objectives. The function is of the form $U(f_1, \tilde{\epsilon_1}, ..., \tilde{\epsilon}_{k-1})$
$\mathcal{X}$	The set constraint or the feasible domain
$\mathcal{X}\epsilon$	Feasible domain for a given upper bound value, ϵ
x	Vector of design variables
$\underline{x}$	Argument of the infimum of the design problem
$\overline{x}$	Argument of the supremum of the design problem
$\underline{\overline{x}}$	Trade-off variable
$\overline{\overline{x}}$	A monotonically decreasing harmonious variable
$\underline{\underline{x}}$	A monotonically increasing harmonious variable
ϵ	A *k*-1 dimensional vector of upper-bound parameters
ϵ_i	Upper-bound parameter for the *ith* bound objective
$\epsilon_{\mathbf{L}}$	Lower limit of objective bounds
$\epsilon_{\mathbf{U}}$	Upper limit of objective bounds
$\tilde{\epsilon}_i$	Reduced-objective variable
$\tilde{\epsilon}^*_{i,j}$	Optimal value of $\tilde{\epsilon}_i$ implied by the activity case where the Pareto-constraint $g_j(\mathbf{x}, \tilde{\epsilon})$ bounds $\tilde{\epsilon}_i$
λ	Lagrange multiplier vector of inequality constraints

Contents

1 Introduction

The need to make trade-offs between different objectives is omnipresent in the design of mechanical systems. No product can be infinitely durable, inexpensive, efficient, sustainable, user friendly, and so forth. While compromise may at times be inevitable, the lack of up-front awareness, understanding, and management of trade-offs during the early stages of design can have dire consequences for the success of the end product and technological progress in general.

An example of such is that one could argue that the industrial revolution might have started sooner or occurred faster. James Watt's invention and development of the Watt steam engine, which occurred in the years between 1763-1775, is commonly cited as a key driver of the industrial revolution [1]. Nevertheless, Watt's invention was not the first steam engine. Back in 1698, Thomas Savery designed and built a steam-driven pump for coal mines. Given that it relied on an imperfect vacuum, Savery's invention had a low lifting height, which limited its areas of application [2]. Inspired by Savery's work, Thomas Newcomen later designed and built the *Atmospheric Engine* in 1712, more than 50 years before Watt's developments. Newcomen's engine solved some of the issues in Savery's design and was used widely in the coal mining industry to drive water pumps, allowing increasingly deep mines.

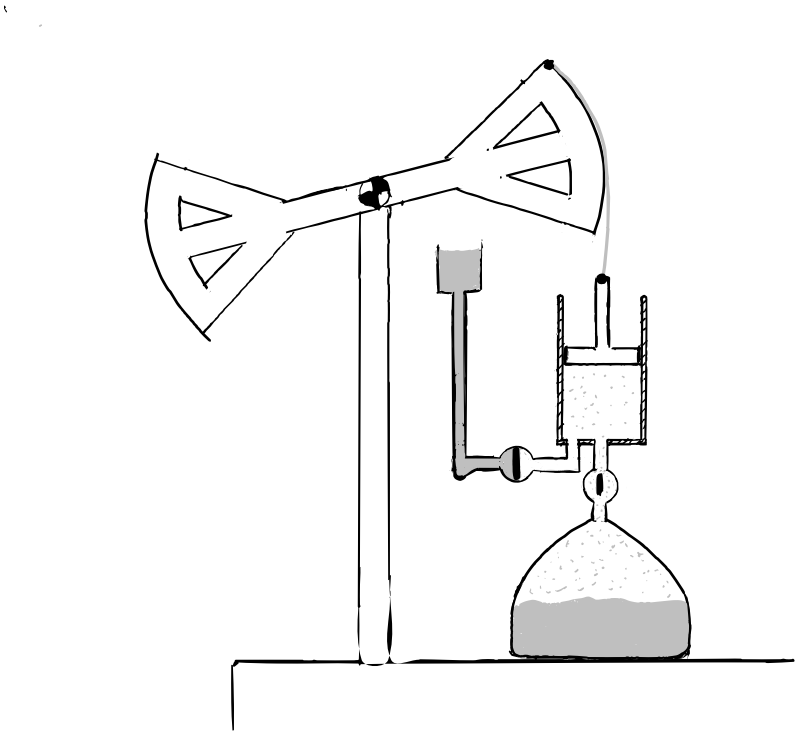


Figure 1.1: The Newcomen steam Engine (adapted from Suh (2001))

Yet, Newcomen's engine, and similar engines, never truly found broader application. With an efficiency of about 1% [2], its coal consumption made it economically unattractive for applications outside the coal sector, given the comparatively low cost of labour. The reason lies in Newcomen's design (see figure 1.4) [3], which has an inherent trade-off that limits its efficiency and energy density. Essentially, the expansion and condensation cycles of the

engine cannot be improved simultaneously, as both cycles occur inside the cylinder. Efficient and fast condensation relies, amongst other things, on a high heat transfer through the outer wall of the cylinder. Meanwhile, the efficiency of expansion relies on no heat transfer at all. As such, the cylinder would ideally be infinitely thermally conductive in one state and infinitely insulated in another. Furthermore, the cylinder would preferably always run warm to maintain a high efficiency, but the condensation cycle cools it down.

Given this dependency, any attempt to change the proportions of the design to improve condensation comes at the cost of a worsened expansion cycle. As a result, the equilibrium between the two determines the engine's overall efficiency. In James Watts's engine, the condensation cycle occurred in a separate chamber, meaning that the piston and the cylinder remained warm throughout the engine's operation. This change drastically increased efficiency, resulting in a 75% reduction in coal consumption. Combined with developments allowing continuous rotational motion, Watt's engine found much broader applications, for instance, replacing water wheels in factories and cotton mills.

We cannot know for sure that a more efficient design during on would have led to an earlier start of the revolution. There were multiple other factors at play, ranging from cultural to socioeconomic influences [1]. However, what is well known is that Watt's invention made the automation of previously manual processes economically feasible. All thanks to the mitigation of a trade-off through design change.

Engineering designers have since made considerable strides in the invention and development of new and improved engine principles. Correspondingly, physicists have to a large extent, developed a substantial understanding of the underlying physical phenomena. However, the more general question of how to identify, understand, mitigate, and manage trade-offs in design remains elusive in product development to this day.

1.1 Trade-offs in Engineering Design

Merriam-Webster defines *trade-offs* as "*A balancing of factors all of which are not attainable at the same time*". First studied in the context of economics - cf. the work by Vilfredo Pareto on *Pareto Efficiency* - trade-off situations are just as present in engineering.

This thesis involves research activities within the *engineering design* domain, focusing on the management of trade-offs in the early stages of product development. Engineering design research largely focuses on the study of the creation of products and systems and the behaviour of designers throughout this process [4]. It is thus a practice-oriented field, with research often relying on the study of industrial practice, with the oft sought but rarely captured aim of identifying a *science of design* [5]. The development of physical products and systems is mainly driven by engineering designers. From initial idea to running production, engineers will need to synthesise, understand, and improve designs, gradually refining them to a point where the end product is ready for launch [6].

An important notion in design science is the distinction between activities and decisions that occur in the earlier and later stages of design. All of the well-accepted design process theories, e.g. Pahl & Beitz [6], Ulrich & Eppinger [7], and Andreassen & Hein [8], state that the earlier stages of design involve the synthesis and refinement of the overall conceptual solution to a given functional intent. This is followed by the synthesis of the geometric realisation(s) of said concept into an embodied design (aka. system design [7], product structure [9], layout design [10], and configuration design [11]). Throughout these phases, the design evolves from a rough initial idea and sketch to a system where the desired functionality is achieved (at least to a certain degree of maturity).

How *good* an end product is overall is difficult to describe objectively. Hence, products are usually designed with a wide range of criteria in mind - be these explicitly stated by the designer or tacit in nature. These criteria describe how desirable or "good" the end product is and are key to a designer's decision-making. In the design optimisation field, these are referred to as *design objectives*. Importantly, it is widely accepted that the decisions made in the aforementioned initial stages of design are by far the most influential on the utility of the final product - i.e. how good an optimum is achievable (e.g. [6, 12–14]).

The importance of the conceptual and embodiment design phases lies in the fact that they determine which physical phenomena influence the behaviour of the end system, how the different system elements fit together and interact. Through decisions such as the selection of working principles [6] and design of the system layout/architecture [6, 10, 15], dependencies between design objectives are created. Also referred to as functional interrelationships [6] and couplings [13], dependencies give rise to trade-offs between the different objectives involved in the design of the product. An example of a dependency that causes trade-offs (sometimes referred to as contradictions [10, 16, 17]) might be a geometric dimension on a component needing to be large and small simultaneously in order to optimize two competing objectives. Given that this is impossible, the designer would need to make a trade-off decision unless the concept of embodiment design itself is changed.

Herein lies the core problem this PhD is aimed at dealing with. An organisation's ability to introduce changes to a design decreases as the design matures, with more time and cost being necessary to implement changes the later they are introduced [18]. At the same time, an organisation's knowledge about and understanding of a design problem increases over time, meaning it can be challenging to make the right decisions and synthesising the "right" designs in the conceptual and embodiment stages.

This is often referred to as the design process paradox [10] (visualised in Fig. 1.2), and it explains why most late design changes and increases in development lead-time can be traced back to decisions made during embodiment [19]. To a large extent, it is this paradox that makes the management of trade-offs challenging. Designers and organisations may simply lack the knowledge required to identify and potentially avoid the contributors to trade-offs in the phases of design where embodiment design changes can be made without substantial cost. Given their potential influence, the identification of dependencies is broadly seen as critical, yet very challenging to do during the conceptual and embodiment design phases [10, 12, 20, 21]. This is perhaps summarised best by Ullman [10, p. 285]:

> "*In early-stage design, the trade-off process is especially challenging, as there is limited knowledge, uncertainties are high, and the decisions made have far-reaching effects on the directions pursued thereafter, and hence the affordability, reliability/safety, and effectiveness of the final product. It is clearly more viable and less expensive to refine a design at the time that it is being conceived. Therefore, efforts toward making good decisions at this stage have high payoffs*."

Further, trade-off situations often involve numerous objectives and contributing factors making them difficult to comprehend [22], and the acceptance of trade-off analyses can also be limited by cognitive biases [23]. Different user/customer groups may also have inherently conflicting needs, further complicating the management of trade-offs.

Beyond this, ill-managed trade-offs can have a drastic influence on the end product's performance [13, 16], time to market [24], robustness [25], and part count and complexity [26]. For an example of the potential influence of trade-offs, consider an every-day trade-off issue in bag based vacuum cleaners [27]:

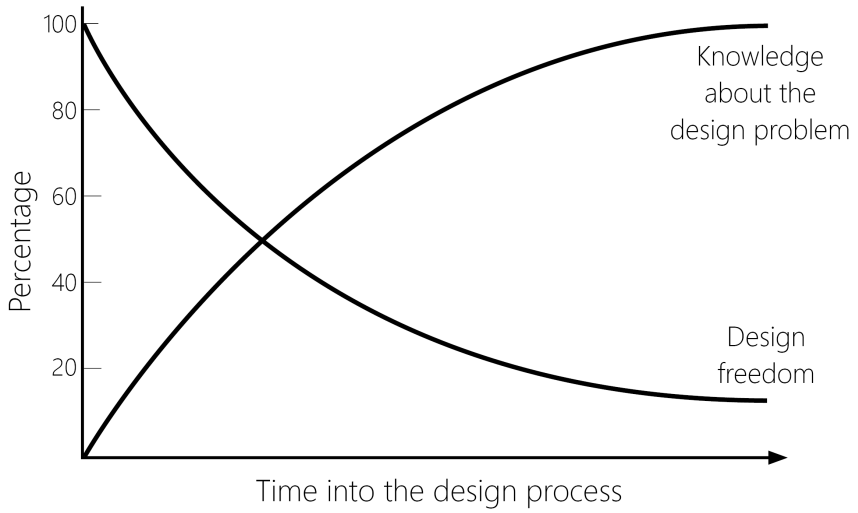


Figure 1.2: The so-called *design process paradox* [10] - figure adapted from Ullman [10]

Example - Vacuum Cleaners [from Paper C]
From the perspective of suction pressure and filtration quality, there is a substantial difference between vacuums with bags and bag-less vacuums, as illustrated in figure 1.3. Regular vacuums create an airflow using an electric motor and fan, which then sucks air through a hose, with dust and debris being filtered out by a bag and perhaps a secondary filter. The filtration in the bag is critical, as it prevents that debris causes damage to the motor while ensuring that dust is captured and not blown out again. Looking at two objectives, suction pressure and filtration quality, a trade-off reveals itself. The more efficient the bag is at filtration, the more resistance it creates, hindering air flow. Correspondingly, the higher the suction pressure, the tighter a filter is required to prevent dust and debris from exiting the bag. In other words, the better the filtration, the more powerful a motor is required to generate a given suction pressure at the end of the hose. This also means that the suction pressure is reduced the more the bag is filled. Bag-less vacuums, meanwhile, commonly rely on cyclonic separation - a process that incurs less loss to the suction path. In fact, the filtration quality increases with pressure, making the two objectives easier to improve simultaneously.

Ultimately, these situations are often inevitable in product development. Over time, competitive pressures drive companies to strive to improve product performance and add integrating more features with each new product generation, all the while keeping costs low [28]. In turn, this drives designers to attempt to realise as much functionality with as few components as possible, with more trade-offs arising as a result [29]. This tendency will only increase in the future, with societal and technological developments creating the demand for new disciplines and new requirements [30]. For instance, with the increasing digitisation of society and the need to transition into a more sustainable and circular economy, products are being designed towards an increasing amount of design objectives and constraints. In a report on the future of Product Development [30], the Design Society predicted that one of the key skills of the design engineers of the future is the ability to manage multi-disciplinary trade-offs systematically, despite increasing product complexity.

There does not seem to be a systematic, scientifically founded approach to managing trade-offs, used widely in engineering practice. Some engineers and designers are taught how to

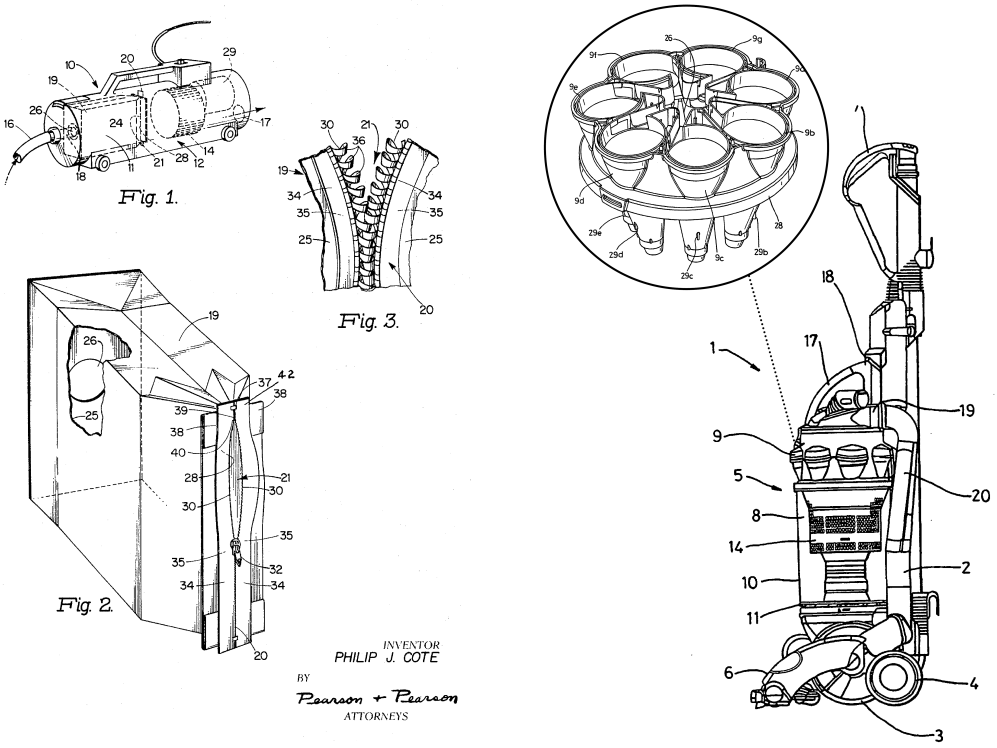


Figure 1.3: *Left:* Patent drawings of a classic bag-based vacuum cleaner (patent nr.US 3755993 A). *Right*: Patent drawings of an example of a bag-less vacuum, along cyclone based filtration system (patent nr. EP 1786568 B1)

identify, and model trade-offs between design objectives using optimisation techniques, and their utility in the design process beyond the identification of the optimum is often touted [12, 31, 32]. While optimization methods are commonly used at the embodiment stage to identify the optimal proportions of the design, systematic, quantitative analysis of trade-offs is less common ahead of important decisions such as concept selection, iterative redesign, or requirement setting [31]. Furthermore, while existing techniques are helpful in quantifying the trade-offs in a system, they do not necessarily concern themselves helping explain *why* the trade-offs exist in the first place. This aspect will be covered in depth in Chapters 3 and 4.

As a consequence of all of these factors, it is not uncommon for experienced designers to rely on intuition over systematic analysis [33, 34]. It is relatively well established that experienced design engineers usually produce *better* designs than novices [35]. Prior research has shown that the management of trade-offs [33] are critical to creative design and a key differentiator between novice and experienced designers. Without an understanding of how to configure a system in a way that limits trade-offs, inexperienced designers and designers facing entirely new design tasks are thus left at a substantial disadvantage [33]. Thus the success of product development projects in today's industrial practice is largely dependent on the tacit knowledge of experienced designers.

1.2 Research Motivation

Before this research project, I worked as a mechanical design engineer in Novo Nordisk A/S, the case company involved in this PhD. My first years working on the early-stage design of medical injection devices had a steep learning curve. My more experienced colleagues'

concepts and solutions were often selected over mine, and with good reason. They seemingly had an innate ability to configure the parts of mechanical systems in a way that achieved desired functionality but with few parts and high achievable performance. When asked how they made these configuration decisions, the answer was always a variation of "*Oh, I have designed something similar before. This is the best way to achieve everything we want at the same time".*

Having already spent years at university, I was dissatisfied with having to spend years or decades learning the underlying "craft of design". Feedback from older colleagues revealed that I was overlooking underlying dependencies between the countless design objectives involved. Therefore, I started taking an active interest in identifying dependencies in designs. Over time, this evolved into a more general interest in the identification and management of trade-offs. Realising that some trade-offs are avoidable, while others have little impact, I decided to pursue this interest through an industrial PhD.

This research was conducted as a part of the Danish Industrial Research program, run by the Innovations Fund of Denmark. Novo Nordisk is a large pharmaceutical company focused on developing treatments for diabetes and obesity (and their complications), haemophilia, and growth disorders. The company develops and manufactures its own medical devices, such as insulin injection pens. More than 30 million people with diabetes worldwide use Novo Nordisk's products, resulting in an annual production volume in excess of 1 bn cartridges of insulin and other peptide-based injectable pharmaceuticals [36]. More than 2/3 of these are delivered in *pre-filled* injection pens. Such pens are disposed of once the cartridge has been emptied. To put this number into perspective, more than 1000 devices, each consisting of 12-40 components, are manufactured and assembled every minute - around the clock.

Figure 1.4: The FlexTouch™- a pre-filled injection pen device using for the delivery of insulin, growth hormone, and other injection-only treatments.

Given the production volume, even minor improvements to these devices have a substantial impact. Being a high volume, low cost, safety intensive application that affects millions of people's health and well-being, the importance of product performance and robustness cannot be understated. It is well described that devices have a key influence on patients' adherence to their treatments [37]. Properties such as dose accuracy, dose setting resolution, dosing speed and pain, device size, and ease of use are critical.

At the same time, Novo Nordisk faces an increasing need for developing medical devices with a high level of functionality. Among other things, this is driven by growing needs for digital health solutions and sustainable, recyclable devices. These products are usually designed with the bare minimum of parts and subsystems to keep costs low and reliability high. Hence, medical devices are becoming increasingly *integrated*. The consequence of these tendencies

is an inevitable increase in the number of trade-offs, as each design variable (i.e. each dimension on the final production drawing) influences more requirements and design objectives.

These have primarily been dealt with through an increasing amount of design iterations throughout the development process. Unfortunately, this has proven time-consuming and risky since parameter variation usually rises as production volume goes up [38], revealing issues late in the process that had not been seen during prototyping. As a result, late design changes are common, which leads to late specification change and tight production tolerances [39]. Lead time and production ramp-up time have hence increased substantially. For pre-filled devices, the time to market has grown to more than six years – in comparison, *Novolet*, the first pre-filled device took 18 months to develop.

The medical devices manufactured by Novo Nordisk will, in many cases, be developed concurrently with new drug candidates, many of which have quite different physical characteristics and treatment regimes. It is hence often more cost-efficient to develop new devices from scratch. Viewed across the entire industry, drug development projects generally have a low likelihood of success. Less than 10% of drug candidates that enter the first round of trials (phase 1) end up on the market [40]. As a result, Novo Nordisk runs a large number of development projects, with a broad pipeline of projects in the conceptual stage, which narrows as products approach running production.

For these reasons, Novo Nordisk is an ideal case company to study how trade-offs can be managed systematically. The company runs a substantial number of early-stage development projects, develops products that are highly interdependent by their very nature, and work with a high degree of documentation due to regulatory requirements. This provides a substantial foundation for exploring how trade-offs might be systematically identified, analysed, and managed throughout the initial stages of product development.

1.3 Aims of this PhD

In their seminal book on engineering design, Gerhard Pahl and Wolfgang Beitz remarked that "*it is impossible to optimise the carrier of several combined functions*" [6, p. 282]. Nevertheless, they also argue that decisions on what parts and subsystems contribute to different aspects of product functionality are made early on. This touches upon what this author views as being one of the basic challenges that exist in engineering design.

Developing multi-functional systems involves dependencies, especially when the number of parts is kept as low as possible to reduce cost. Dependencies create trade-offs, which in turn determine the achievable performance of a product. Techniques for the formal analysis of trade-offs do not concern themselves with questioning why the trade-offs exist in the first place. Yet, trade-offs are typically embedded in a system based on the decisions made at a very early stage of development [12, 13].

This PhD project aims to develop systematic methods for the management of trade-offs in early mechanical design. This aim is motivated in part by the research gaps discussed in the preceding sections and in part by the challenges in industrial practice that the PhD fellow observed during his time as a mechanical design engineer.

Herein, trade-off management is taken to mean the identification and quantification of the trade-offs in a design and the subsequent identification and mitigation of their root cause(s). Specifically, the project aims to develop rigorous support for the decision making processes involved in the embodiment design stage. The hope is to support design engineers that lack the domain-specific knowledge required to avoid or mitigate trade-offs in the synthesis and improvement of new mechanical products. Thus, the developed design support is both

aimed at novice designers but also at designers facing non-trivial design challenges they have not met before. Due to the iterative and concurrent nature of design, the research also has implications for both conceptual design and detail design.

Success is achieved when contextually independent, quantitatively founded methods for identifying, evaluating, and solving the underlying issues in a system that drive trade-offs between design objectives have been developed. These methods are to be verified in theory and practice and to be proven to result in concepts and solutions that are less prone to trade-offs, with the potential to achieve robustness and potentially reduce lead time. These results are to be internally and externally consistent and have gained a degree of acceptance in the case company.

1.4 Research Questions and Hypothesis

Based on an initial literature study and the PhD fellow's experience from industry, the research was conducted with an overall hypothesis in mind along with several working hypotheses. One of these working hypotheses was refuted during the research, which significantly impacted the directions that the subsequent research took.

General hypothesis
The end performance of mechanical systems is ultimately determined by the trade-offs that affect the design of the system. It is hypothesised that the decisions made in conceptual and embodiment design in effect determine the trade-offs that exist between the different functionalities and objectives that a product is designed towards. Some of these trade-offs can be limited or prevented through deliberate attempts at avoiding certain detrimental dependencies during design synthesis and change.

Research Questions
To design a research process that allows the exploration of these hypotheses and to meet the aforementioned aims, a set of research questions were formulated. These were identified after an initial literature study, the results of which are described in the preceding sections. Thus, this PhD project seeks to answer the following research questions:

RQ1 How can the trade-offs between design objectives be identified in the concept and embodiment design phase?

The purpose of this RQ is to establish a baseline for the subsequent research, exploring existing analysis methods that allow the identification of the objectives in trade-off. If these methods exhibit limitations preventing their application to subsequent research questions, new trade-off identification methods may need to be developed.

RQ2 How can the root causes of these trade-offs be identified systematically?

The general idea is that once the objectives in the trade-off have been identified, one can concentrate analysis efforts on ascertaining *why* the trade-off exists in the first place. The answers to this research question will be utilised to explore how trade-off knowledge can be leveraged to identify design improvements.

RQ3 What approaches and solutions can be used in design to remove, mitigate, or reduce the influence of trade-offs?

In answering these three research questions, the PhD would result in an overall methodological framework that design engineers can employ in order to 1) identify the trade-offs in their designs, 2) build an understanding of what causes them through rigorous analysis, 3) utilise this knowledge to apply a systematic approach which allows the elimination or reduction of

these trade-offs, and support decision making overall. These research questions were addressed through a range of sub-studies that were collated into four work packages. These are described in detail in chapter 2, which covers the research approach employed in this PhD project.

Delimitation
This research is limited to the study of mechanical design, meaning that it concerns itself with physical products involving the analysis of mechanical properties. Given that the focus is on the study of trade-offs, a fundamental prerequisite is also that multiple design objectives are at play, or else there would be no trade-offs. Hence, this research is also limited to multifunctional products and multiobjective design tasks.

Another delimitation is that the research focuses on product development contexts that actually involve *original* or *adaptive design* [6]. This is driven by the initial literature study revealing that most trade-offs are imbued into a system by the decisions made during the early phases of design. Product development projects that do not involve the design of- or changes to the overall working principles or system structure simply do not permit design change on a level that drastically affects the trade-offs that affect the system. Hence, it does not make sense to explore methodology aimed at these types of design change in product development contexts that do not involve conceptual or embodiment design.

Finally, this research concerns itself with objectives and characteristics which are quantifiable or measurable. While one can validly argue that trade-offs also exist between less measurable (and more subjective) characteristics such as *user friendliness*, *aesthetics* and *complexity*, such aspects are not within the scope of this research. When relevant, these characteristics are replaced by more quantifiable proxies - i.e. objectives at a lower level of abstraction that contribute to a less measurable design objective. Examples of such would be to model the aspects of a product's performance that contribute to the product's user-friendliness (e.g. product mass or size, the magnitude of interaction forces, size of interaction interface, etc.), rather than attempting to describe user-friendliness explicitly.

1.5 Thesis Format

Structure
This style of this thesis is a mixture of the two most common formats; monograph and paper-based. As opposed to the standard paper-based format (which are the norm at DTU), this thesis consolidates the content of several papers written as a part of the PhD project, along with previously unpublished work. These papers are:

Paper A
Multiobjective Monotonicity Analysis: Pareto Set Dependency and Tradeoffs Causality in Configuration Design (2021)
Sigurdarson, N.S.; Eifler, T.; Ebro, M.; Papalambros, P.Y.
Published in the ASME Journal of Mechanical Design. The appended manuscript is the post-print of the final paper.

Paper B
A Novel Approach to Configuration Redesign: Using Multiobjective Monotonicity Analysis to Alter the Pareto-set (2021)
Sigurdarson, N.S.; Eifler, T.; Ebro, M.; Papalambros, P.Y.
Submitted and approved pending revisions in the ASME Journal of Mechanical Design. The appended manuscript is the revision currently under review.

Paper C, Supplementary Paper

Functional Trade-offs in the Mechanical Design of Integrated Products - Impact on Robustness and Optimisability (2019)
Sigurdarson, N.S.; Eifler, T.; Ebro, M. Published in the Proceedings of the 22nd International Conference on Engineering Design (ICED 19).

Paper D, Supplementary Paper
Limitations of Design Space-based Indicators for Early Robustness Assessment (2021)
Juul-Nyholm, H. B.; Sigurdarson, N. S.; Ebro, M.; Eifler, T.
Published in the Proceedings of the 23rd International Conference on Engineering Design (ICED 21).

Of these, Papers A and B comprise the core contributions of the research, along with the contents of chapter 6. For the sake of the readers comprehension of the overall focus and results of this PhD project, most of the content in these two papers has been built into chapters 4, 5, and 6, with a substantial amount of additional details and further methodological developments added. The supplementary papers are referenced in this thesis but are not a part of it. Paper C was written before the developments in Papers A and B were made, meaning it does contain inconsistent terminology.

The thesis is split into eight chapters. This introductory chapter is followed by a brief chapter on the research methodology of this PhD project. Subsequently, the third chapter covers the theoretical foundation of this thesis, providing a background on the well-established theories and methodologies that this thesis builds upon. This chapter is a partial answer to research question 1, as existing methods for trade-off analysis are described.

Research questions 1 and 2 are primarily addressed in chapter 4, presenting a range of theoretical and methodological developments within trade-off analysis. These allow the rigorous identification of trade-offs and their root causes in early mechanical design. This chapter is primarily an adaptation of the content of paper A, containing additional detail and methodological developments.

In chapter 5, research question 3 is addressed through methodological developments aimed at trade-off mitigation supported by analysis. The theoretical developments made in the previous chapter are leveraged to derive a novel redesign procedure. This procedure results in improvements in product performance that exceed those that may be reached through classic proportional and parametric optimization methods. This is achieved through guided configuration redesign, reducing or eliminating trade-offs among design objectives while also improving optimality overall. The chapter is an adaptation of the content of paper B, containing additional detail and methodological developments.

Chapter 6 describes the perspectives of the methodological developments in the broader engineering design process. Specifically, it addresses the applications of the analysis and redesign methods throughout different stages of the process. In doing so, a set of opportunistic approaches and heuristics are introduced, which are founded on the developments in the previous chapters. These aim to allow the underlying rationales in the redesign methodology to be used in design synthesis and decision making.

Finally, the results and limitations of the research are discussed in chapter 7, followed by the concluding remarks and suggestions for further work in chapter 8. In most of these chapters, the same case from an ongoing device development project in Novo Nordisk is used to demonstrate and contextualise the research. In Appendix 1, an overview is given of the terminology used in this thesis. Here it is worth noting that the terms *configuration design* and *embodiment design* will be used interchangeably throughout the thesis, given that this thesis lies in the intersection of two fields with different terminological traditions.

Specifically, these are the US-dominated mechanical design optimization research community to the more European leaning engineering design research community.

The format of the thesis was necessitated by some of the challenges faced in the research. Firstly, the systematic analysis and management of trade-offs in mechanical design necessitates relatively expansive cases, which are not necessarily well suited for the brief format of scientific journals. Simple design problems will commonly involve simple trade-offs, which can be trivial to identify, and may be caused by intrinsic relationships in the design problem (e.g. mechanical stiffness vs mass) or caused by an obvious design error. Thus, the monograph format permits a more expansive treatment of a real-world case, which demonstrates the important and challenging nature of trade-offs in product development.

Secondly, the PhD project involved the use of proprietary data from Novo Nordisk, which greatly complicated the writing and submission of academic papers. As the ensuing chapters hopefully reflect, the application of the theoretical and methodological developments to a real design problem resulted in several inventions and design improvements in ongoing development projects. As a result, several patent applications needed to be prepared and filed before any papers describing the work could be submitted for publication. This resulted in a timeline issue, where it was simply not feasible to get patenting documentation in place in time to allow the submission of 3+ papers in sequence. As paper B builds upon paper A, and the content in chapter 6 builds upon both, the contents of neither could be submitted for review in a scientific journal before paper A was approved. For the same reason, it was decided to prepare and submit papers A and B for review and cover the remainder of the research work in the thesis itself.

1.6 Case - Design of a Self-Orienting Millimeter Scale Applicator

	Part name	Suffix
1	Top Housing	*H*
2	Hub	*NH*
3	Plug	*P*
4	Power Spring	*PS*
5	Sealing ring	*S*
6	Needle	*N*
7	Valve	*V*
8	Base housing	*B*

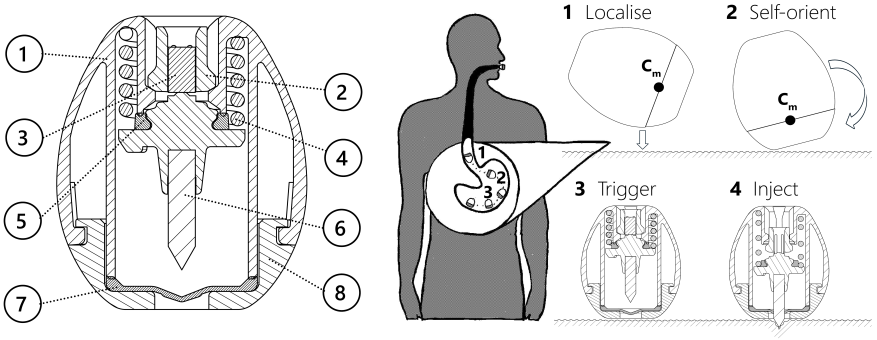


Figure 1.5: *From Paper A and B*: An overview of the SOMA device (in part) adapted from [41]. The patient swallows the device, which self-orients inside the stomach and injects a needle of pure API into gastric tissue. Here the needle dissolves, resulting in systemic uptake while the device passes through the gastrointestinal tract and out of the body.

As mentioned, a running case is used throughout this thesis. The following is an expanded description of the case given in Papers A and B. First published by Abramson et al. [41], the SOMA device (**S**elf-**O**rienting **M**illimeter-scale **A**pplicator) is a drug delivery device currently in development through a private-public collaboration between Novo Nordisk A/S and Massachusetts Institute of Technology.

The SOMA is designed for oral delivery of large proteins such as insulin, which cannot otherwise be administered orally, as the stomach breaks them down, and as they have poor permeability across the intestinal barrier. This substantially reduces the efficacy of such drugs,

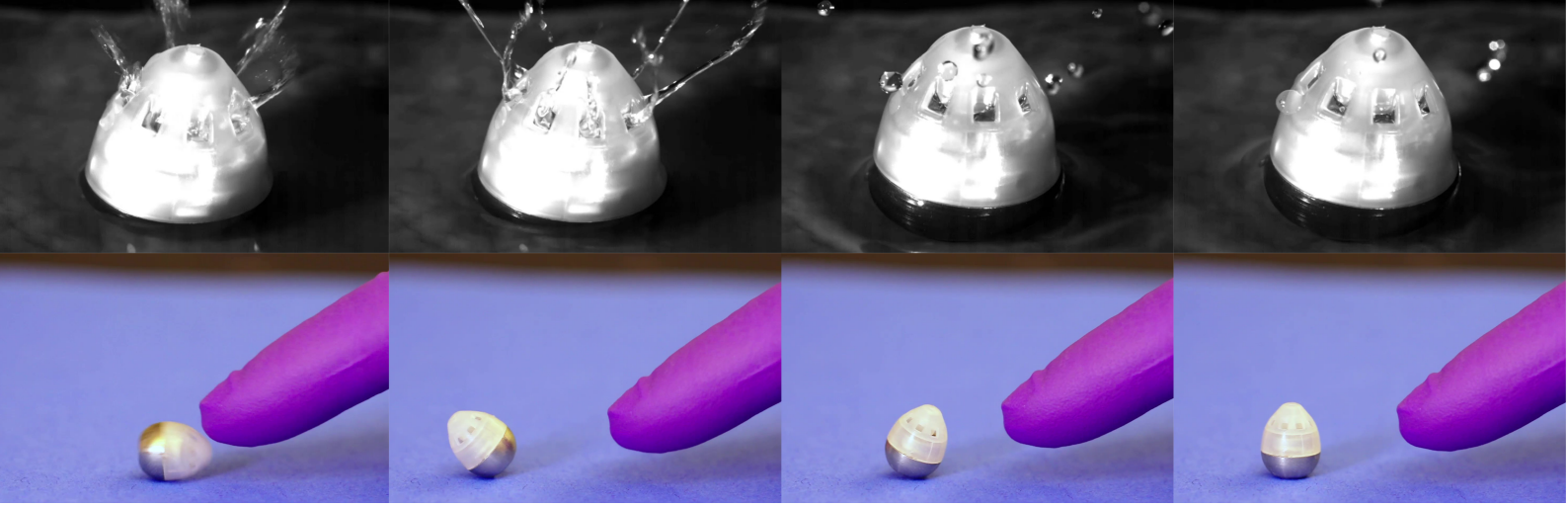


Figure 1.6: *Above*: Images from a high-speed recording of a prototype of the SOMA device triggering once exposed to liquid. Note how the recoil of the device in the third image from the left *Below:* Images from a recording of experiments done on a SOMA prototype to test its self-orientation performance. After being perturbed (far left), the device takes around one second to return to a stable position (far right).

meaning they are mostly delivered via subcutaneous injections today, e.g. with injection pen devices. In the chapters of relevance (Chapters 3-6), the design of the SOMA device will be used to demonstrate the application of the methodological developments made in the research and to illustrate their value and limitations.

Essentially, the SOMA is a pill-sized device designed to be swallowed by the user. Once in the stomach, the SOMA self-orients to a stable position due to a low centre of mass and an outer shape inspired by that of leopard tortoises [41]. Once oriented, the device injects a biodegradable needle loaded with active pharmaceutical ingredient (API) into the *submucosa* tissue-layer of the stomach (see figure 1.7 for an overview of the layers of the stomach), which has a high density of blood vessels, allowing systemic uptake.

This functionality is currently embodied with a linear spring actuator, held in place by a triggering mechanism (see Fig.1.5). The API mixture is shaped into a needle-like geometry (6) and is attached to a hub component (2) which is pre-loaded by a compression spring (4). The hub is held in place by two snap features, which are press-fit against the housing (1) by a plug (3), made out of isomalt, a dissoluble solid poly-alcohol. Once in the stomach, where the device is submerged in stomach fluid, this plug starts dissolving to a point where the spring force pushes the snap features out of engagement.

This triggers the device, with the spring pushing the needle into the stomach lining through a hole in the base (8) of the device. Until injection, the needle is kept dry in the hostile environment of the stomach by a silicone O-ring (5) and valve (7) that seal the needle inside the SOMA. The position of the centre of mass is low, as the base (8) is denser than the other parts, which aids self-orientation.

At the time of this PhD, the device was in the preliminary phases of design, still in the process of configuration, prototyping, and testing [41], and was yet to be tested on humans. While the development of the SOMA was not strictly a part of the research aims of this PhD project, it was the subject of an action research study, being used as a test case in the methodological developments presented in Chapters 4, 5, and 6.

The SOMA project commenced around the same time this PhD project began. Its outer shape was originally derived through optimization [41], but the inner configuration was iteratively developed within the constraints defined by the outer shape. Numerous configurations had been designed and built, with the one shown in Fig.1.5 showing the most promise. At the start of the action research study, the SOMA project was in the process of testing this

configuration, with the aim of manufacturing it for a "*first-human-dose*" trial.

When the PhD fellow's involvement with the SOMA project began, the project team faced several challenges that could be attributed to trade-offs between the key design objectives. Their causes were not fully understood, nor had the project had the time to quantify their influence. Given this, and that the project was in the early phases of development, the SOMA device was seen as an ideal case study. In particular, the internal configuration design of the SOMA posed several interesting trade-off challenges.

For an oral device to be viable, it needs to deliver an amount of API comparable to dosing with injection devices. This implies a dose of at least 80 units of insulin, which equates to a payload of approximately 2.8 mg of pure crystalline insulin. At the same time, the needle needs to be delivered reliably into a tissue layer deep enough to enable systemic uptake. The properties of the stomach lining are such that a large injection force is required to deliver the needle at the right depth. Hence, the challenge is to design a device that is small enough to be swallowable while reliably self orienting and injecting a sufficient amount of API deep enough. Furthermore, low cost and robust performance are essential. If only 1% of the world's 400M+ diabetics were treated with long-acting once-daily insulin from a SOMA, the annual production volume would be over 1.46bn devices. Hence, even slight improvements to the configuration may have a vast financial and societal impact.

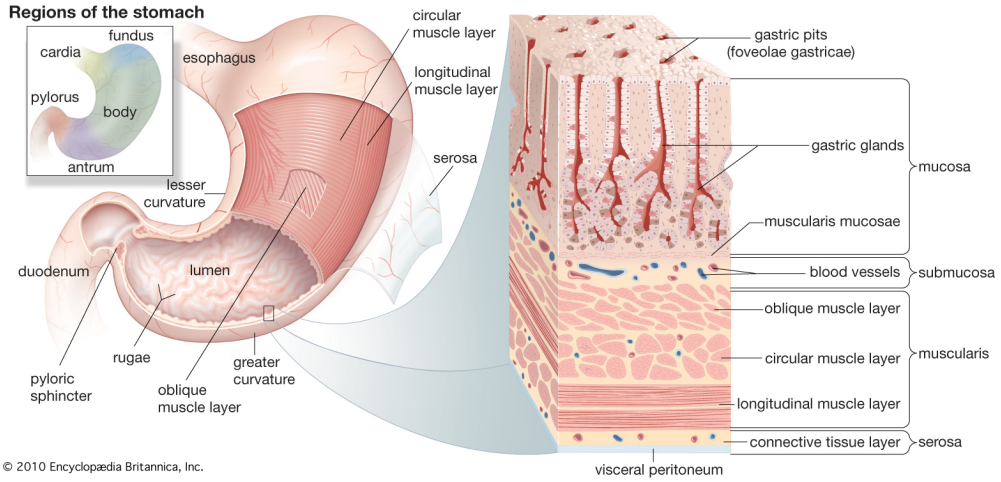


Figure 1.7: An overview of the human stomach, and its tissue layers. [42]

Thus, the action research study was conducted with the aim of identifying and understanding the trade-offs involved in the design of the SOMA device and systematically applying this knowledge to identify improvements to the configuration design. The aim was not only to use the SOMA as a test case in method development but also to provide support to the SOMA project in the process of doing so, allowing real-world verification of the value of the developed methods. As such, the study involved several sub-studies, each of which related to the different methodological developments made in this PhD project:

1. Multiobjective Optimization:
As the first step, data was gathered from the SOMA project team in order to construct a multiobjective optimization model. The data collection, model construction, and numerical solution are described at the end of chapter 3.

2. Trade-off Root cause analysis:
As a part of the development of the methods described in chapter 4, pre-and post optimality analysis of the SOMA device was performed, using said optimization model. This is described

at the end of chapter 4.

3. Systematic Design Improvement:
Using the insights gained from said methods, a novel redesign procedure was developed and applied to the SOMA device. This is described at the end of chapter 5.

4. Redesign Evaluation:
To assess the validity of the redesign procedure and to better understand the redesigns, some of the generated redesigns of the SOMA device were analysed. Amongst other things, this involved the construction and solution of a comparative optimization model. This, and supporting redesign evaluation activities, are described at the end of chapter 6.

Beyond the theoretical contributions of this work, the research was of value in the development of the SOMA device, and it resulted in several patents. So, while the invention of the SOMA device itself is by no means a part of the contributions of this research, the research did contribute to several redesigns. As much of this work is highly sensitive from an intellectual property perspective, it is unfortunately not possible to include all of the data gathered and results reached in this thesis. However, to the extent possible, measures have been taken to include as much content of interest as possible. In some cases, this does involve a degree of anonymisation, with sensitive details left out.

2 Research Approach

This chapter describes the research approaches applied throughout this PhD, intending to illustrate the scientific principles used to reach the results presented in the subsequent chapters. This chapter begins with a description of the overall research methodology. Subsequently, the methods used to conduct the research are outlined, followed by an overview of the research plan and studies of the PhD. The chapter is concluded with a description of the approach to result verification and validation.

Research practices in *design science* differ from those seen in natural sciences and in large parts of engineering science. This is not driven by a lack of scientific rigour but is rather necessitated by the challenges one faces in design research. The more *explanatory sciences* [43] (e.g. natural sciences) largely aim to describe, explain, and predict the natural world. Meanwhile, as argued by Van Aken [43], the ultimate goal of design research is to develop knowledge that design professionals can use to design solutions within their specific context.

Nevertheless, design engineers still develop artefacts within the confines of the natural world. The working principles behind the products that designers create, manufacturing processes used to realise them, and the behaviour of end-users, can largely be explained and predicted through classical explanatory sciences such as physics, chemistry, anthropology, and psychology.

As a result, the study and support of the processes involved in synthesising, refining, and realising products can, in part, involve research practices similar to those seen in the explanatory sciences. Nevertheless, design research is affected by a lack of pristine laboratory conditions. Products and systems are mostly developed by organisations, involving time-consuming and costly processes. No two product development projects are the same, given that most organisations strive to develop their own unique products, with the context and the individuals involved having a substantial influence on the end result. These projects can span from weeks to decades in length, depending on the complexity of the application. As such, it is neither temporally nor economically feasible to conduct controlled experiments in practice. Correspondingly, given the influence of context and the behaviour of the designers involved, such experiments would probably fail to produce reproducible results.

For these reasons, design science has its own research methodologies and approaches to verification and validation. This PhD relies in part on the seminal *Design Research Methodology* (DRM) developed by Blessing and Chakrabarti [4], shown in figure 2.1. DRM consists of four distinct research stages, each with its own defined means and intended outcomes. It is aimed at identifying the *current situation* in the design context being studied and developing and testing *support* which can be used to create the transition to a *desired situation.*

During the first stage, the *Research Clarification*, one seeks to find evidence or indications that the assumptions about the current state of design are valid in order to formulate a realistic, worthwhile research goal. Subsequently, a more elaborate understanding is sought in *Descriptive Study I* through more in-depth study of existing literature, analysis of empirical data, interviews, or the like. This is done to gather enough evidence to determine the crucial factors that affect the current state. In the *Prescriptive Study*, the understanding reached is

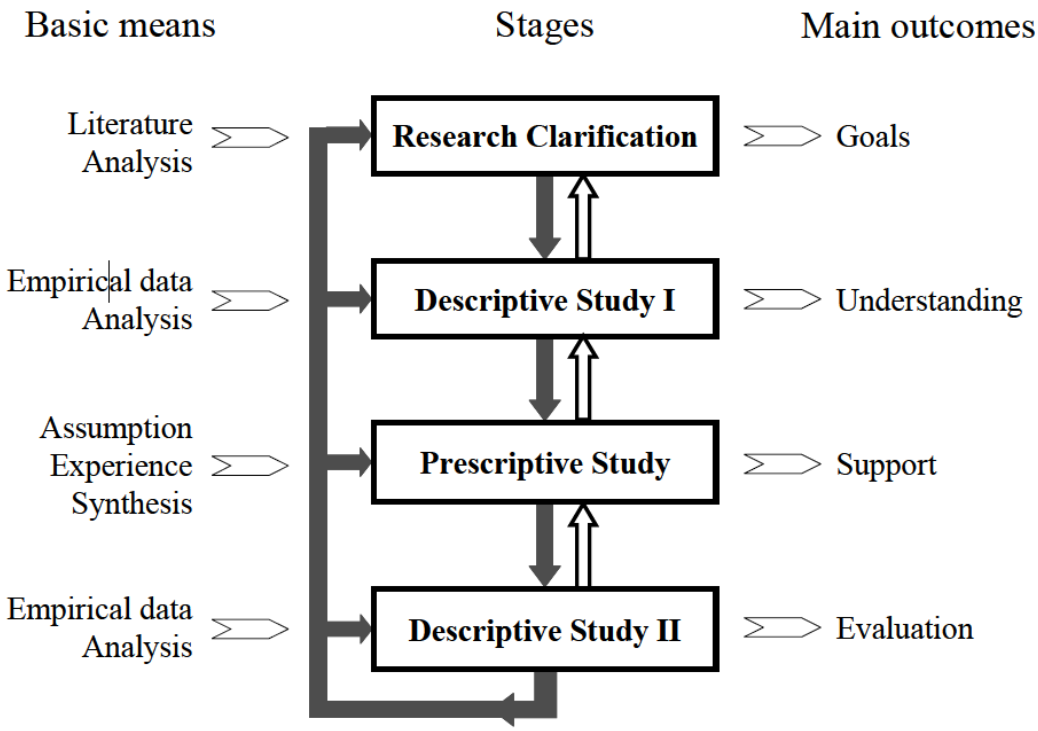


Figure 2.1: An overview [4] of different stages in the Design Research Methodology put forward by Blessing & Chakrabarti

used to develop and define so-called support, i.e. new methods, tools, processes or knowledge that might support designers or organisations in reaching the aforementioned desired situation. In the final research stage, *Descriptive Study II*, the impact, validity, and value of the support developed in the prescriptive study is assessed by analysing whether the desired situation identified in the research clarification has been achieved.

The DRM-framework has been used to structure the research plan of the PhD and guide the underlying activities. As such, the PhD was divided into five studies, each fitting within the DRM stage framework, in an effort to find answers to the research questions described in chapter 1. However, while DRM was used to structure the PhD, a degree of pragmatism has been involved in the execution of the research activities to take advantage of certain opportunities that arose during the project. As this is an industrial research project, the timing and prioritisation of the research activities were, for instance, affected when opportunities presented themselves to gather data or apply novel methods in ongoing product development projects in Novo Nordisk. Furthermore, the industrial context also resulted in unforeseen challenges, affecting the publication schedule and completion of some of the envisioned sub-studies. The use of proprietary data and the patentable inventions that arose from the research has also prevented the inclusion of all of the practice-related results in this thesis.

The research approach in this PhD also resembles the *Problem-based, Theory Based* approach put forward by Jørgensen [44]. Jørgensen argues that achievements in research can originate from different but equally valid paradigms, either from an initial problem statement, the cause of which needs to be identified, or from an initial theory or hypothesis that needs to be proven or disproven. In practice, both paradigms are seen in research and often exist concurrently or sequentially in a given research project. The application and refinement of a theory can, for instance, reveal novel problem statements, just as novel theories can be synthesised

in the study of a problem statement. Jørgensen's research framework is of relevance, as this PhD mainly originates at a fundamental problem statement: that the management of trade-offs is challenging in early mechanical design. Yet, it also involves an initial theory/hypothesis - that better products can be developed if trade-offs are managed systematically from an early stage. Correspondingly, the developments presented in Chapters 5 and 6 involve theory which was synthesised due to realisations made in the analysis of the design problem presented in Chapter 4.

2.1 Research Methods

This PhD project involves applied research; it is conducted in an industrial context focusing on mechanical design in early product development. Hence, as in most design science research, the target is to develop useful and relevant methods that support designers and product developing organisations, specifically methods that allow the management of trade-offs at an early stage of mechanical design.

As opposed to basic research, the aim is thus not to provide an exhaustive answer as to *why* trade-offs occur in development projects or to present the one and only valid approach to handling trade-offs early on. Instead, this research seeks to develop *a* generalized and rigorous approach to identifying the root causes of trade-offs and their subsequent mitigation during the initial phases of development. As shall be shown in chapter 3, such methods do currently not exist, despite the needs discussed in chapter 1.

The applied nature, and industrial context of this research, has thus influenced the selection of the research methods used to find answers to the research questions. With this in mind, the foundational research methods of this PhD are:

(Systematic) Literature Studies
The review of state-of-the-art publications and older, more foundational literature serves multiple purposes. Firstly, literature reviews serve the essential purpose of establishing a knowledge base that permits the researcher to target and scope the research towards making novel contributions. Secondly, they help contextualise the research within the broader body of research and within the existing research community. Furthermore, literature reviews may help reveal new insights that are not obvious from the study of individual contributions. Instead, they emerge from tendencies that are apparent in a specific body of work. Literature studies were used extensively throughout this research, mainly in a semi-structured form. The results of this work are primarily reflected in chapters 1 and 3.

Case Studies
Being a part of most design research, the use of case studies allows the exploration of research hypotheses in the real world. In isolation, individual case studies cannot be used to test causal relationships. However, they can be "used for exploratory research or for pre-testing some research hypotheses" as argued by Blessing and Chakrabarti [4]. They are also helpful in the absence of controllable laboratory conditions. In this research, case studies were mostly used to provide test cases in the development of new methodology.

Archival Analysis
As argued by Blessing & Chakrabarti [4], retrospective analysis data collection is cardinal in design research. Development projects often occur over long periods of time, meaning the collection of data from multiple projects, in multiple organisations, in multiple contexts is not practical unless it is done retrospectively. The analysis of archival data, such as sketches, engineering drawings and CAD models, analysis models, experiment and simulation results etc., thus allows the exploration of and description of patterns and tendencies in the design

process. This yields valuable information upon which to test or build hypotheses, theories, or new models. In this research, archival analysis was used extensively to better understand the challenges that arise throughout the development process in the case company.

Formal Analysis
This PhD project relies heavily on the rigour of formal design analysis methods - i.e. mathematics - as a means to identify and understand the design challenges observed in the case studies. Applied mathematical analysis is foundational to many of the natural sciences. It is used in this research with the hope that it might bring a degree of generality and rigour to the results, which would not be achievable through purely qualitative work within the project's time frame. This PhD especially relies on multiobjective design optimization, on monotonicity analysis [45]), and their related mathematical foundations (theorems and proofs). Yet, optimization techniques are used here as a form of design analysis, specifically what the optimization results and the model itself reveal about a given system, rather than aiming to simply identify its optimal proportions. The methodology presented in chapters 4 and 5 is developed on a mathematical foundation and relies on a set of theorems, corollaries, and proofs.

Action Research
Originating from social sciences, *action research* is increasingly prevalent in design research [4]. It is essentially a structured approach to iteratively introducing and evaluating change in organisations, relying on the researcher's participation in the organisation being studied. As such, the subjects being studied (the design engineers) become stakeholders in the research. Hence, the researcher helps convert descriptive research into practice or action and evaluates the effect concurrently. In action research, research is "*directed to solving problems in the world.... Together, stakeholders and researcher co-create knowledge*" [46]. This helps ensure the practical relevance of the results, at the potential sacrifice of generality, unless action research is conducted across a broad slate of organisations. In many ways, action research is somewhat analogous to the latter stages of DRM, as these also involve iterative introduction and evaluation of actions (i.e. design methods) that create change. The running case included in most of the chapters of this dissertation was also a subject of more targeted action research throughout the latter two-thirds of the PhD. The PhD fellow participated in the SOMA project in a role resembling that of a consultant, tasked with helping the project identify the causes and solutions to technical challenges which were hindering progress. This task was used as an opportunity to build and test novel analysis and redesign methods, using insights gained from literature review and prior case studies. To ensure generality, this work focused on developing mathematically founded and context-independent approaches.

2.2 Studies and Research Plan

To find answers to the research questions defined in chapter 1, the PhD project was divided into several sub-studies, each with specific research methods and empirical data involved. These sub-studies were collated into four distinct work packages. The resulting overall structure of the research project is illustrated in figure 2.2.

The work packages were mostly carried out in sequence. However, they did involve overlaps to allow pre-evaluation of new methods, additional literature reviews to qualify ongoing method development further, and exploration of the effect of using well-known methods in a new context. Thus, three work packages involve combinations of descriptive, prescriptive, and clarifying activities, while each sub-study fell into distinct stages of the DRM model. This also reflects the iterative nature of the DRM framework, where the researcher may "loop" back and forward between building understanding, developing design support, and evaluating said support.

Work package 1 - Trade-off Identification

Clarifying and descriptive research of the current state

Carried out in 2018 and early 2019, this work package involved the exploration of the current state of methods and practices involved in early-stage trade-off management. The purpose of this work was to identify research aims, build a more detailed understanding of existing research work within the field, and ensure the relevance of the research to design practice. This work started with an initial literature review and the PhD fellow's *on-boarding* process in the part of the case company that deals with mechanical analyses and technical challenges in medical devices. This allowed the collection and analysis of data such as issue slides from project milestones and the related technical reports and involvement in ongoing trade-off management activities. This resulted in a deepened understanding of the challenges with trade-offs faced in practice and already described in literature.

This work informed the identification of the research questions and core hypotheses, resulting in the overall research plan, and led into a set of subsequent sub-studies, with the intent of exploring existing trade-off identification and analysis methods. These studies involved archival analysis of a product from the case company that is already in the market and a case study in an ongoing product development project. These studies led to the realisation that one of the core assumptions of the research was invalid, leading to re-scoping of the project towards more mathematically, and less heuristically founded design methods. These sub-studies and the subsequent transition are briefly described in section 4.1.

Work package 2: Trade-off Root cause Analysis

Descriptive-prescriptive research within analysis methods

Carried out in 2019, the purpose of this work package is to identify a systematic approach to trade-off root cause identification. This involved the application of existing design optimisation methods to an early-stage design. Much of this work was carried out in preparation for and during a research stay at the Optimal Design Laboratory at the University of Michigan. In this work, several gaps in existing methodology were identified, relating to the early stage applicability of optimisation methods and their focus on the optimal result itself, rather than more holistic design analysis. These resulted in the subsequent synthesis of novel design optimisation methods to meet the purposes of this research. This work resulted in paper A and most of the content of chapter 4 (excluding section 4.1.) and sections 1.6 and 3.5.

This research relied heavily on mathematical analysis and formal proofs, optimization literature, and upon the collection of the archival data from the SOMA device. Much of this work package was conducted in action research form, with an optimisation model being built. This model was used to support the SOMA project with trade-off identification decision making. It was also used as a test case in the development of novel root cause analysis methods, which later provided the project with input for redesign.

Work package 3: Trade-off Mitigation

Mostly prescriptive research on design method synthesis

Carried out in late 2019 and most of 2020, this work package aimed to develop systematic and rigorously founded methods and design principles that might *support* designers in reducing or avoiding contributions to trade-offs between design objectives through design change. This involved the development of redesign methods that could utilise the outputs of the aforementioned analysis to identify promising redesigns. By ensuring that the analysis methodology was general to most early-stage design optimisation problems, the underlying mathematics could be used to derive a set of principles that result in design improvement when applied systematically. These were not only mathematically consistent but also consistent with numerous existing heuristics. Thus, the research method is a combination of formal

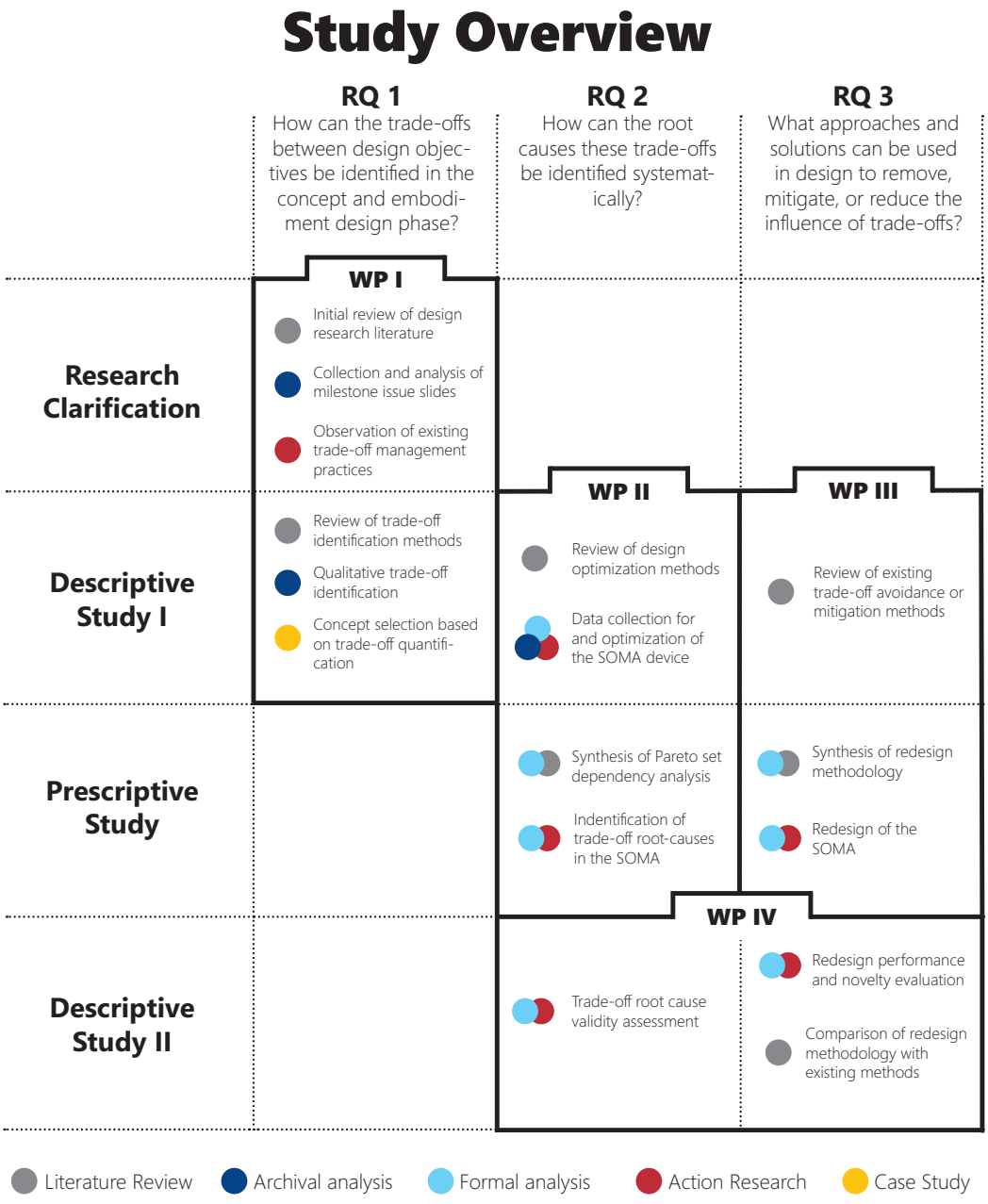


Figure 2.2: How the DRM stages, research questions, and sub-studies (with associated research methods) fit together into a set of work packages, forming an overall research design

analysis and literature review.

This work resulted in most of the content of paper B and the content of Chapter 5. The redesign methodology was applied to the SOMA design (without the involvement of the stakeholders in the development project) to derive a series of redesigns. If the method were valid, this would substantially reduce trade-offs and improve overall performance compared to the initial design analysed in the preceding study.

Work Package 4: Method Evaluation and Result Implications in Design
Descriptive work to evaluate the new state created by the prescriptive work

The final work package of the research aimed at evaluating the developed analysis and redesign methodology for the purposes of verification and validation. Evaluating the success of research results is somewhat less straightforward in design science than in the more explanatory sciences. This evaluation was conducted in different ways. First, the validity of the trade-off root cause analysis method was assessed by comparing the outputs of the root cause analysis with the optimization results from the SOMA case. Secondly, the impact of the redesign process of the SOMA device was assessed quantitatively. This assessment sought to answer whether the redesigns were better than the original design (the starting point of the study). The idea here was that if it could be proven that the new designs exhibited measurably lessened trade-off and improved performance, then one could claim with high confidence that the methodology is valid in this specific context. Combined with the generality of the mathematical foundations of the methodology, this evaluation of success in the SOMA case would also support the claim that the methodology is generally valid and valuable.

Furthermore, the consistency and value of the analysis and redesign methods was assessed through comparison with existing methods. Combined with the previous efforts, these evaluation activities lead to an exploration of how the developed methodology and its implications in design could be expanded to be employed beyond supporting mere redesign activities. This specifically focused on the implications of the methodology in the context of design synthesis and decision making through the product development process. The output of this work resulted in the content of Chapter 6 and also contributed to the case study in paper B.

2.3 Verification and Validation

As mentioned, design research can present substantial challenges in verification and validation. How does one empirically measure the influence of novel design support upon the design process without being able to perform controlled experiments within product developing organisations?

This has been a topic of quite some discussion in design research. Several different approaches have been suggested, which allow a degree of verification and validation despite this fundamental challenge; examples include the *Validation Square* by Pedersen et al. [47], Frey and Dym's medical science inspired approach [48], and the aforementioned DRM framework [4]. As discussed by Blessing and Chakrabarti [4], the understanding of what verification and validation (V&V) involve is inconsistent when viewed across different fields of research. Here, V&V is perceived in the meaning that is common in systems modelling [49]; namely that verification refers to the assessment of internal consistency, while validation refers to the justification of the new knowledge claimed.

The primary contributions of this PhD project are founded upon novel mathematical developments - specifically, these are extensions to monotonicity analysis which allow systematic trade-off root cause analysis. As such, certain aspects of the results of this research are simpler to verify and validate than much design research. As the research is delimited to mechanical design and involves properties and characteristics which can be expressed through mathematical models, we can, to a certain extent, rely on quantitative analysis to assess validity. As the ultimate goal for this research is to support designers in reducing or eliminating trade-offs between design objectives, during the conceptual or embodiment design phases, we can in part validate the results by answering two simple questions:

1. Does the application of the developed methodologies result in designs that have measurably improved performance and reduced trade-off?

2. Do the optimization models used to identify redesigns, and subsequently, compare

them, yield repeatable results, real design variable values (e.g. non-negative or infinite dimensions) and do they sufficiently approximate the behaviour of the real system?

The first aspect is addressed in Chapter 6, while the second is addressed in Chapter 4. In both cases, there is already a wide range of well-accepted approaches in existence. Yet this assessment can only be performed on a case basis. How do we then sufficiently verify and validate the generality of the research?

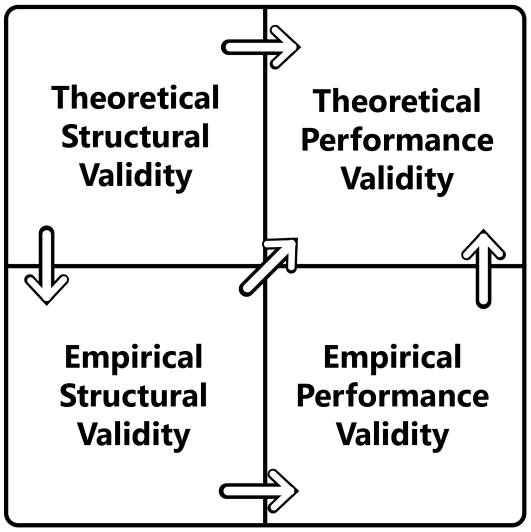


Figure 2.3: The *Validation Square* put forward by Pedersen et al, reproduced from [47]

One approach to this is assembled in the "Validation square" introduced by Pedersen et al. (51) (illustrated in figure 2.3), aiming at assessing whether a novel design method meets its purpose in an effective and efficient manner. Based on the notion that design methods are commonly applied to test cases or examples, the validation square presents a systematic approach to validating novel design methods. The validation square divides V&V into four parts, relating to different aspects of the methodology and the case(s) used to test it:

1. *Theoretical structural validity:* Is the underlying theory behind the design method well accepted, and is the method consistent? Are the constituent elements of a method and the method as a whole internally consistent?
2. *Empirical structural validity:* Are the example problems used to verify the method appropriate?
3. *Empirical performance validity*: Are the results of the application of the method useful in the studied case(s) - does the method meet its initial purpose?
4. *Theoretical performance validity*: Is the usefulness of the support from the empirical case study generalisable beyond the case?

This approach will be used in Chapter 7, where the validity of the research is discussed.

3 Theoretical Foundation

In this chapter, the theoretical foundations for this PhD is described, with the aim to position the research in the body of existing knowledge. Thus, it also provides a non-exhaustive overview of the state of the art, in part providing an answer to RQ1. Furthermore, this chapter aims to give sufficient background to support the reader. This is done to the extent necessary for readers with an engineering background to comprehend the ensuing chapters. As many design engineering focused readers might be unfamiliar with design optimisation techniques, the basics and the details of relevance are covered in greater detail; much of this content is based upon the excellent textbooks by Papalambros & Wilde [12], and by Arora [50]. For even further details, one can refer to the cited literature. The chapter also contains sections on dependency analysis, heuristic design methods, and quantitatively founded design methods, which is content that has been adapted from Paper B. The chapter is concluded a section on the application of existing optimization methods for the construction of a multiobjective optimization model for the SOMA device. This content on adapted from Papers A and B, albeit with a much higher degree of detail.

3.1 The Design Process

Industrial practice is characterised by a varying degree of formalisation and systematisation of the design process, with practices of individual organisations being defined by the design context, the corporate culture, and the resources available. Multiple design process theories have been put forward in an attempt to describe and generalise the activities and processes involved in bringing a product from idea to running production and use. Examples include models by Pahl & Beitz[6], French [14], Ullman [51], Ulrich & Eppinger [7], and Andreassen & Hein [8]. While these differ in various ways, they do have the common trait that the design process is split into a series of distinct phases involving certain activities, deliverables, and milestones, starting at the initial idea and continuing until running production. There is largely agreement that the initial stages revolve around the synthesis and refinement of the overall concept and the product structure (aka. embodiment design [6], scheme[14], layout [10], configuration design [11], system architecture [15]).

These are largely idealisations of processes that may be far from linear in industrial practice [24]. This is also acknowledged by most of the well-accepted design process theories [6, 7, 14]. In industry, the process/sequence can be more complex, with activities and deliverables falling in a different sequence, for instance, due to partial design reuse (e.g. of parts or modules), the concurrent development of multiple competing concepts or options for market launch, front-loading of certain activities that might be critical in the given context, and loop-backs due to manufacturing issues or dissatisfactory performance revealed in late-stage tests. Depending on the type of development project, certain phases might also be skipped entirely. For instance, the development of a new aircraft does not necessarily involve changes on a conceptual level to the overall system, but certain sub-systems or modules might be changed or entirely new. Thus, it is important to acknowledge that design processes in practice involve different degrees of design change and development at different levels of abstraction. Pahl and Beitz [6] hence distinguish between design projects involving *original designs*, *adaptive designs*, and *variant designs*.

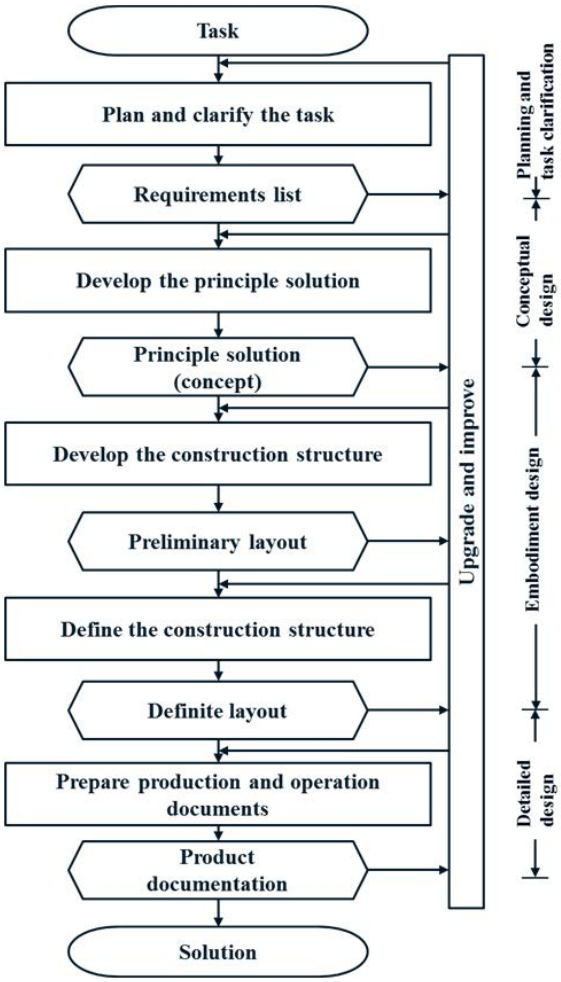


Figure 3.1: The phases of the engineering design process, as put forward by Pahl & Beitz [6])

For the sake of consistency, this thesis will mostly rely on the terminology and concepts introduced by Pahl & Beitz [6]. The focus of most of the research questions and the stated purposes of the research project is on the activities and decisions that usually occur during the embodiment design phase. Nevertheless, many of the activities in the conceptual design phase can be difficult to separate from embodiment, as it can be difficult for design engineers to synthesise a concept without thinking in terms of a geometric realisation [6]. As a result, some of the work in the thesis also relates to what Pahl & Beitz termed *working principles* and *working structures*, which they defined as: "*the combination of the physical effect with the geometric and material characteristics (working surfaces, working motions and materials) allows the principle of the solution to emerge... The combination of several working principles results in the working structure.*" [6].

3.2 Design Optimisation

Papalambros and Wilde [12] define design optimisation as a mathematical approach to finding the *best* design within the *available means*; i.e. the highest performing solution achievable within the limitations of the physical world. This is done by modelling the measures of goodness through explicit or numerical functions. These describe the relationship between the properties the designer can manipulate directly, referred to as *design variables* and *design parameters*, and said measures of goodness, referred to as *design objectives*. These objectives may describe how *well* the product performs its functionality, or more general characteristics

that are desirable to the organisation developing the product, the user/customer, or society as a whole (e.g. cost, maintainability, environmental impact, and so forth).

Design variables are typically continuous or discrete properties where the designer can change freely. Parameters, meanwhile, are fixed values determined by decisions made between a set of alternative options (e.g. material selection, safety factors, part shape parameters, and so on). The available means, meanwhile, are modelled through so-called *constraint functions,* which describe the conditions the design needs to meet in order to be feasible. Examples include one part fitting inside another, the components not breaking or yielding when loaded during operation, and the parts having at least a certain wall thickness in order to be manufacturable. Design objectives are typically either maximised or minimised, meaning that *nominal is best* requirements can be modelled by minimising the deviation from a given target.

Solving optimization problems

Optimisation problems can be stated in numerous standard forms, where all objective and constraint functions are stated in the same mathematical form, allowing it to be solved numerically. The most common of these is the so-called *negative-null* formulation:

$$\begin{aligned} \text{min.} \quad & \mathbf{f}(\mathbf{x}) && (3.1)\\ \text{subject to} \quad & \mathbf{g}(\mathbf{x}) \leq 0 && (3.2)\\ & \mathbf{h}(\mathbf{x}) = 0 && (3.3)\\ & \mathbf{x} \in \mathbb{P} && (3.4) \end{aligned}$$

Here **f(x)** is a vector of the design objectives, f_i, $i = [1, 2, ..., k]^T$, that describe the performance of a product as a function of the vector of design variables **x**, and the vector of fixed design parameters **P**. Any maximization objectives are transformed into minimization objectives by simply multiplying the term by -1. Meanwhile **h(x)** and **g(x)** are the equality and inequality constraints respectively. If $\mathcal{X}$ denotes the feasible domain - i.e. the set of values of **x** that fulfill the constraint functions, then the attainable set $\mathcal{A}$ contains all values of $\mathbf{f}(\mathbf{x})|x \in \mathcal{X}$. If no trade-off exists between the design objectives, then a single optimum $\mathbf{f}^*(\mathbf{x}^*)$ exists. In this situation, it would be sufficient to identify the optimum of one objective f_i:

$$\begin{aligned} \text{min.} \quad & f_i(\mathbf{x}; \mathbf{P}) && (3.5)\\ \text{subject to} \quad & \mathbf{g}(\mathbf{x}; \mathbf{P}) \leq 0 && (3.6)\\ & \mathbf{h}(\mathbf{x}; \mathbf{P}) = 0 && (3.7)\\ & \mathbf{x} \in \mathbb{P} && (3.8) \end{aligned}$$

Such single-objective problems can be solved using gradient based solution methods. Identifying the optimum of unconstrained problems is a question of identifying the values of **x** where the conditions of optimality are fulfilled:

1. **First order necessary condition** (*FONC*): A function $f(\mathbf{x})$ has a local minimum at $\mathbf{x}^*$ if $\nabla f(\mathbf{x}^*) = 0$
2. **Second order sufficiency condition** (*SOSC)*: A stationary point $\mathbf{x}^*$ is a minimum if the Hessian of the f is positive definite at $\mathbf{x}^*$

It is not a given that points that fulfil the FONC are minima; other stationary points such as maxima and saddle points also fulfil this condition (see figure 3.2). Thus, it is only a necessary condition - we are not sufficiently satisfied a candidate optimum is, in fact, a minimum. As such, confirming that a candidate optimum fulfils the SOFC assures that an identified candidate optimum point is actually a minimum. In nonlinear, multivariate problems, the identification of minima that fulfil these conditions is typically done using an iterative line search approach, using, e.g. *quasi-Newton methods* or *steepest decent* to identify the direction in which a function is decreasing.

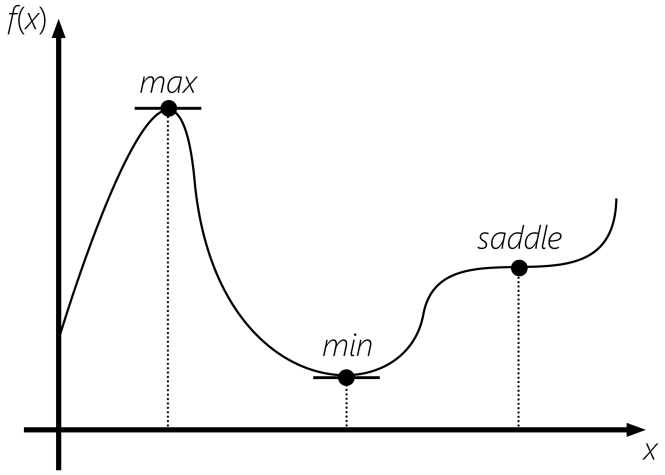


Figure 3.2: A stationary point is not necessarily a minimum. Hence, it is necessary to evaluate whether a candidate optimum point fulfills the necessary and sufficient conditions for optimality. Adapted from Papalambros & Wilde [12]

Yet, for all intents and purposes, optimisation in engineering design involves constraint, as the design of products obviously involves limitations as to what can be achieved. These can range from practical considerations such as limitations to part shape and size stemming from a given production process to unavoidable limitations defined by the physical world such as material properties and geometric fits between parts.

The issue with constrained problems, is that the optimum will not necessarily exist at a stationary point of the objective; the constraints may define $\mathbf{x}^*$. Consider the problem:

$$\begin{aligned} \text{min.} \quad & f(x_1) = x_1^2 + 3 && (3.9)\\ \text{subject to} \quad & g_1(x_1) = 10 - 2x_1 \leq 0 && (3.10)\\ & g_2(x_1) = x_1 - 20 \leq 0 && (3.11)\\ & \mathbf{x} \in \mathbb{P} && (3.12) \end{aligned}$$

Disregarding the constraint functions, the minimum of $f(x_1)$ exists at $x_1 = 0$. Yet, as this violates a inequality constraint, g_1, this minimum is not feasible, and therefore not optimal. In fact, the minimum value of $f(x_1)$ exists at the minimum value of x_1 that fulfills g_1, which is $x_1 = 5$, meaning $g_1(x_1) = 0$. Hence, g_1, which is satisfied at equality, affects the optimum of f. In such situations, the inequality constraint is said to be *active*.

Now obviously, the the gradient of f in this example is not 0 at the optimum. Hence, an adaptation to the necessary and sufficient conditions of optimality is required, to account

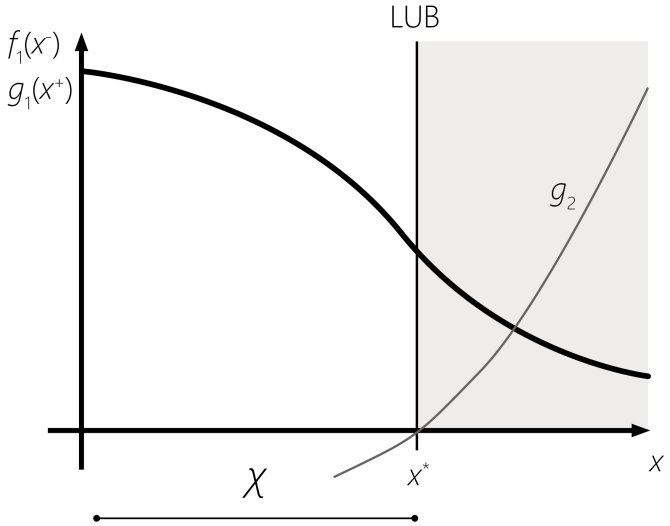


Figure 3.3: The effect an active constraint has on the location of the optimum. Notice how g_1 determines the *lowest upper bound* of *x*. Adapted from Papalambros & Wilde [12]

for the influence of constraints. Instead, the problem in the form shown in eq. 3.10-3.12 is transformed into a *Lagrangian function*:

$$\mathcal{L}(\mathbf{x}, \lambda, \mu) = f(\mathbf{x}) + \lambda \mathbf{h}(\mathbf{x}) + \mu \mathbf{g}(\mathbf{x}) \tag{3.13}$$

where λ and μ are the so-called *Lagrange multipliers*. This transformation allows the computation of the gradient, which in fact does allow the identification of the optimum of a problem subject to constraints. In this form, a candidate optimum $\mathbf{x}^*$ must fulfil the so-called Karush-Kuhn-Tucker (KKT) necessary conditions [52, 53] to be a minimum:

1. $\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda, \mu) = 0$
2. $\mathbf{g}(\mathbf{x}) \leq 0, \mathbf{h}(\mathbf{x}) = 0$
3. $\lambda \neq 0, \mu \geq 0$
4. $\mu^T \mathbf{g}(\mathbf{x}) = 0$

These conditions ultimately mean that $\mathcal{L}(\mathbf{x}^*, \lambda, \mu)$ is a stationary point, that the constraints are fulfilled, and that the active constraints have non-zero Lagrange Multiplier (i.e. that the active constraint affects the optimum). Correspondingly, the SOSC is that the Hessian of the Lagrangian is *positive definite in the feasible direction.* For more on this, see chapt. 5 in Papalambros & Wilde [12] or chapt. 4 in Arora [50].

Evaluating points at random to identify values of $\mathbf{x}$ that fulfil these conditions is, of course, not an efficient approach. In linear problems, these conditions can be used to solve for minima algebraically. This is not possible for nonlinear problems. Here, the most common approach is to utilise an initial guess $\mathbf{x}_0$, and then compute the gradient and Hessian, which indicates in which direction each variable should be adjusted, in order to move in a descending (i.e. minimizing) and feasible direction, and how *far* they could be adjusted. This yields a new point $\mathbf{x_1}$, after which the process is repeated until the gradient information points to that an optimum has been found. This computation of a suitable direction and step size is called *line search*. Computing the gradient of the Lagrangian, especially in large problems with hundreds or thousands of variables and constraint functions, can be extremely computationally expensive. For the same reason, there are numerous alternative algorithms that

approximate these in a very efficient manner or which use other means to reduce computational cost. Examples include sequential quadratic programming, interior-point methods, active set strategies, and barrier methods [12, 50].

Solving Multi-objective Problems

Now, as mentioned, this is only strictly possible for a single objective problem or for multiobjective problems without any trade-offs. In the presence of trade-offs, the notion of optimality changes somewhat, as we cannot unequivocally define what *best* constitutes when there are multiple, conflicting objectives. That is unless we have a clear and unambiguous expression of their relative importance. This might for instance be done by transforming the problem into a single objective problem, introducing a *global criterion* at a higher level of abstraction, e.g. *utility*, *profit*, or *customer satisfaction* maximisation.

This requires a substantial amount of data to ensure that all aspects of utility and contributors to it are accounted for. This might not be achievable, and will in many cases, inevitably involve subjective assessment. Hence it is common to avoid attempting to find a single optimal solution but rather to identify a set. Here, the notion of *Pareto Optimality* comes in. A point $\mathbf{f_0}(\mathbf{x}^*)$ in the attainable set $\mathcal{A}$ is said to be Pareto-optimal if and only if there exists no point in $\mathcal{A}$ that fulfils:

$$\mathbf{f}(\mathbf{x}) \leq \mathbf{f_0}(\mathbf{x}^*) \qquad \wedge \qquad f_i(\mathbf{x}) < f_i(\mathbf{x}^*) \tag{3.14}$$

Expressed in less formal terms, a Pareto optimal solution is one where no single objective can be improved further, without worsening another objective. The set of all Pareto-optimal points is the Pareto set $\mathcal{C}$, which exists of the boundary of $\mathcal{A}$, facing origin in a negative-null formulation.

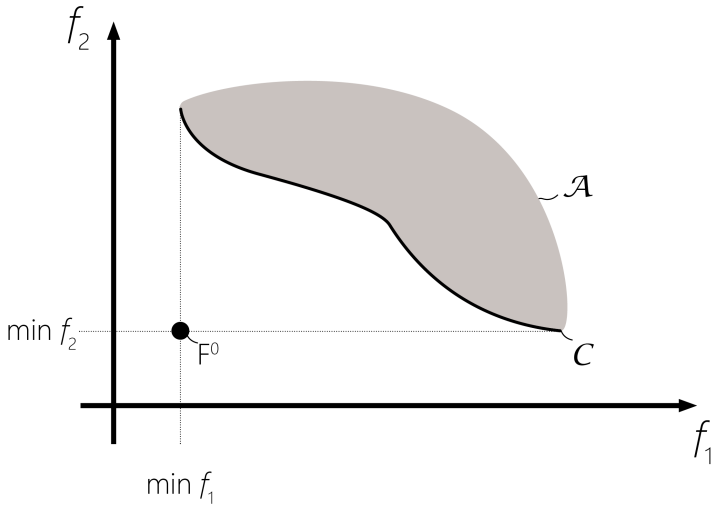


Figure 3.4: An example of the location of the Pareto set ($\mathcal{C}$) in the attainable set ($\mathcal{A}$.) in a bi-objective minimisation problem

The gradient-based methods discussed can only find the optimum of a single objective function subject to constraints. To identify a set of optima rather than a single one, it is hence either necessary to transform the multiobjective problem into a single objective problem that can be solved iteratively to identify different Pareto points or rely on methods that do not compute a gradient. For gradient-based methods, this may involve exploring a region of the Pareto set based on a decision maker's preference (some regions of the set might simply be known to be unacceptable a priori). Alternatively, one might attempt to exhaustively

identify the contours of the Pareto set itself. For a full overview of the most common solution methods, see Marler and Arora [54]. Optimization algorithms also exist that do rely on the computation of a gradient, meaning that they can be as effective at solving multiobjective problems as single objective ones [54]. These non-gradient based algorithms essentially search for optima within the objective space, either through a stochastic or deterministic approach. Common non-gradient methods include genetic algorithms, direct search, simulated annealing, and EGO (efficient global optimisation) [12, 50].

Two of the simplest solution methods are the weighted sum formulation and the ϵ-constraint method (aka. the upper-bound formulation[12] or the bound objective method [50]). These exist in various alternative algorithmic implementations and also have related formulations that have different benefits, depending on the size of the problem and the shape of the attainable set (whether it is convex or not). In broad terms, however, weighted sum formulation involves gathering the objectives in an aggregate function:

$$\text{min.} \quad U = \sum_{i=1}^{k} w_i f_i(\mathbf{x}) \tag{3.15}$$

$$\text{s.j.t} \quad \mathbf{g}(\mathbf{x}) \leq 0 \tag{3.16}$$

$$\mathbf{h}(\mathbf{x}) = 0 \tag{3.17}$$

$$\mathbf{x}, \mathbf{w} \in \mathbb{P} \tag{3.18}$$

where **w** is a vector of weighting coefficients w_i for each objective f_i. When this formulation is solved iteratively for different values of **w**, then Pareto points may be identified. There are different strategies for selecting combinations of weighting coefficients, just as one can use quasi-random sampling methods to get a set of uniformly distributed weighting coefficients within a given range, to achieve a low discrepancy Pareto set. Related formulations include the weighted product method, and the exponential weighted criterion [55].

The ϵ-constraint formulation meanwhile, involves converting the problem in Eq. 3.5-3.8 into a single objective problem, where the remaining objectives are represented though additional constraints:

$$\text{min.} \quad f(\mathbf{x}) \tag{3.19}$$

$$\text{s.j.t} \quad \mathbf{c}(\mathbf{x}; \epsilon) \leq 0 \tag{3.20}$$

$$\mathbf{g}(\mathbf{x}) \leq 0 \tag{3.21}$$

$$\mathbf{h}(\mathbf{x}) = 0 \tag{3.22}$$

$$\mathbf{x}, \epsilon \in \mathbb{P} \tag{3.23}$$

Originally put forward by Haimes [56] and later refined by Carmichael [57], this formulation includes $\mathbf{c}(\mathbf{x}; \epsilon))$ which is a $k-1$ dimensional vector of *bound objectives* in the form $c_i(\mathbf{x}, \epsilon_i) = f_{i+1}(\mathbf{x}) - \epsilon_i \leq 0$ or $c_i(\mathbf{x}, \epsilon_i) = \epsilon_i - f_{i+1}(\mathbf{x}) \leq 0$, $i = [1, 2, .., (k-1)]$. The vector ϵ represents the upper bound parameters of the objectives. When $f(\mathbf{x})$ is minimised for given values of ϵ, then the solution $\mathbf{x}^*$ is Pareto optimal if all of the bound objectives are active with non-zero Lagrange multipliers. By varying ϵ systematically between lower $\epsilon_{\mathbf{L}}$ and upper limits $\epsilon_{\mathbf{U}}$, one can identify Pareto points.

Yet, these formulations have benefits and limitations. In fact, any optimisation algorithm and formulation comes with its own strengths and weaknesses - this is often referred to as the *no free lunch* theorem. See Marler and Arora [54] for an overview of the most common multiobjective formulations, their strengths and weaknesses, and for references to works on the underlying mathematics.

3.2.1 Pre and Post Optimality Analysis

Recall the notion of constraint activity; that an inequality constraint is satisfied at equality, i.e. $g_i(\mathbf{x}^*) = 0$, and determines the location of the optimum. Such constraints are common in design [58]. Consider the mechanical stress in a component. Engineers do not, per se, have any preference as to what the stress level should be, so long as it stays below a certain limit (e.g. a yield, creep, or fatigue limit), often with a certain safety factor attached. Yet, some load cases will end up determining how a component is dimensioned, to a point where the load case ends up influencing the design of the overall system. This is essentially a question of constraints that are active, and some engineers simply know intuitively [59], which load cases determine the dimensioning of a system. This, in turn, allows them to design feasible systems with less trial and error.

Ultimately, active constraints actually remove *degrees of freedom* from a design problem. In the context of optimisation, this is not meant in the same sense as in kinematic design or structural mechanics. Rather, it reflects the number of variables the optimisation algorithm can adjust freely, as the value of the remaining variables is determined by active constraints. For instance, in a problem with five variables, and four active constraints, the value of four variables can be expressed as a function of the fifth variable. These relationships will be determined by the active constraints. Thus, in problems where the sum of active inequality constraints and equality constraints equals the number of design variables, the optimal solution will be defined by the constraints, as the number of unknowns (design variables) is the same as the number of functions (active constraints). These are known as *constraint bound solutions* [60]

Correspondingly, an active constraint reveals the relationships between the design variables that exist at the optimum. They are, in other words, dependencies specific to the optimal design, which is useful information about the nature of the design problem, as with the example of load cases. Identifying active constraints also has computational benefits; as a variable is no longer freely adjustable when a constraint is active, the constraint function itself can be solved w.r.t. one of its dependent variables. Substituting this solution back into the optimisation problem eliminates the active constraint and the bound variable from the problem. Thus, the size of the problem is reduced, reducing the computational cost correspondingly. This is known as *back-substitution* or *partial minimization*. Correspondingly, the eliminated variable is said to have been *optimized out* of the problem.

First put forward by Papalambros and Wilde [45], *Monotonicity Analysis* (MA) is a rigorous analysis method that allows the identification of active constraints without performing computation. It thus provides a systematic approach to performing these model reductions.

MA involves assessing the *monotonicity* of the optimisation problem. A function is said to be monotonically increasing with respect to a variable x if $f(x_2) > f(x_1)$ for any $x_2 > x_1$. This relationship is denoted $f(x^+)$. Correspondingly, if $f(x_2) < f(x_1)$ for any $x_2 > x_1$, the function is said to be monotonically deceasing w.r.t. x, which is denoted as $f(x^-)$. In the presence of monotonicity in the objective and constraint functions„ the following principles [12] can be used in single-objective problems to find active constraints:

First monotonicity principle (MP1)
In a well-constrained minimization problem, every increasing variable is bounded below by at least one non-increasing active constraint.

Second monotonicity principle (MP2)
In a well-constrained minimization problem, every nonobjective variable is bounded both below by at least one non-increasing semi-active constraint and above by at least one non-

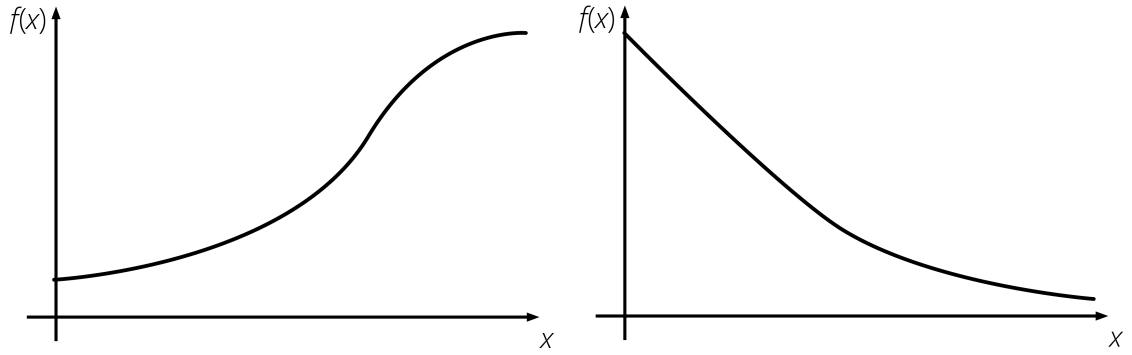


Figure 3.5: Examples of functions that are monotonically increasing (left) and monotonically decreasing (right).

decreasing semi-active constraint.

Using these principles and a set of concepts that can be used to identify the constraint that bounds a variable among a set of potentially active constraints (see chapt. 3 in Papalambros and Wilde [12]), single-objective optimisation models can be reduced systematically [58]. This also acts as a form of model verification; if an optimisation problem contains monotonically decreasing variable $f(x_i^-)$ where x_i is not bound from above by any constraint, then $x_i^* = \infty$. Correspondingly, the optimum will exist at 0 or $-\infty$ for monotonically increasing variables. Given that most design variables represent some form of geometric or physical properties in a design, such optima do, of course, not exist in reality. Furthermore, optimisation algorithms tend to fail to converge in these situations, resulting in a waste of computational resources. Thus, such models are said to be poorly bounded, which can be difficult to spot in most real optimisation problems prior to computation without performing monotonicity analysis. Typically, the optimisation models can thus be checked and reduced prior to computation by performing MA using so-called monotonicity tables.

MA is somewhat unique in the design optimisation field in that it is one of the few pre-optimality analysis methods. Some work has been done on problem partitioning and decomposition for the sake of computational efficiency and ease of model construction. Works include efforts to use the Design Structure Matrix [61], target cascading [62], and functional dependence trees [12], to support model construction. Yet, MA is the only method that reveals information that is specific to the optimal design. However, MA is largely reliant on manual, algebraic manipulations or implicit numerical reductions. This limits its application, as optimisation research today is more focused on allowing the solution of increasingly large and complex problems. Examples of this include works on multidisciplinary optimisation techniques [63], meta-modelling [64], non-hierarchical coordination algorithms [65], and techniques for dealing with coupling [66], and topology optimisation [67].

Work has been done to expand the applications of MA, allowing the analysis of more complicated problems in a simpler manner. Expansions include work on regional [68] and local [69] MA, work on automated [69, 70] and interactive MA [71]. Furthermore, Michelena and Agogina expanded MA to multiobjective problems [72], albeit without introducing a systematic procedure, while Gobbi et al. [73] later applied MA to allow the explicit derivation of the Pareto set for simple problems.

While pre-optimality analysis methods are relatively scarce, there is a substantial amount of post-optimality analysis methods. Broadly speaking, these aim to either help the analyst/designer understand the results even better or generate additional data to support decision

Table 3.3 Monotonicity table for model 2 (with model repairs in parentheses)

	Variables							
Functions	d_1	d_2	d_3	L_2	L_3	R	N_S	N_T
f_0	+	+	+	+	+			
g_1		−		+	+			
g_2	−					−		
g_3				−	−			
g_4				+	+			
g_5				+				
g_6				−				
g_7					+			
g_8					−			
g_9		+						
(g_{10})		−	+					
g_{11}	+	−						
g_{12}						+		−
g_{13}		U				+		−
g_{14}								+
g_{15}						−	−	
g_{16}							+	
(g_{17})			−					
(g_{18})		−	+					−

Figure 3.6: An example of a monotonicity table, by Papalambros [58]

making.

Once the Pareto set has been identified, there are numerous approaches for assessing the degree of competition between objectives. The often-used utopia point F^0 is a k dimensional point consisting of all the single-objective minima, $[\min(f_i(\mathbf{x})|i = 1..k, \mathbf{x} \in \chi]$, as illustrated in figure 3.4. By measuring the proximity of each individual Pareto-point to this Utopia point, for instance, Euclidean distance [74], compromise solutions can be identified. In a design with a high degree of competition, the Pareto points would be further away from the Utopia point. This is, however, only a measure that can be applied to individual Pareto points. Frischknecht and Papalambros [75] took this notion of the degree of competition and applied it to Pareto sets, describing the alignment of objectives using the effective curvature, area, and sensitivity of Pareto frontiers. This allows the comparison of alternative configurations.

Examples of research aiming at providing additional information post initial computation, includes the aspect if sensitivity analysis [76], and the assessment of robustness [77], uncertainty [78], reliability [79], and the effects of imprecise models upon decision making [80]. To inform decision making and aid the identification of a suitable solution amongst numerous alternatives in a Pareto set, efforts have also been made to allow the preferences of the decision-maker or user to be modelled [81–83]. Strategies for making trade-offs aggressively and conservatively [84], and identifying *efficient compromises* in the Pareto set [54], have also been developed.

Finally, approaches have been introduced to allow the comparison of the Pareto sets of different concepts, such as Mattson & Messac's S-curves [85], Athan and Papalambros' notion of the meta-Pareto set [86], and the numerous quantitative metrics to assess and measure

the quality of a Pareto set [74].

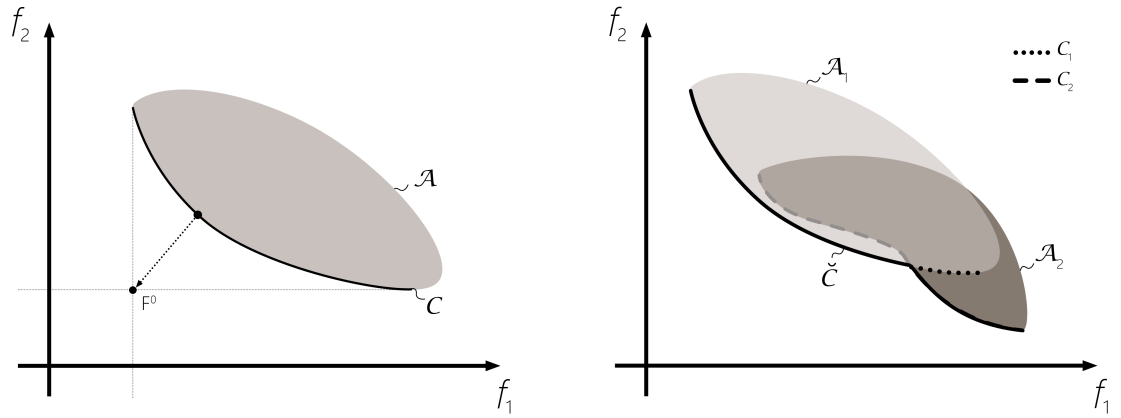


Figure 3.7: *Left:* The *Compromise solution* is the point in the Pareto set which has the smallest normalised Euclidian distance to the *Utopia point*. *Right*:An illustration of the meta Pareto concept introduced by Athan [55]. Here, $\breve{\mathcal{C}}$ is the set of Pareto optimal solutions for the union of the Pareto sets $\mathcal{C}_\infty$ and $\mathcal{C}_\in$, each representing an alternative "strategy" (e.g. concept, configuration, etc).

3.2.2 Design Space Exploration

Also known as Design by Shopping [87], design space exploration (DSE) is quite different from design optimisation. While it is not strictly an optimisation method, it can be used to identify an approximation of the Pareto set and is not affected by issues with convergence, handling discrete variables, or non-convex attainable sets.

Essentially, DSE is a *brute force approach,* which involves sampling a large number of designs and subsequently evaluating whether they are feasible, and if so, how they perform. There are various implementations with different benefits, but using the notation form equations 3.5-3.8, the general approach is as follows:

Eliminate $\mathbf{h}(\mathbf{x})$, by algebraically solving for one of their dependent variables and back-substituting
Sample the values of $\mathbf{x}$ between predefined limits, $\mathbf{x_L}$ and $\mathbf{x_U}$.
Store these in $\mathbf{X}$, a $[n, i_{iter}]$-dimensional matrix
for i = 1..i_{iter} **do**
  $\mathbf{x} = \mathbf{X}(i, :)$
  Evaluate whether $\mathbf{x}$ fulfils the constraints, i.e. whether $\mathbf{g}(\mathbf{x}) \leq \epsilon$.
  IF not, then end the iteration and proceed to the next one.
  Evaluate $\mathbf{f}(\mathbf{x})$
  Store the outputs: $\mathbf{F}(i, :) = \mathbf{f}(\mathbf{x})$, and $\mathbf{G}(i, :) = \mathbf{g}(\mathbf{x})$.
end for

Here, *n* is the number of design variables, i_{iter} is the number of samples, ϵ is the allowable residual error for the constraint functions (typically $\epsilon \leq 10^{-6}$). The equality constraints are eliminated prior to sampling, as this avoids the generation of a large number of designs that will never fulfil an equality constraint. Depending on the purpose of the DSE, one can use different distributions to sample $\mathbf{x}$. If one, for instance, wishes to identify the attainable set, $\mathcal{A}$, a quasi-random uniformly distributed set as suggested by Athan [86], would be suitable.

Depending on the nature of the problem, DSE can either be performed iteratively, as illustrated above or through a vectorised approach. The vectorised approach will, in most cases,

be much more computationally efficient. With a wide enough sampling range and a high number of samples, the whole attainable set can be approximately identified. Thus, we can apply a non-dominance filter to $\mathbf{F}(i,:) = \mathbf{f}(\mathbf{x})$, e.g. as the one suggested by Mattson et al [88]. Such filters evaluate each point in a data set to assess whether they meet the condition for Pareto optimality shown in equation 3.14. Thus, DSE allows the identification of solutions that are potentially Pareto optimal; whether or not depends on whether $\mathcal{A}$ has been exhaustively identified or not. The combination of DSE and non-dominance filters is commonly referred to as *Trade-space Exploration* [89, 90]

The benefit of DSE is first and foremost that it allows the identification of the attainable set and the feasible domain of the design variables. Thus, it can be used for many other purposes besides identifying *good* solutions - one can, for instance, use it to assess how restrictive the constraints are or to explore how the objectives interact without necessarily considering optimality. DSE does not suffer from convergence issues, and it is completely insensitive to discontinuous design spaces, non-convexity, and discrete variables. Thus, it also allows the comparison of alternative configurations and concepts, so long as the mathematical models to describe each of them have been constructed. This simplicity avoids many of the methodological barriers that optimisation may pose, meaning designers without much or any optimisation experience can use DSE and still gain valuable insights.

The main strengths of DSE also drive its key weakness. In problems with a large number of variables and constraints, or in problems with very small and discontinuous attainable sets, a substantial amount of samples is needed to approximate the attainable set. In large problems, this might result in a higher computational cost than optimization would. Furthermore, much of the information gained through optimization, such as Lagrange multipliers and more detailed constraint activity data, is not gained through DSE. Thus, one will at times need to rely on data visualisation to reach some of the insights that are otherwise reached through design optimisation. Thus, a lot of the research within DSE revolves around dimensional reduction [91], multi-objective visualisation technique [78, 87, 89, 91, 92], and post-computation indicators [22].

3.3 Dependency Analysis

As discussed in chapter 1, modern products are becoming increasingly integrated and multidisciplinary [30], leaving designers with increasingly complicated dependencies and corresponding trade-offs. As such, the notion of interdependencies existing between the different goals involved through the design process is an important one.

Numerous qualitative methodological frameworks exist that concern themselves with different aspects related to dependencies in design. Some, such as the Design Structure Matrix (DSM) [7] and the quality function deployment (QFD) [93] are purely focused on providing an approach to mapping out dependencies, which have especially found acceptance in the study and design of complex or multidisciplinary systems, e.g. mechantronics[94]. Others, such as Axiomatic Design [13] and TRIZ, make prescriptive rules for "good" design that are related to some form of dependency avoidance. To facilitate this, both also provide their own approaches for the identification of dependencies. Related to this, the Contradiction Index approach developed by Göhler et al. [21] attempts to integrate the perspectives and prescriptions from Axiomatic Design, TRIZ, and DSM into one, to allow the specific identification of contradictory dependencies and their relationship with the complexity of the system.

As most of these methods have not been used to reach the results in this thesis, we will, for the sake of brevity, not go into detail by describing each of these methodologies and exploring their differences. They are merely included here as they present the only existing

alternatives to the methods developed later in this thesis. As will be covered in more detail in the next chapter, these existing dependency analysis techniques are not suitable for the purposes of this research. This primarily comes down to the fact that they do not necessarily account for the influence of active and inactive constraints. In fact, they generally do not distinguish between objectives and constraints at all. Axiomatic design, for instance, bundles them under so-called *Functional Requirements* (FR). DSM, meanwhile, studies designs at a certain level of abstraction (e.g. components or sub-assemblies) [95], meaning the dependencies identified can be of any form and not necessarily to dependencies between design objectives.

As a result, the designer's knowledge determines which insights are reached if these dependency analysis methods are used in the effort to identify trade-offs and the dependencies that cause them. If the dependencies in a system are assessed using a reference design (i.e. an embodied and dimensioned system, e.g. in CAD), then the dependencies identified between design objectives would essentially be *local* to a specific point in the attainable set. Alternatively, the designer might be aware of certain relationships that exist between different criteria through shared design variables, irrespective of constraint activity. This essentially involves the identification of dependencies that exist *globally* in the objective space. Yet, using existing dependency analysis methods to understand the dependencies that exist at the optimum - which is surely of most interest in the context of embodiment - the designer would need to know a priori which constraints are active.

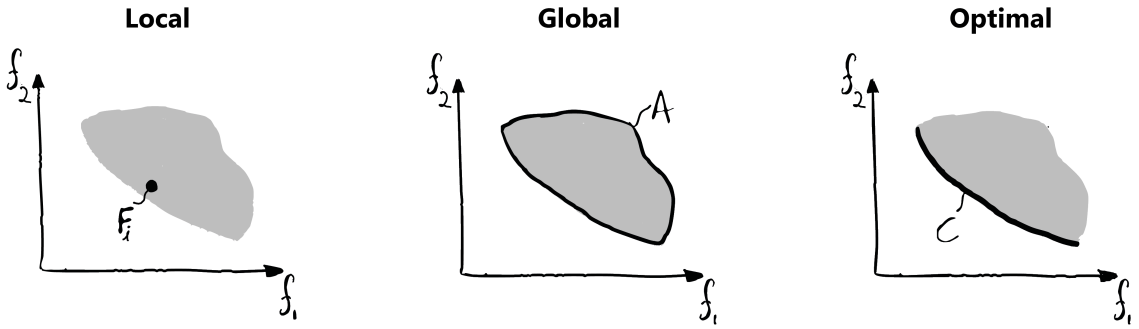


Figure 3.8: In qualitative dependency analysis, the knowledge of the designer determines whether the dependencies are local in the objective space (*left)*, exist globally in the objective space (*center*), or are specific to the Pareto Optimal set.

3.4 Design Methods

In this section, an overview is given of existing qualitatively and quantitatively founded design methods aimed at embodiment design or in some way involving achieving a "good" solution. The techniques discussed until now rely on the designer actually synthesising a concept and an embodiment and then analysing or optimising it. When it comes to changing the design, however, one moves beyond the description of the design problem defined in the design optimisation model or the dependency analysis. For the same reason, most design methods aimed at mechanical design during the embodiment stage are heuristic in nature, either focusing on avoiding common errors made by the designer, on a specific sub-discipline of design, or are more broadly oriented towards identifying/synthesising a "good design".

3.4.1 Qualitative and Heuristic Design

Historically, engineering design research has focused on developing a wide range of tools and methods aimed at supporting designers in achieving certain desirable characteristics in the end product. These are often presented in the form of collections of heuristics - i.e. design

principles, guidelines, rules, tools, and *good* practices. For examples of such, see works by e.g. Pahl & Beitz [6], Skagoon [96], Altshuller [16], and French [14, 97, 98]. In particular, Pahl & Beitz and French have published a wide array of heuristic design methods aimed at the embodiment design stage. Of note in that regard are the oft-cited *The Principles of Embodiment Design* [6].

Over time, these developed into collections of heuristics within different groups of desirable characteristics related to different sub-disciplines of design or adjacent fields. Examples include Design for Manufacture, Design for Reliability, Design for Robustness, Design for Sustainability, Design for Recycling. These characteristics often referred to as "*ilities*"[99, 100], and are referred to in unison as *Design for "X"* (DfX). Yet, trade-offs between "X's" will occur, and the question of how to design toward achieving as many of these goals as possible, and balancing them systematically, is less well described.

As mentioned in the previous section, some methodologies that involve some mode of dependency analysis also involve prescriptive rules for "good" design. If note here is TRIZ (the Theory of Inventive Problem Solving), which was originally developed by Altshuller [16] through the study of commonalities in patents. TRIZ ultimately concerns itself with the avoidance of *contradictions* between generic design objectives in inventive design, with Altshuller arguing that all novel inventions solve these contradictions through a finite amount of principles. Hence, TRIZ consists of an approach to identifying these contradictions, and a set of inventive principles that may be applied, to avoid these contractions. One could argue that TRIZ is essentially a heuristic approach to trade-off identification and avoidance, based on a large data-set of patents, targeted towards the synthesis of new concepts. Yet, TRIZ has a relatively limited uptake in practice and academia, largely due to a perceived degree of "mysticism" surrounding it [101].

Another prescriptive design methodology of note here is Suh's Axiomatic Design [13]. In short, Suh put forward a general theory for systems design that prescribes two axioms for achieving the *best* design. Axiomatic design decomposes all design problems into a set of *functional requirements* (FR) describing the goals a system is designed towards and a set of design parameters (DP) describing the properties of the system that the designer can manipulate. The so-called *design matrix* maps what DPs each FR is dependent on, with the first axiom essentially stating that all FRs should ideally be uncoupled (share no DPs), and barring that decoupled (share DPs but have enough independent DPs to adjust the value of each FR). Meanwhile, the second axiom states that the best design has a minimal amount of DPs. Put into the context of trade-off avoidance, an uncoupled design would indeed be free of trade-offs, with axiomatic Design essentially suggesting one should design toward achieving no interdependence and no complexity.

Nevertheless, Axiomatic design is not without its limitations and criticisms. Viewed from a multiobjective optimization perspective, dependencies do not always detrimentally affect the location of the optimum. As argued by Gohler et al. [21], dependencies can actually contribute to lowering complexity. Ultimately, the number of design variables in a system is correlated with the number of parts and sub-systems, meaning one can to an extent relate them to cost, as also suggested by Suh [13], meaning that the two axioms are inherently conflicting. Similarly, Frey et al. [26] found that there is no direct correlation between the degree of dependency and the achievable performance in the end product when studying part-reduction efforts in the design of jet engines. Of equal importance is that design variables (which Axiomatic design calls DPs) themselves are not always independent from each other - the existence of equality constraints and active inequality constraints can determine their relative values. To an extent, Axiomatic Design can capture this, but only if all active

constraints are known and included in the set of FRs. If an inactive constraint is included as a *Functional requirement*, then one would find "couplings" (the axiomatic design term for dependency) that have no bearing on the optimum. Correspondingly, if an active constraint is left out, then couplings are overlooked.

3.4.2 Quantitatively Driven Design - Adapted from Paper B

It is relatively well accepted that the decisions made in conceptualisation and embodiment design of a system largely determine the achievable optimal result [12, 31, 32, 102]. Yet, engineers are not taught *how* to design for a good optimum, avoiding trade-offs and overly restrictive constraints. As mentioned, current optimisation methods provide little systematic support in identifying when and how to improve the problem formulation by changing the embodiment design or conceptual, as they mostly deal with proportional or parametric changes. The closest approximation of this is TRiZ [16].

The fundamental challenge is the lack of appropriate mathematical modelling capabilities: different configurations require different mathematical problem formulations. Without this ability, it is challenging to explore alternative configuration designs simultaneously. In mechanical design, one of many reasons for this is that the underlying analysis models often rely on constraints and boundary conditions without which the governing equations cannot be solved. For instance, in changing the configuration of components in a system, said constraints and boundary conditions inevitably change. Existing methods, therefore, fall into one of two strategies:

Firstly, one can use design optimization techniques to identify the optimal proportions of an already embodied design and reach a better understanding of the design problem and use this to compare alternative configurations. Yet, this has no direct impact on the configuration of parts in the system or their topology unless when the designer is able to use the optimization results to extract insights that guide the identification of configuration design improvements or entirely new concepts. This requires that the designer understands the problem and the results well enough to know *what* to change in a configuration or in the overall concept in order to achieve improvement, and *how* to change it. Work has been done, primarily in the 1980s and 1990s, to leverage monotonicity analysis with such goals in mind. Examples include Cagan & Agogino's work on redesigning multistage gearboxes using MA [32], Jain & Agogino's work on an optimisation based theory of design [102], and Ishii and Barkans work on building expert systems that use MA to reveal the properties of the optimum that create "bottlenecks in early design" [103]. Furthermore, Deb [104] used the same rationales similar to MA to identify *innovations through optimisation*, which seeks to identify common patterns in Pareto optimal designs that may lead the designer to make more informed decisions.

Secondly, one could attempt to optimize a functional representation without the actual embodiment of the functions, the output of which would inform configuration or conceptual design. The book edited by Antonsson and Cagan [11] provides an old but excellent overview of formal techniques for the synthesis of configuration designs. Further, the review by Chakrabarti et al. [105] provides an overview of computational synthesis methods that lie more within the conceptual design domain.

There are a couple of notable achievements in this direction. Some limited success has been achieved through combinatorial methods such as grammars[105], optimal configuration design [106], and network [107] and decomposition-based methods [108] for automated configuration synthesis and optimisation. What these methods have in common is that they on some form of algorithm or combinatorial approach to synthesise alternative configurations from a predefined set of elements.

Furthermore, structural topology optimization (TO) [67] methods are a broad and quite successful research field, primarily as they actually involve a unified mathematical model across configurations. TO essentially involves a bi-objective optimization problem, minimising the compliance and mass of a loaded structure subject to material failure constraints. The use of a tensor field allows the optimal allocation of material based on a predetermined physical criterion, allowing applications beyond structural design (e.g. heat transfer, electrical resistance, optical properties). Somewhat uniquely, TO optimizes a functional representation of the design without the actual embodiment of the functions, and the results inform configuration design. For the same reason, the TO field has seen a lot of research attention since the seminal paper by Bendsø & Kikuchi [67].

Despite these achievements, the ultimate goal of broadly applicable methods for generative design, and actual optimal configuration design, remains somewhat elusive. Topology optimization approaches have limitations in the context of early iterations in configuration design, as they rely on a predefined set of boundary conditions and loads for numerical solutions. While decomposition methods allow TO of assemblies [109], the need for well-defined boundary conditions and loads implies a static configuration of components. Further, TO deals with configuration design problems where the physics change from configuration to configuration. Combinatorial optimization techniques, meanwhile, can only capture configurations accounted for in the model. Thus, they are limited by the capability of the designer to include all the relevant system elements and build a sufficiently flexible model to allow the optimisation of any and all permutations that are synthesised. That said, both TO and combinatorial methods have found considerable success in lightweight design and automated morphological studies, respectively.

3.5 Constructing a Optimization Model for the SOMA Device

The following is an adaptation of the case related content in Paper A, albeit with a much greater level of detail than a journal paper permits.

As discussed in chapter 1, the design of the SOMA device presents an interesting configuration design task, given the substantial envisioned production volume, the safety-intensive application, and the numerous trade-offs involved. Thus, it was used as a test case in the method development described in the coming chapters. To do so required the construction of a multiobjective optimization model, which had two purposes;

1. To identify and quantify the key trade-offs involved in the early design of the SOMA device, the root causes of which would subsequently be studied.
2. To provide the SOMA project with data on optimal solutions, constraint activity and screen for previously unidentified design issues

In order to build such a model, data was first collected from the SOMA project. This served to build an understanding of the overall design problem to inform model construction and provided the parametric values to be used in numerical solution. The data types and collection methods took several forms:

- *CAD data* - Archival data from the company's PLM system, covering each major design iteration (e.g. change in part structure, addition of new functionality, or the like) was collected and studied.
- *Design History* - Based on an initial analysis of this archival data, unstructured interviews were conducted with the design engineers involved to gain an understanding of the reasoning behind each of the major design iterations since the initiation of the project.

This especially focused on what improvements they had been attempting to make, what issues they had been trying to eliminate, and the data or information that was used to qualify decision making in this process. At the time of this work, the incumbent design of the SOMA device was as shown in chapter 1.

- *Test data from explorative testing* - Based on these interviews, the results from some of the explorative testing done by the project on prototypes was reviewed. This helped build an understanding of the physics of the SOMA and what phenomena one would need to model in order to build a valid optimization model.
- *Material data* - Subsequently, the properties of the materials used in the device were identified. This either came from supplier data sheets or from internal work done on material characterisation. Proprieties such as density, Young's- and shear moduli, yield properties, and coefficients of friction were of special importance.
- *Manufacturing process capability* - Finally, the expected achievable tolerance grades, the requirements for gripping and orientation surfaces for handling in assembly, and shape requirements were provided by DfMA and process simulation engineers.

Due to the IP-sensitive nature of most of this data, it is not included explicitly in this thesis. After this process, the key design objectives for the optimization model were selected in collaboration with the technical project manager involved in the SOMA project. This selection was primarily based on the challenges faced by the project but also upon an overall goal to model as much of the in-use behaviour of the device as possible. As illustrated in figure 3.9, the SOMA device goes through different functional states; it is handled and swallowed by the user, after which it passes into the stomach where it self-orients. Subsequently, the dosing mechanism is triggered by the dissolution of the plug, injecting the API needle into the submucosa, where it dissolves. Finally, the SOMA passes onward through the digestive system.

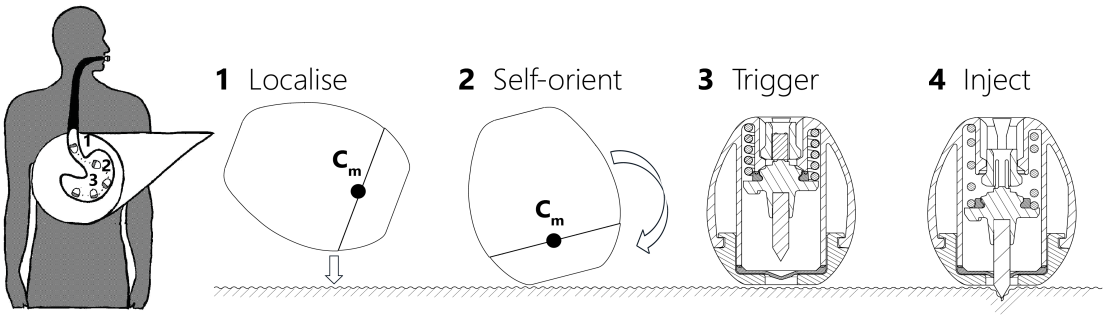


Figure 3.9: An overview of the functional states of the SOMA device. In part adapted from Abramson et al [41]

The *performance* of the device, in each of these states, can be simplified down to four objectives. In order for the device to be swallowed and pass safely, it needs to be small, meaning device size needs to be minimized. To reliably deliver API, it needs to self orient well and inject the needle far enough into the stomach lining, meaning self-orientation and injection depth need to be maximized. To have a therapeutic impact, it needs to deliver a sufficiently large amount of API, meaning the mass of the needle needs to be maximised. Testing activities had at the time already revealed that the triggering and sealing functionality (which keeps the API dry until injected) could be achieved without issue, meaning these were not included as optimization objectives. Given the early stage of development, model construction focused on developing the simplest *meaningful* model; i.e. one which would capture the trade-offs

involved and the constraints and variables that contribute to them, without necessarily focusing on capturing all non-linearities (e.g. using finite element based models to compute the stress in the components).

In the following, the construction of the optimization model is described. The non-reduced model consists of 4 objectives, 45 design variables, 52 inequality constraints, 7 equality constraints, and multiple parameters. Hence, the model construction is not covered exhaustively but rather to the detail required for the comprehension of the subsequent chapters and to illustrate the overall model fidelity. First, the decomposition of the SOMA device into meaningful variables and parameters is described, followed by the derivation of the design objectives and an overview of the classes of constraints involved in the model.

3.5.1 Design Variables and Parameters

As mentioned in chapter 1, the SOMA and its invention was first described in a paper by Abramson et al. [41]. Cardinal to the concept is the self-orienting functionality of the device, which is achieved through a mono-monostatic body inspired by the shape of leopard tortoises (who are able to self-right when lying on their backs). The outer shape of the device, which allows this functionality, was derived by the project team using an optimization model that minimized self-orientation time[41] while maximizing stability after orientation. From very early on, it was hence decided to represent the outer shape of the device through equality constraints, as the outer shape of the device was already optimal. Instead, the optimization model focused on varying the relative sizing of the internal components and how the external components contribute to the outer shape.

As a result, the device was decomposed into a set of design variables, allowing the optimization model to adjust the relative sizing of the components without affecting the overall outer shape. The variables most important variables (the ones involved in the subsequent analysis) are illustrated in figure 3.10. These represent the dimensions of the internal components and the dimensions of the external components (top and base housing) that do not relate to shape. This decomposition relies heavily on the (mostly) rotational symmetry of the SOMA device, meaning that most of the design can be represented through diameters, denoted d, and axial lengths, denoted l. Widths and overlaps that are not rotationally symmetric are denoted w and δ, respectively. Furthermore, a few of the design variables denote movements and clearances in an axial direction, and these are denoted as z (e.g. the spring pretension, z_{pre}). The subscripts on most variables denote which part they are attributed to. When the design variables are referred to in unison, the more common notation **x**, denoting the vector of design variables, will be used. The use of separate notation for lengths, diameters, and so on is used for the sake of clarity, allowing the analysis to be related to the design.

Thee dimensions on the top and base housing, relate to their internal geometry, the overall sizing of the device (l_{t1}, d_{t1}, d_{b2}), and the dimensions that determine the split line between the two housings ($l_{t2}, l_{b1}, l_{b2}, d_{b5}$). The outer shape of the device is determined by the equality constraints and a set of shape parameters (e.g. the height-width ratio, C_t). Further design parameters include material properties, manufacturing limits such as radial clearances, wall thicknesses, and radial overlaps in assembly interfaces that are not load-bearing.

3.5.2 Modelling of Objectives

Self orientation - Calculating the System Center of Mass

Given that this optimization study does not involve the outer shape of the device, we can model the self-orientation performance by simply optimizing the location of the centre of mass. The lower the centre of mass, the quicker the self-orientation. Due to the (roughly) rotationally symmetric design, the centre of mass is positioned in the centre axis going through the length of the device.

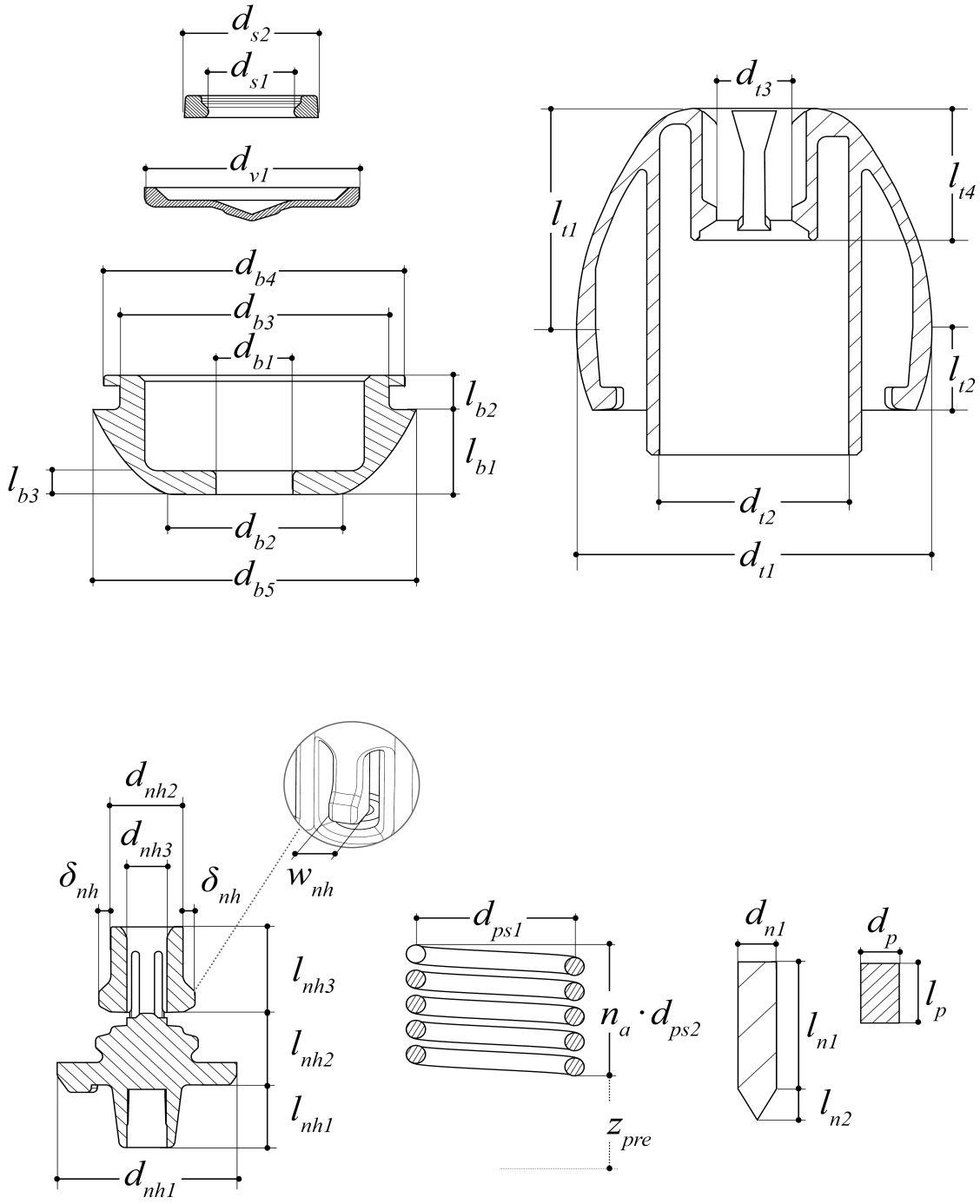


Figure 3.10: Design variables of relevance to the constraints used to demonstrate the trade-off root cause analysis. *l* denotes length variables, *d* -diameters, δ -overlaps, and z denotes vertical positions. Adapted from Paper A

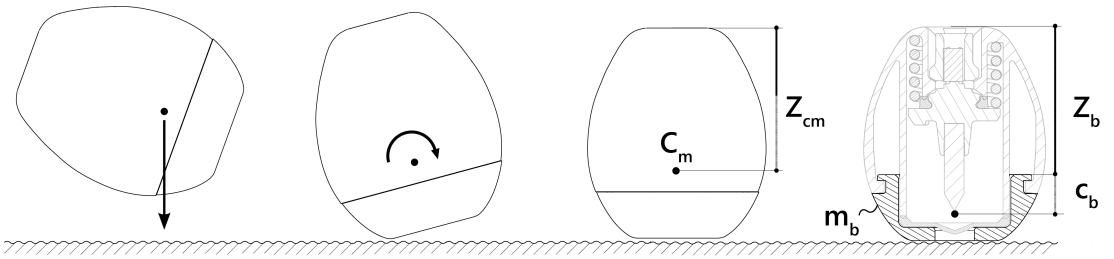


Figure 3.11: Objective 1: The self-orientation performance of the SOMA device is modelled using the vertical position of the centre of mass relative to the top of the device. This is done by calculating the mass and centre of mass of each component (as exemplified for the base component on the right) and using this to calculate the position of the system centre of mass.

As such, it is sufficient to develop a model describing how close the centre of mass is positioned to the bottom of the device. To derive an analytical expression of this, the SOMA device was viewed as a *system of particles*, with each particle representing a part in the SOMA assembly with its own mass and position on the global coordinate system. Correspondingly, each component was decomposed into a set of idealised geometric elements (e.g. ellipsoids, cylinders, beams) while accounting for features such as rounds, draft, and snap interfaces.

This allowed the calculation of the volume, mass, and centre of mass of each part, which could then be used to calculate the position of the centre of mass of the system as a whole. This position is measured relative to the top of the device, meaning that the self-orientation is optimized by maximising the distance between the top and the centre of mass. This was done to preserve the monotonicity of the objective function. If we were to measure the centre of mass relative to the middle of the device, or the bottom, certain design variables, such as the height (l_{b1}) and thickness (l_{b3}) of the base housing would have become non-monotonic.

This resulting objective function is:

$$Z_{cm} = -\frac{\sum_{p=1}^{p=8} m_p \cdot (C_p + Z_p)}{(l_{t1} + l_{t2} + l_{b1}) \cdot \sum_{p=1}^{p=8} m_p} \tag{3.24}$$

where Z_{cm} is the distance between the top of the device and the system centre of mass, C_m, relative to the total height of the device, $l_{t1} + l_{t2} + l_{b1}$. By normalising Z_{cm} relative to the device height, avoids the optimal result always involving the largest possible device, as this would maximise the distance Z_{cm}.

In this expression, m_p, C_p, and Z_p are the intermediate functions, with m_p describing the mass of each part in the device, C_p the centre of mass in each part, and Z_p the axial distance of each part from the top of the device. This is illustrated using the base housing as an example, in figure 3.11. The objective function was then verified against CAD models of the SOMA in several sizes, with a max. observed deviation of 0.83% compared to the mass distribution analysis done in the CAD system (PTC Creo 4.0). This was deemed to be acceptable for the purposes of the optimization study.

Device Size

The device size is of importance to the design of the SOMA device, as it needs to be swallow-

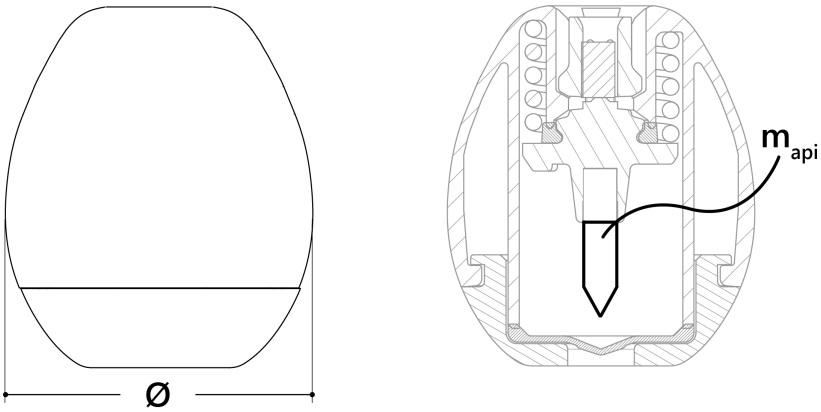


Figure 3.12: Objectives 2 and 3 - Device Diameter and API Payload

able without risk to the user. The risk of complications is generally proportional to the minor diameter of pills and capsules [110]. The US-FDA, therefore, recommends that pills and capsules generally stay below a standard 00-size [111], which has a minor diameter of 8.35mm, while the largest standard size, 000 capsules, are 9.91mm in minor diameter. Complications from swallowing pills and capsules start to arise at a size above 8mm in minor diameter. These increase substantially beyond a minor diameter of 11mm [110]. Hence, it is sufficient to minimize the outer diameter of the top housing, d_{t1}, to optimize swallowability. Hence the optimum of this objective will be defined by the constraint(s) that determine the lower bound of d_{t1}. How this is handled will be shown in chapter 4.

API Payload

The SOMA device can be seen as a platform product in that it could, in principle, deliver any protein-based compound which cannot otherwise be delivered orally. Yet, the size of the therapeutic dose of different treatments varies wildly, with some compounds being delivered as less than a milligram, while others are counted in the tens or even hundreds more. Hence, the larger the payload of API contained in a swallowable SOMA device, the larger the variety of pharmaceutical compounds might be delivered using it. Hence, a maximisation objective describing the mass of deliverable API in the needle was modelled:

$$M_{api} = \rho_{api} \frac{\pi}{4} d_{n1}^2 \left(l_{n1} + \frac{1}{3} \cdot l_{n2} \right) \tag{3.25}$$

This is a simply volumetric expression, where the needle is viewed as a composite geometry, consisting of a cylindrical main body with a cone-shaped tip. The density parameter, ρ_{api}, describes the amount of active pharmaceutical ingredient pr. unit of needle volume.

Injection Depth - Calculating the Impact Velocity

The injection depth of the SOMA determines how much of the delivered API reaches systemic circulation. This depth is ultimately determined by the mechanical properties of gastric tissue, the impulse with which the needle impacts gastric tissue, and the sharpness of its tip. The needle is made from compacted protein, meaning that the sharper the tip, the more costly and sensitive the production process. Hence it is preferable to achieve a sufficient depth with a large impulse and not rely on sharpness.

The impulse, however, is affected by the API payload. As we wish to study the relationship between API payload, and injection performance, we need to isolate the two. Furthermore,

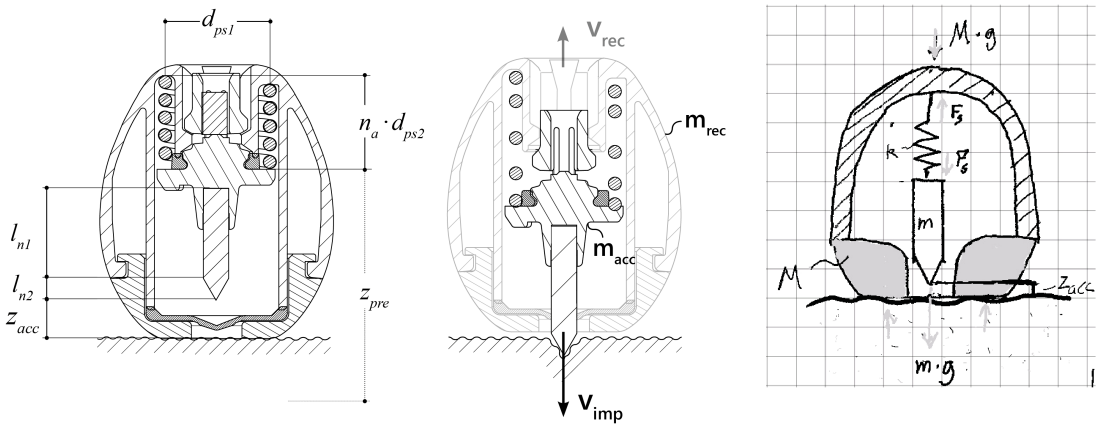


Figure 3.13: The impact velocity between needle tip and tissue is used to model the injection depth performance of the SOMA device. This velocity is dependent on the design of the spring, the initial distance between needle tip and tissue (z_{acc}), the accelerated mass m_{acc} (needle, hub, and spring) , and the mass of the mass

the mechanical properties of the different layers of the stomach (illustrated in Fig. 1.7) are not well described, given that living tissue has highly variable characteristics over time and from person to person. Its properties are also different depending on whether it is tested in-vivo or ex-vivo. Hence, it is challenging to model the change in the linear momentum that occurs at impact, given that the damping and stiffness of the stomach tissue are unknown.

As a result, we instead describe the injection depth performance through the impact velocity, rather than the impulse, modelled as a maximizing objective since the injection depth increases monotonically with velocity. Thus, we seek an expression for the velocity of the needle tip at the point of impact with the stomach. Here, we can rely on Newtonian mechanics, as we are studying the motion of a body influenced by a spring force and gravity. In the following, the air and hydraulic resistance overcome by the actuator and the frictional loss of that occurs then of the valve component are disregarded for the sake of simplicity. These contributions had already been found to be negligible in tests done by the SOMA project on prototypes, and omitting them greatly simplifies the analysis of monotonicity performed in chapter 4.

As such, we seek the velocity $V_{imp} = a \cdot t$, at the point in time, t, where the needle impacts the gastric tissue. As shown in the system sketch in figure 3.13, the spring accelerates the mass of the hub and needle towards the tissue (m_{acc}), and the rest of the device m_{rec} away from it, when the device is triggered. This results in a small recoil effect, where the device "jump" once triggered. Given that $m_{acc} << m_{rec}$, the effect of this recoil is minimal but will be accounted for through a slight correction of the kinematic equations.

First, we look at the movement of the needle. We know that the acceleration a is determined by two contributions; gravity and spring force:

$$F_s(z_n) + g_C \; m_{acc} = m_{acc} \, a \tag{3.26}$$

$$\Leftrightarrow a = \frac{F_s(z_n)}{m_{acc}} + g_C \tag{3.27}$$

$$\tag{3.28}$$

where g_c is the gravitational acceleration, m_{acc} is the mass being accelerated by the actuator, and $F_s(z_n)$ is the spring force as a function of travel, z_n. Given that we know the travel

distance before impact, which is equivalent to z_{acc}, we can solve for *t*:

$$z_{acc} = \frac{1}{2}a \cdot t^2 \Leftrightarrow t = \sqrt{2\frac{z_{acc}}{a}} \tag{3.29}$$

$$\Leftrightarrow V_{imp} = a\sqrt{2\frac{z_{acc}}{a}} = \sqrt{2az_{acc}} = \sqrt{2(g + F_s(z_n)/m_{acc})z_{acc}} \tag{3.30}$$

The spring force is not constant; it decreases from its initial state, $F(0) = z_{pre} \cdot k_{spring}$ to its state at the point of impact, $F(z_{acc}) = (z_{pre} - z_{acc}) \cdot k_{spring}$. Under the assumption that the SOMA is at rest when it is triggered, and disregarding hysteresis in the spring, we use the conservation of energy set up an integral, describing the velocity as a function of position, z_n:

$$E_k = \int_0^{z_{acc}} \sum F(z_n) = \int_0^{z_{acc}} (g + F_s(z_n)/m_{acc})m_{acc} = \frac{1}{2}m_{acc}V_{imp}^2 \tag{3.31}$$

$$\Leftrightarrow \frac{1}{2}V_{imp}^2 = \int_0^{z_{acc}} g_c + (z_{pre} - z_n)\frac{G_{st}d_{ps2}^4}{8d_{ps1}^3 n_a}\,\mathrm{dz_n} \tag{3.32}$$

$$= \int_0^{z_{acc}} g_C \mathrm{dz_n} + \int_0^{z_{acc}} z_{pre}\frac{G_{st}d_{ps2}^4}{8d_{ps1}^3 n_a}\mathrm{dz_n} - \int_0^{z_{acc}} z_n\frac{G_{st}d_{ps2}^4}{8d_{ps1}^3 n_a}\mathrm{dz_n} \tag{3.33}$$

$$= g_C z_{acc} + z_{acc}z_{pre}\frac{G_{st}d_{ps2}^4}{8d_{ps1}^3 n_a} - \frac{1}{2}z_{acc}^2\frac{G_{st}d_{ps2}^4}{8d_{ps1}^3 n_a} \tag{3.34}$$

$$\Rightarrow V_{imp} = \sqrt{2z_{acc}\left(g_C + \frac{z_{pre}G_{st}d_{ps2}^4}{8d_{ps1}^3 n_a m_{acc}} - \frac{z_{acc}G_{st}d_{ps2}^4}{16d_{ps1}^3 n_a m_{acc}}\right)} \tag{3.35}$$

In the above, we have inserted the expression for the stiffness of compression springs [112], where G_{st} is the shear modulus of the spring steel, while the variables in question are illustrated in figure3.13. Yet, there are still two contributions we need to derive in order to have a sufficiently accurate expression. Firstly, we need to account for the mass of the spring, when considering the acceleration of the needle, given that the mass of the spring, m_{ps} is comparatively large when related to the mass of the needle, m_n, and hub, m_{nh}. We can, in other words, not neglect the effect of the spring also accelerating itself. Given that $m_{acc} << m_{rec}$, we can simplify this question by viewing the needle end of the spring as being in motion while the opposite end is stationary. In this scenario, we know that the velocity, u, of each coil of the spring is proportional to its distance from the stationary end, i.e. $u = \frac{z}{L_{ps}}V$. Thus, we can identify the effective mass of the spring by looking at the kinetic energy of each winding:

$$E_{k_{ps}} = \frac{1m_{ps}}{2L_{ps}}u^2 = \frac{1m_{ps}}{2L_{ps}}\int_0^{L_{ps}}(\frac{z}{L_{ps}}V)^2\mathrm{dz} \tag{3.36}$$

$$= \frac{1m_{ps}}{2L_{ps}^3}V^2\int_0^{L_{ps}} z^2\mathrm{dz} \tag{3.37}$$

$$= \frac{1m_{ps}}{2L_{ps}^3}V^2\left[z^3/3\right]_0^{L_{ps}} \tag{3.38}$$

$$= \frac{1}{2}\frac{m_{ps}}{3}V^2 \tag{3.39}$$

From the above, we can infer that see that one-third of the spring's mass is accelerated along with the needle and hub. Hence, we can conclude that:

$$m_{acc} = \frac{1}{3}m_{ps} + m_n + m_{nh} \tag{3.40}$$

Finally, there is the question of recoil. Given that the rest of the device, the mass of which is m_{rec}, is accelerated in the opposite direction, some of the spring energy is lost to accelerating the device upward. This acceleration results in the reduction of the spring force exerted onto the needle system, as spring pretension is lost as the device moves upwards. Thus, we seek to express how high the device is pushed upward, at the point of impact between needle and tissue:

$$m_{rec}\, g_c - F_s(z_n) = m_{rec}a_{rec} \Leftrightarrow a_{rec} = g_c - \frac{F_s(z_n)}{m_{rec}} \tag{3.41}$$

$$\Rightarrow \frac{V_{rec}}{V_{imp}} = \frac{g_c - \frac{F_s(z_n)}{m_{rec}}}{g_c + \frac{F_s(z_n)}{m_{acc}}} \tag{3.42}$$

Here, we consider that the mass of the SOMA device is very small - not more than a few grams. Meanwhile, the force exerted by the spring is comparatively large - at the time of the study, the SOMA device was embodied with a spring exerting an initial load of 14 N. Thus, $g_c \cdot m_{rec} << F_s(z_n)$, meaning we can we can assume that:

$$\frac{V_{rec}}{V_{imp}} \approx \frac{\frac{-F_s(z_n)}{m_{rec}}}{\frac{F_s(z_n)}{m_{acc}}} \tag{3.43}$$

$$\Leftrightarrow V_{rec} = -V_{imp}\frac{m_{acc}}{m_{rec}} \tag{3.44}$$

As the upward velocity of the rest of the device is proportional to the impact velocity and the relative difference in mass, the recoil distance, z_{rec} will also be defined by this relationship, meaning $z_{rec} = z_{acc}\frac{m_{acc}}{m_{rec}}$. This means, that the spring force is further reduced in the state at which the needle impacts the tissue: $F(z_{acc}) = (z_{pre-z_{acc}}(1 + \frac{m_{acc}}{m_{rec}}))k_{spring}$. Adding this to the impact velocity term yields the final, unreduced form of the impact velocity objective:

$$V_{imp} = \sqrt{2z_{acc}\left(g_C + \frac{z_{pre}G_{st}d_{ps2}^4}{8d_{ps1}^3 n_a m_{acc}} - \frac{z_{acc}\left(1 + \frac{m_{acc}}{m_{rec}}\right)G_{st}d_{ps2}^4}{16d_{ps1}^3 n_a m_{acc}}\right)} \tag{3.45}$$

3.5.3 Identification and Modelling of Constraints

As mentioned, the full model involves 59 constraints. While the constraints of relevance to the monotonicity analysis are shown in chapter 4, the full set of constraints in the model

will not be included in this thesis. Many of these involve a quite extensive derivation effort (over 60 pages of handwritten algebraic derivations), meaning that including them in detail would distract from the actual research contributions. All the constraints in the model involve the use of well known algebraic expressions from geometry, structural mechanics, machine element theory, along with common manufacturing considerations. Hence, a brief overview is given of the different types of constraints involved in the model will be given instead. Beyond the desire for brevity, this is also driven by the challenge that the model contains quite a lot of proprietary data (material selection, process capability and design, data from experiments), which could unfortunately not be included in the thesis.

Inequality Constraints

The inequality constraints, $\mathbf{g}(\mathbf{x})$, represent a broad range of conditions relating to the geometric relations in the SOMA, its manufacture, and avoidance of structural failure (yield, creep, and buckling).

Radial constraints - The set of radial inequality constraints ensure that the parts actually fit together in the radial direction and that all the radial dimensions are well-bounded. These are mathematical representations of many of the relations in the CAD model, which define how the parts fit together, i.e. the top housing being able to snap onto the base during assembly, and that the cylindrical geometry inside the top housing fits into the hole in the base housing. Further, the hub needs to fit inside the trigger geometry in the top housing, the plug and needle inside the hub, the spring around the trigger, and so on. In combination with some of the axial constraints, these radial constraints also prevent part collisions - e.g. that the needle goes through the hole in the base, that the trigger arms on the hub are able to pass through the top housing, and so on.

Axial constraints - In the axial direction, the constraints represent several dimension *chains* that amount to tolerance stack-ups and the minimum and maximum axial overlaps required to ensure correct interface connections between the parts. An example of such is illustrated in figure 3.15; the internal components must be dimensioned relative to the outer components in a way that prevents the needle from ever protruding through the valve in the device's steady-state when accounting for the total contribution of axial tolerances, represented through the tolerance parameter Z_{tol}. Other examples of axial constraints include that the trigger geometry on the hub cannot protrude out of the top housing, that there needs to be sufficient overlap between the top and base housing to create a sealing interface in which the valve component is mounted, and that there is sufficient overlap between the plug and the trigger arms.

Manufacturing Constraints - As mentioned, the SOMA device would involve a substantial manufacturing volume, under very strict quality management requirements, due to the regulated nature of the pharmaceutical industry. Therefore, a set of constraints were introduced to represent the manufacturing requirements that the SOMA project needed to design towards. These were modelled to represent the intended manufacturing setup, meaning alternative joining methods and manufacturing processes than the ones already used in the existing design were considered.

Ultimately, these constraints represent the limits of what can be done in manufacture without incurring excessive manufacturing cost (e.g. through long cycle times in injection moulding) or the need for uncontrollable or low yield processes (e.g.). As a result, they, in unison, represent unmodelled manufacturing cost and scrap rate objectives.

Thus, the manufacturing processes involved - e.g. injection moulding, coiling, and compaction - were all represented through the constraints. Injection moulding constraints include minimum wall thickness requirements (which affect almost all design variables in the model),

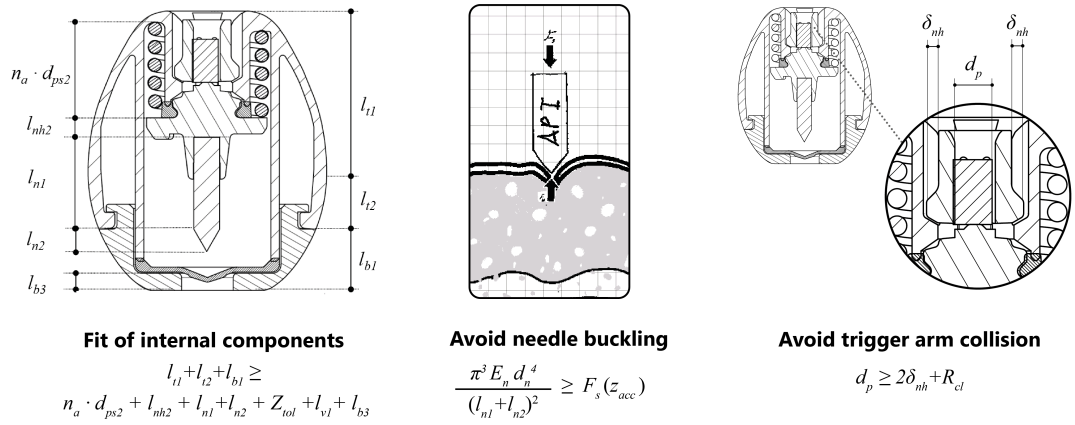


Figure 3.14: Examples of the constraints derived for the optimization of the SOMA device. *Left*: The internal components need to fit inside the outer components of the SOMA device. *Middle:* The needle must not buckle upon impact with gastric tissue. As the damping properties of the tissue and needle are unknown, a simple critical buckling load criterion is applied. *Right:* When the plug dissolves, the trigger arms are pushed inward by the spring force acting on the ratchet surface. In order for the trigger arms to pass through, they must not collide with each other upon triggering. This imposes a constraint on the relationship between the plug diameter, d_p and the trigger overlap δ_{nh}, involving a clearance parameter, R_{cl}.

minimum radial clearances, draft angles on long surfaces, and the geometric constraints that prevent the need for undercuts in the mould. The spring design is also constrained by spring manufacturing constraints such as minimum and maximum spring index [112], limits on its aspect ratio and the number of dead windings to prevent tangling, and a maximum axial pretension relative to its size, based on the limits of assembly equipment used on this scale. All of these constraints were defined with input from DfMA engineers and process simulation specialists within the case company.

Structural Constraints - The SOMA device is exposed to a set of static and dynamic load cases during its operation. The SOMA device is statically loaded by a compression spring, meaning that the load-bearing components need to be dimensioned to withstand the resulting stress.

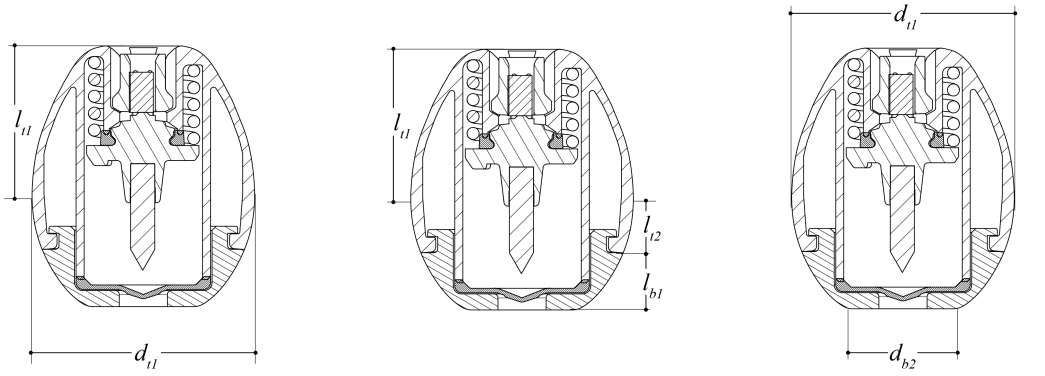
First and foremost, the compression spring can only be loaded to a certain point; to account for this, a yield constraint was imposed, involving a *Von Mises* failure criterion, a *Wahl* correction factor to account for any stress concentrations, and a safety factor to account for variation and avoid relaxation. These allow the accurate assessment of stress, given that the optimum might exist at spring proportions resulting in a high pitch or a low spring index. Springs with a high pitch are loaded in shear and bending [112], meaning the standard shear based expressions for stress in compression springs do not apply, while a low spring index (ratio between d_{ps1} and d_ps2) results in stress concentrations. The safety factor was selected based on the standard requirements in the case company, stemming from the achievable tolerance grades and the storage conditions of assembled devices containing pre-stressed springs.

The resulting loads on the trigger arms and trigger surface in the top housing were also accounted for through constraints, as the bucking of the trigger arms and creep fracture of the top housing need to be avoided. Finally, the trigger arms are deflected inward during

injection, and the needle impacts tissue. A simple beam equation was used to define a stress constraint for the deflection of the trigger arms. Meanwhile, to avoid the needle shattering upon impact, a simple buckling criterion was defined. As the needle is an anisotropic material, this is only a conservative estimate of the limits as to the load the needle can withstand.

As the SOMA will be contained in a dissoluble capsule or covered in a powder when packaged, it comes with a degree of protection against misuse by the user. Hence, scenarios such as the user accidentally chewing on the SOMA, dropping it before ingestion, or the like, was left out of the optimization model, especially given that this is an early-stage design study, where such situations are rarely dealt with in-depth.

Equality Constrains



Height width ratio
$l_{t1} = C_t d_{t1}$

Base-top height ratio
$l_{t1} = C_{tb}(l_{t2} + l_{b1})$

Base-top diameter ratio
$d_{b2} = C_b d_{t1}$

Figure 3.15: Examples of the equality constraints that determine the shape of the SOMA device in the optimization model.

Shape functions - As mentioned, the outer shape of the device is represented through equality constraints, which ensures that the overall outer shape of the device remains unchanged, as the rest of the design variables are resized during optimization. These equality constraints take the form of basic relative relationships between certain outer design variables, combined with shape parameters determining their relative sizing. In numerical solution, these ratios were taken from by measuring the shape of the SOMA design generated through optimization of stability and self-orientation speed.

Part fits - Certain part-part relationships could only be expressed through equality constraints. Firstly, as the valve and o-ring are used to prevent liquid from entering the device, they need a specific interference fit (nominally) in order to function as specified. While there is a tolerance on said interference, the band is so small that it would not make sense to explore proportional changes that affect the interference fit. In a similar manner, the plug needs to fit snugly into the hub in order to keep the trigger arms in place until it is dissolved. The top and base housings also need to snap together, creating an exact fit resulting in a smooth outer surface to avoid edges affecting self-orientation. Finally, the acceleration stroke, z_{acc} is determined by the relative sizing of the height of the device and the length of the internal components. As the acceleration stroke cannot exceed the amount of clearance left inside the device, an equality constraint is necessary to constrain the model from exploring non-real results.

3.5.4 Optimization Model

Transforming all of these in negative-null, minimization form, the resulting initial optimization model is

$$\text{min} \qquad f_1(\mathbf{x}) = -\frac{\sum_{p=1}^{p=8} m_p \cdot (C_p + Z_p)}{(l_{t1} + l_{t2} + l_{b1}) \cdot \sum_{p=1}^{p=8} m_p} \tag{3.46}$$

$$f_2(\mathbf{x}) = d_{t1} \tag{3.47}$$

$$f_3(\mathbf{x}) = -\rho \frac{\pi}{4} d_{n1}^2 \left(l_{n1} + \frac{1}{3} \cdot l_{n2} \right) \tag{3.48}$$

$$f_4(\mathbf{x}) = -\sqrt{2 z_{acc} \left(g_C + \frac{z_{pre} G_{st} d_{ps2}^4}{8 d_{ps1}^3 n_a m_{acc}} - \frac{z_{acc} \left(1 + \frac{m_{acc}}{m_{rec}} \right) G_{st} d_{ps2}^4}{16 d_{ps1}^3 n_a m_{acc}} \right)} \tag{3.49}$$

$$\text{s.j.t} \qquad \mathbf{g}(\mathbf{x}) \leq 0 \tag{3.50}$$

$$\mathbf{h}(\mathbf{x}) = 0 \tag{3.51}$$

$$\mathbf{x} \in \mathbb{P} \tag{3.52}$$

As mentioned, the self-orientation objective in this model was verified through comparison with a CAD model of the SOMA device. In general, the reference design was used in a variety of ways to verify the model. The velocity objective was cross-checked with numerous measurements and experiments made by the SOMA project, and the objective function was within the measurement uncertainty when the design variable values of the reference design were input into the objective function. Correspondingly, the constraint functions were checked through a range of check-sums, which were cross-referenced with the CAD model, to ensure that the shape functions, fit constraints, and the like matched the geometric representation of the SOMA device seen in the CAD model.

In chapter 4, the multiobjective monotonicity analysis and numerical solution of the model are presented. This includes a transformation to a form that can be solved using gradient-based solution methods, specifically an ϵ-constraint formulation, which will be solved with the sequential quadratic programming (SQP) subroutine in the fmincon function in MATLAB.

4 Trade-off Identification and Root-cause Analysis

This chapter presents the descriptive and prescriptive research involved in the exploration and development of methods for the identification of trade-offs and their underlying causes. The chapter starts with an overview of initial sub-studies related to RQ1, which led to an important set of realisations which formed the genesis of the decisions made in answering RQ2. This is followed by content from Paper A in sections 4.2-4.3, covering novel mathematical developments within Monotonicity Analysis. These allow a set of analysis procedures, collectively referred to as Pareto-set Dependency Analysis, for the identification of trade-off root causes in a design. The chapter concludes with the application of this new methodology to the SOMA case. Sections 4.1.3. and onward are adapted from Paper A and expanded further, providing additional details and perspectives.

Initially, this research project was largely shaped by the assumption that trade-offs in design are generic; i.e. that certain fundamental common causes exist (certain forms of design error and common avoidable dependencies) which cause trade-offs of a similar type (e.g. efficiency vs. accuracy, stiffness vs. weight) in a wide range of systems. A pair of initial studies tested this initial assumption, and revealed that one of the hypotheses of the research was probably invalid. The realisations made in this process shaped the subsequent research activities, and an overview of these initial studies is given in the following to provide a background for the rationales behind the decision to further develop monotonicity analysis [45] for the purposes of the research project.

4.1 Initial Studies

The initial research of this PhD project involved efforts to explore existing analysis methods that would be suitable trade-off identification for an early stage of product development. The hope was that these could be developed further to address the aspect of cause.

This was done through two studies:

1. The first was a study on the application of the *Contradiction Index* method by Göhler & Howard [17] to Novo Nordisk device, which was already in the market. The aim was to identify drivers of trade-offs between design objectives and assess the influence of these could be traced in the final specifications of the product, the instructions for use, and the challenges seen in the production of the device.
2. The second, on the application of design space exploration methods, in an ongoing product development project in Novo Nordisk. This was conducted as action research, with the aim of supporting decision making and requirement specification based on trade-off analysis

Neither study yielded the expected results and failed to meet their original intents. While they might have been seen as failed studies, they did result in important learnings, which informed the developments ultimately made in this PhD project. Hence, these studies will be touched upon in the following, as they help illustrate the reasoning behind the subsequent

methodological developments presented in this dissertation. The details of the methods applied, the results, and the cases involved, are not touched upon in much detail; they are included here in brief form in the service of showing the rationales behind the subsequent methodological developments.

4.1.1 A Qualitative Approach - The Contradiction Index

The first study of the PhD explored potential extensions and applications of the Contradiction Index [17, 21, 25]. The Contradiction Index (CI) approach involves an analysis method that combines some of the rationales in Axiomatic Design [13], TRiZ [16], and the design structure matrix [61]. Using terminology consistent with Axiomatic Design [13], the method involves qualitative identification of design parameters (DPs) with contradictory design intents when viewing the different *functional requirements* (FRs) they contribute to. While Axiomatic Design largely prescribed that all "coupling" (i.e. dependencies) should be avoided, Göhler et al. [17] drew on perspectives from TRIZ to suggest that one should distinguish between *positive* and *negative* coupling. They proposed that positive couplings are, in fact, not detrimental, as they allow more functionality (i.e. FRs) with less complexity (i.e. DPs). Thus, the CI method focuses on identifying these negative couplings and counting the amount of them to measure how *contradicting* the design is. Having used the CI method design engineer with some success, the author at first aimed to further develop such qualitative methods to gain a better understanding of common drivers of trade-offs between design objectives. This might then have been used to prescribe a set of heuristics with which to identify and avoid trade-offs at an early stage of design.

Yet, applying the Contradiction Index method to an initial case - the Novo Nordisk FlexTouch device shown in Chapter 1 - revealed that qualitative analysis was insufficient for the purposes of this research. The research plan for the study was as follows:

1. Identify functional requirements based on the product specification document and based on input from the designers involved in the development of FlexTouch.
2. Decompose the design to a level detail where the variables involved in existing mechanical analyses (e.g. dose accuracy, and dosing speed calculations) were included.
3. Apply the CI method
4. Map out all of the functional requirements in trade-off and how the variables shared between them influence the problem (i.e. whether they are negative or positive dependencies)
5. Use this insight to identify the underlying cause of the trade-offs in the design
6. Perform structured interviews with the product development team regarding the key challenges and issues that arose throughout the development of FlexTouch
7. Assess whether the results of these interviews correlate with the observed drives of trade-off. Is there a link between the trade-offs in the system, the dependencies that drive them, and the issues seen throughout the development process?
8. Repeat this process in ongoing development projects and identify commonalities.

Steps 1-4 resulted in a design structure matrix describing the number of contradictions between each pair of FRs. This primarily revealed a substantial contradiction between dosing speed and dosing accuracy, which was not unexpected. Yet, step 5 resulted in some key realisations that meant that the rest of the study was abandoned. In inspecting the results of the analysis, it became clear that there was a need to involve more rigorous aspects in the work. The reason is twofold.

Firstly, there is the question of the **influence of constraints**. Mechanical design involves the synthesis and improvement of mechanical systems, subject to numerous design objectives and design constraints. When prescribing the axioms of Axiomatic Design [13], Suh bundled these into one aspect; functional requirements. Constraints are the limitations imposed upon a design problem by the desire to avoid failures caused by fundamental physical phenomena such as plastic yield of materials, geometric constraints such as part needing to fit together, and practical considerations such as manufacturability [12]. When these are *active*, they influence the achievable performance of a system, i.e. the optimum of each objective. This causes dependencies that are specific to certain regions of the objective space. If one, for instance, simultaneously minimises the mass of a part and increases the load the system exerts upon it, then at some point, the limits of the material will determine the dimensions of the component. In becoming active, the constraint thus reduces the degrees of freedom in a design problem, locking the relationship of its dependent variables relative to one another. In such situations, the constraint matters - it becomes a requirement that influences the design problem. By bundling constraints and objectives together, the notion of constraint activity is lost in Axiomatic Design, meaning one might overlook important contributors to trade-off unless the designer knows which constraints are active a priori.

Secondly, there is the aspect of the **degree of trade-off** between objectives. A key initial hypothesis in this research was any dependency that contributes to trade-off is detrimental and should be avoided. The analysis of the FlexTouch led to the realisation that this is not necessarily given; a dependency can contribute to a trade-off without causing much/any loss of utility. The CI of the FlexTouch pointed to several contributors to trade-offs that potentially have no bearing on the achievable performance due to their small influence on the performance of the system. In other words, a trade-off might be insignificant, and qualitative methods such as the CI cannot capture the extent to which an objective pair is in trade-off and how this affects the achievable performance (i.e. the location of the optimal set). No existing qualitative technique for dependency analysis - neither DSM, AD, TRiZ, or QFD - allows the systematic assessment of the degree of trade-off or the identification of active constraints and their effect on the design problem. This ultimately requires quantitative analysis.

4.1.2 A Sampling-based Approach - Design Space Exploration

Given these realisations, the focus of the research shifted towards more quantitative methods that might be applied from a very early stage of product development to ensure that any outputs of analysis might be applied to actually inform embodiment or even conceptual design. Hence a case study was performed in an ongoing product development project in Novo Nordisk, which was in the process of choosing between two competing embodiment designs. Thus, an action research study was designed, where the author was embedded in the product development project, with the aim of applying design space exploration (DSE) techniques to compare the alternative embodiments. Essentially, the ongoing development project involved the design of a new injector pen device, where each of the two embodiments had quite different arrangements of the internal components and relied on different working principles. Not all of the functionality had been realised, nor had all requirements had been specified. A more detailed explanation of this development project and the embodiments involved has been omitted due to intellectual property restrictions.

The hypothesis of the study was that combining simple DSE models with the CI method might reveal the trade-offs present in each embodiment, the active constraints, and an approximation of the Pareto set. It was thought that this would reveal enough information to drive a redesign process for each embodiment to reduce their respective trade-offs, resulting in an improved Pareto set. When subsequently updating the analysis to account for the changes made in design, the new Pareto sets could then be used to compare the redesigned embod-

iments to support the selection decision. The basic idea was that the *better embodiment* would have a larger design space, better achievable performance, and lesser trade-offs and that this would be measurable, even with incomplete designs and requirement uncertainty.

A model was built to describe one of the embodiments, w.r.t. a set of objectives of importance, identified based on input from the product development project team. These objectives were expected to be the selection criteria in the decision between the two embodiments. Correspondingly, a set of sampling variables and sampling ranges were identified with input from the team. These were selected based on:

1. A desire to sample the variables that represented the key differences between the two embodiments. This was to extrapolate the effect of the differences between the two designs.
2. Uncertainties; the project wanted to gain input on the effect of different decisions being made. An example of such was the magnitude of the allowable spring pre-load, which affects the complexity of the assembly processes and the shelf life of the assembled device (both are cost drivers), but also affects device size and dosing speed.
3. A rough application of the Contradiction Index method, which revealed several contradicting variables in each design; examples include the spring dimensions, the lead screw pitch angle and diameter, and several variables in the activation mechanism.

However, before modelling of the second embodiment could be finished, the targeted launch date of the injection device was moved forward. With a much tighter timeline, the project team no longer had the time to develop both options concurrently. Hence, it was simply the embodiment that had most of the sub-functionality realised that was selected. As a result, the initial purpose of the action research study was no longer feasible. This is not an uncommon challenge in engineering design research - performing research in an industrial setting simply does not provide pristine laboratory conditions in which to explore the application and development of new methods [4].

While the original intent of the study was no longer feasible, the modelling efforts for the first embodiment revealed that the study might, in fact, never have reached the desired goals anyway. Essentially, DSE is a brute force approach, which involves sampling a large number of designs, and subsequently evaluating whether they are feasible, and if so, how they perform. There are various implementations with different benefits depending on the nature of the problem. The following DSE approach, shown as a pseudo algorithm in iterative form for the sake of clarity, was used:

Eliminate h$(\mathbf{x};\mathbf{P})$, by algebraically solving for one of their dependent variables and back-substituting
Sample the values of **x** between predefined limits, $\mathbf{x_L}$ and $\mathbf{x_U}$.
Store these in **X**, a $[n, i_{iter}]$-dimensional matrix
for i = 1..i_{iter} **do**
  $\mathbf{x} = \mathbf{X}(i,:)$
  Evaluate $\mathbf{f}(\mathbf{x};\mathbf{P})$ and store the outputs in $\mathbf{F}(i,:) = \mathbf{f}(\mathbf{x};\mathbf{P})$
  Compute $\mathbf{g}(\mathbf{x};\mathbf{P})$ and store the output in $\mathbf{G}(i,:) = \mathbf{g}(\mathbf{x};\mathbf{P})$
  if $\mathbf{g}(\mathbf{x};\mathbf{P}) \leq \epsilon$ **then**
    Q(i)=i,
    else **Q**(i)=0;
  end if
end for
Q=**Q**>0;

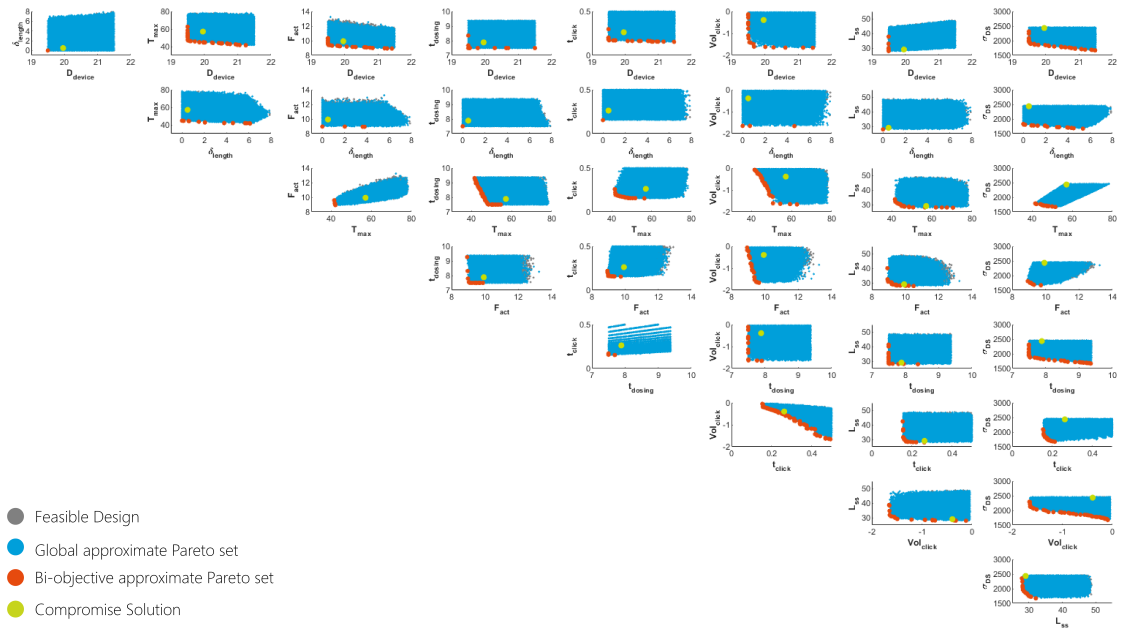


Figure 4.1: 2D result plots of the objective space, from the design space exploration of the first embodiment design, with all design objectives transformed into minimisation form

Apply a non-dominance filter to **F**(**Q**, :)

Here, **x** is the vector of variable values for each sample, **P** is a vector of design variables, **f(x;P)** the vector of design objectives, **g(x;P)** and **h(x;P)** are the inequality and equality constraint vectors expressed in negative null form, **Q** is the vector of feasible indices, *n* is the number of design variables, i_{iter} is the number of samples, ϵ is the allowable residual error for the constraint functions (typically $\epsilon \leq 10^{-6}$). The equality constraints are eliminated prior to sampling, as this avoids the generation of a large number of designs that will never fulfil an equality constraint. A few combinatorial sub-routines were built into the model to allow exploration of the effect of certain design decisions. Examples include the use of a rectangular spring wire over a round wire, alternative material selections, alternative mechanism designs, and altering the radial arrangement of the components (i.e. one part inside the other and vice versa). This yielded the objective spaces shown in figure 4.1 in a run with 10^7 samples.

As the model leaves out numerous design variables by design, it does not strictly identify Pareto points. That said, it does reveal several trade-offs. Their cause is, however, not as clear, given the limitations of the approach, which became evident upon the solution of the model:

1. As can be seen in the result plot, the amount of non-dominated samples is almost equal to the number of feasible samples. The a priori identification of variables that contribute to trade-off and their use as sampling variables means that most feasible samples will be non-dominated.

2. Numerous constraints were found to be active in all samples, while others were only active in certain regions of the objective space. Yet, due to the dimensionality of the problem, it is difficult to identify regional constraint activities that occur due to specific combinations of values of more than two objectives

3. While the inequality constraints are evaluated, it is impossible to determine which vari-

ables the bound in each sample, without further analysis. Thus, it is difficult to identify the relationships that exist at the bi-objective approximate Pareto frontiers, which would help explain the causes of the pair-wise trade-offs.

Thus, a better approach to analysing the influence of constraint activity across the objective space was necessary in order to identify the global and local dependencies that exist between design objectives which drive trade-offs.

4.1.3 Resulting Research Efforts

As one might gather from the limitations of the two aforementioned studies, the challenge in studying the contributors to trade-offs in a design at an early stage of design is that trade-offs are specific to optimal designs. Clearly, a trade-off can only exist when one objective cannot be improved further without detriment to the other. This implies Pareto optimality, meaning that all variables values are either contributing to the trade-off, at the edge of their feasible domains, or at an interior optimum, which is shared by all objectives.

The approaches in the two studies were flawed in that the Pareto set is created through two distinct types of dependency, one of which cannot be handled qualitatively nor through sampling. Consider that the Pareto set $\mathcal{C}$ exists on the boundary of the attainable set $\mathcal{A}$ but is not necessarily defined by the constraints alone, as unconstrained multiobjective problems also yield Pareto sets [113]. It follows that the occurrence of Pareto sets must have two causes:

1. *Trade-off variables* - Global Dependencies In negative-null form, a variable x that is shared by two objectives, $f_1(x)$ and $f_2(x)$, causes a trade-off if $\arg\min f_1(x) \neq \arg\min f_2(x)$. This can only occur if the objectives are either oppositely monotonic, or when one or both are non-monotonic w.r.t. x.

2. *Active constraints* - Regional or Local Dependencies Active constraints reduce the degrees of freedom (DOF) in optimization problems, affect the feasible domains for the remaining DOF, and change the optimum. Suppose we are able to identify an active constraint and solve it w.r.t. the bounded variable. Back-substituting, the resulting expression into the objectives, may introduce new variables into an objective or change its monotonicity w.r.t the original variables. Some of these may be trade-off variables, meaning that active constraints can create dependencies specific to the optimum, resulting in trade-offs between the objectives. To understand such relationships, we need to find an explicit expression of the relationships between the objectives at the optimum.

Existing qualitative methods and heuristics do not account for the effects of such situations, meaning the focus of the PhD needed to turn toward multiobjective design optimization methods. Yet, design optimization research generally does not concern itself with understanding *why* the optimum is a set rather than a single dominant solution, and *what* defines the shape of this set. Rather, current methods are more preoccupied with modelling and solving increasingly complicated optimization problems, identifying and selecting desirable points within the set, or developing measures to describe the set. Selecting a point in a Pareto set includes work on modelling preferences [81–83], identification of compromise solutions by measuring the distance to a utopia point [54], scaling methods to account for objective weighting [114], and strategies for making trade-offs aggressively or conservatively [84]. Substantial work also exists for sensitivity, robustness [77], uncertainty [78], visualisation [92], dimensional reduction [91], and identification of competing objectives in a *n*-dimensional objective space [81].

Thus, for the purposes of this research, there are three challenges with the approach of

current multiobjective design optimisation methods to design. First, the main focus is on optimizing a fixed design rather than questioning why the objectives compete. Second, the analysis done at earlier time points in the product's evolution may become obsolete during a later design phase. Finally, if the Pareto set contains no points acceptable to the designer, e.g., due to non-modelled considerations, there is little guidance for what to do next. A rigorous approach to gain insights into the root cause of the trade-offs inherent to the design would substantially increase the value of optimization at an early stage of product development.

Furthermore, there is currently no well accepted and consistent terminology to describe the drivers of trade-offs in a design problem. Axiomatic design uses the notion of coupled design parameters, TRiZ "contradictions", whereas DSM based methods largely concern themselves with the notion of "positive" and "negative" dependencies [115]. No source in optimization literature was identified that distinguishes between the manner in which shared variables and constraints affect the relationships between design objectives.

Thus, the development of a mathematical foundation for the identification of dependencies, and analysis of their influence on a design problem, was deemed to be necessary for the purposes of the research. In this regard monotonicity analysis, provides promising perspectives. It has previously been applied in a design context, utilising the knowledge gained from identifying the properties of the optimum in single-objective problems to support decision making. Yet, most of this prior work is outlined in section 3.4.1. [32, 102, 104] has focused on understanding the common characteristics of Pareto-optimal designs to allow reuse in future designs but within a single configuration. Yet, if a configuration has inherent limitations, one would simply find the best compromise for a poor design. If MA can identify relationships for the design variables at optimality, then arguably, it might also be able to identify relationships that *limit* optimality. In a multiobjective formulation, such analysis could lead to the discovery of the root cause for trade-offs between objectives.

Thus, it was hypothesised that monotonicity analysis could be used as a dependency analysis method for multiobjective problems. Specifically, it might be used with the intent of reaching a better understanding of the relationships that drive trade-offs, rather than merely using it as an approach to model reduction and identification of optima without computation. The reasoning behind this was threefold. Firstly, monotonicity analysis reveals dependencies. When setting up a monotonicity table, variables shared between expressions are identified, and their influence on the function is assessed. It is decreasing, increasing, or non-monotonic? If we applied this to multiobjective problems, we would identify the shared variables that contribute to trade-offs and the variables that do not. Secondly, monotonicity analysis reveals the effect of constraints. It involves the identification and back-substitution of active constraints revealing the relationships that are specific to the optimum. In multiobjective problems, these might help explain the cause of the shape of the Pareto set - i.e. reveal regional relationships or explain the existence of the Pareto set itself.

Finally, monotonicity analysis is unique in an optimization context in that it can be applied without computation and even before a full optimization model has been constructed. Thus, it was envisioned that as opposed to more computationally intensive methods for large, complicated models, MA might be applied in early design using simpler analytical models to better understand the trade-offs. In engineering practice, analytical models are often constructed with the aim of better understanding the design problem, rather than reaching an accurate result [116]. Thus, simple analytical models may still capture the drivers of trade.off, despite not necessarily accurately identifying the true optimum set by accounting for any and all non-linearities and 2^{nd} order effects. These realisations thus lead to an adapted version of research question 2, which will be addressed in the remainder of this chapter:

RQ2* How can conceptual or configuration design limitations reflected in the Pareto set be identified rigorously? In particular, what specific design dependencies and constraints cause trade-offs?

4.2 Extensions to Monotonicity Analysis

Upon these realisations, it became clear that new developments to monotonicity analysis (MA) would be necessary. MA has not seen broad application to multi-objective problems, nor has anyone developed a systematic approach to multi-objective MA with the intent of identifying the drivers of trade-off. Based on a literature search, only two prior works were identified. Michelena & Agogino [72] expanded the MA to multiobjective problems by applying MA to a weighted sum formulation, and demonstrating this in the parametric design of a hydraulic cylinder. While this allowed the identification of different Pareto optimal activity cases, and a reduction in computational effort, they did not present a systematic reduction process for larger multi-objective problems. Gobbi et al. [73] later applied MA to multi-objective problems stated in ϵ-constraint form, to support the analytical derivation of an expression of the Pareto set. Yet as they discuss, their approach is strictly only in low-dimensional problems.

Thus, the following developments to MA were needed, in order to apply it as a dependency analysis method with the aim of identifying trade-offs their root causes:

1. A systematic reduction procedure for multi-objective problems
2. A mathematical foundation for the identification and description of the variables and constraints that cause trade-offs between design objectives.
3. An approach to studying the relationships that exist between design objectives, locally or regionally in the objective space.

4.2.1 Selection of a Multiobjective Formulation

First, a suitable multi-objective formulation needed to be selected, as there are numerous alternatives with different benefits and limitations [54]. For the purpose of multi-objective monotonicity analysis (MOMA), the ϵ-constraint method [57], also known as the upper-bound formulation [12] was selected. As discussed in Section 3.3, this involves converting the a standard multiobjective optimization problem into a single objective one:

$$\begin{aligned} \text{min.} \quad & f(\mathbf{x}) && (4.1)\\ \text{s.j.t} \quad & \mathbf{c}(\mathbf{x};\epsilon) \leq 0 && (4.2)\\ & \mathbf{g}(\mathbf{x}) \leq 0 && (4.3)\\ & \mathbf{h}(\mathbf{x}) = 0 && (4.4)\\ & \mathbf{x}, \epsilon \in \mathbb{P} && (4.5) \end{aligned}$$

As mentioned, the vector ϵ of parameters ϵ_i represents the upper bounds of the bound objectives. When $f(\mathbf{x})$ is minimised for given values of ϵ_i, then the solution $\mathbf{x}^*$ is Pareto optimal if all of the bound objectives are active with non-zero Lagrange multipliers. Pareto points are thus identified by varying ϵ systematically between lower $\epsilon_{\mathbf{L}}$ and upper limits $\epsilon_{\mathbf{U}}$. See [54, 57, 117] for an overview of works on the upper bound formulation, the underlying mathematics, and approaches to defining suitable limits for ϵ. In this work, the Pareto set is constructed by sampling a set of ϵ parameter values:

$$\mathbf{E} = (\epsilon_{\mathbf{U}} - \epsilon_{\mathbf{L}})\mathbf{R} + \epsilon_{\mathbf{L}} \tag{4.6}$$

where **R** is a matrix of uniformly distributed quasi-random numbers between 0 and 1 of the dimension [*k*-1; *j*], where *j* is the number of computational iterations, and *k* is the number of objectives. A low discrepancy quasi-random set (e.g., a Halton set) can be used to reduce bias in **R** to reduce the computational cost of achieving a Pareto set with low sparsity [55]. After sampling, the optimization problem is solved iteratively, as in the following pseudo-code:

for i = 1..j **do**
  Set upper bound on constrained objectives, $\epsilon = \mathbf{E}(:, i)$
  Solve optimization problem w.r.t ϵ
  Store optimum, $\mathbf{F}^*(:, i) = [f^*, \epsilon^T]^T$
  Store arguments, $\mathbf{X}^*(:, i) = \mathbf{x}^*$
  Store Lagrange multipliers, $\Lambda(:, i) = \lambda$
  Store constraint values ,$\mathbf{G}^*(:, i) = \mathbf{g}(\mathbf{x}^*)$ and $\mathbf{H}^*(:, i) = \mathbf{h}(\mathbf{x}^*)$
end for

The sparsity of the approximated Pareto set decreases as *j* increases, while the span increases with *j* and the difference between ϵ_U and ϵ_L. With an increased *j*, one identifies more Pareto points resulting in a more dense Pareto set. A high *j* can be necessary to approximate the shape of the Pareto set, should it have interactions between the objectives that exist locally in the attainable set, for instance creating knee like shapes [118]. Beyond a certain limit, the Pareto set will have been exhaustively constructed, meaning no additional feasible solutions can be found by further increasing the difference between ϵ_U and ϵ_L. Thus, one can also solve the MODO problem multiple times with a relatively low *j*, increasing the difference between ϵ_U and ϵ_L, until the boundaries of the Pareto-set seem to have been identified, and then subsequently increasing *j* to the desired level of density.

This approach was selected, as it is very well suited to MA, as it ensures that the extension into multiple objectives is relatively straightforward. The principles and procedures originally developed by Papalambros and Wilde [12] mostly still apply. In this form, multi-objective MA merely involves handling more constraints, albeit ones that represent bound objectives. Furthermore, the ϵ-constraint formulation has additional benefits:

1. *Maintaining monotonic properties*
   Converting a set of objectives into a composite function, e.g., a weighted-sum as done by Michelena & Agogino [72], can result in loss of monotonic properties when the objectives share variables. The proof for this is as follows:

   Let f_1 and f_2 be two differentiable functions wrt. a design variable x_1. If f_1 and f_2 are oppositely monotonic wrt. x_1, i.e. $f_1(x^+)$ and $f_2(x^-)$, then a weighted composite function of the two $U = f_1 w_1 + f_2 w_2$ will have the following partial derivative:

   $$\frac{\partial U}{\partial x_1} = \frac{\partial f_1 w_1}{\partial x_1} + \frac{\partial f_2 w_2}{\partial x_1} \tag{4.7}$$

   Given that monotonicity by definition implies that the partial derivative of f_1 wrt. x_1 will be strictly positive, while it will be strictly negative for f_2, the sign of the partial derivative of the composite function U wrt. x_1 (and hence its' monotonicity) will therefore depend on the value of the partial derivatives, and upon the weighting. U will as such only be monotonic if it holds that $\frac{\partial f_1 w_1}{\partial x_1} \leq \frac{\partial f_2 w_2}{\partial x_1}$ for all values of x_1, $w_1$and w_2, or it holds that $\frac{\partial f_1 w_1}{\partial x_1} \geq \frac{\partial f_2 w_2}{\partial x_1}$ for all values of x_1, w_1 and w_2. This means that U may be non-monotonic, complicating monotonicity analysis. Using an ϵ-constraint formulation avoids this issue.

2. *Objective elimination*
   Introducing objectives as constraints in an optimization model allows one to parametrically study the *activity* of the bound objective across the attainable set using monotonicity analysis. If a bound objective can be determined to be active through monotonicity analysis, the objective itself can be 'optimized out' of the model through back-substitution [12], revealing how the objectives affect each other at the Pareto frontier. This is similar to the developments by Gobbi et al [73], but does require some extensions to account for regionally active constraints.

3. *Sensitivity data*
   Solving a constrained optimization problem yields non-zero Lagrange multipliers for active constraints, revealing the local sensitivity of the optimum w.r.t. changes in each active constraint. In the upper-bound formulation, the Lagrange multipliers of the bound objectives describe whether and to which degree the bound objectives compete with the primary objective – what Haimes and Hall [56] call the *trade-off ratio*. Similar information can be gained for an objective that has been optimized out, c_i, by computing a partial derivative of the remaining bound objectives, $\mathbf{c}$, w.r.t. the bound objective parameter:

$$\frac{\partial \mathbf{c}(\mathbf{x};\epsilon)}{\partial \epsilon_i} \tag{4.8}$$

It is often suggested that the most important objective should be modelled as the function being minimised [54], while the remaining objectives should be bound. To simplify monotonicity analysis, however, the most suitable approach would be to select the objective that has the greatest number of design variables. This allows the broadest application of *MP1* in problem reduction and lowering the required number of back-substitutions as a result.

4.2.2 Multiobjective Monotonicity Analysis (MOMA)

Definitions, Theorems, and Proofs

As mentioned, the use of the ϵ-constraint method makes MOMA relatively similar to MA of single-objective problems. The exception is that the bound objectives, $\mathbf{c}(\mathbf{x};\epsilon)$, cannot be treated as traditional inequality constraints. Firstly, as we wish to vary the upper-bound values, ϵ, these cannot be regarded as fixed parameters when performing monotonicity analysis. Secondly, we seek to partially minimize the bound objectives, which has implications for the use of MP1 and MP2. Recall that these are:

First monotonicity principle (MP1)
In a well-constrained minimization problem, every increasing variable is bounded below by at least one non-increasing active constraint.

Second monotonicity principle (MP2)
In a well-constrained minimization problem, every nonobjective variable is bounded both below by at least one non-increasing semi-active constraint and above by at least one non-decreasing semi-active constraint.

As the activity of $\mathbf{c}$ will depend on the values of ϵ, it is necessary to introduce some theorems of relevance to how $\mathbf{c}$ is handled.

Definition 1 Trade-off Variables
If an objective pair f and c_i have a variable x_1 in common, but differ in monotonicity w.r.t. x_1, e.g., $f(x_1^+)$ and $c_i(x_1^-)$, then x_1 is said to be a trade-off variable, denoted $\overline{\underline{x}}_1$. Correspondingly, an objective pair of like monotonicity w.r.t. a common variable, indicates that the variable is harmonious and can be used to partially minimise both simultaneously.

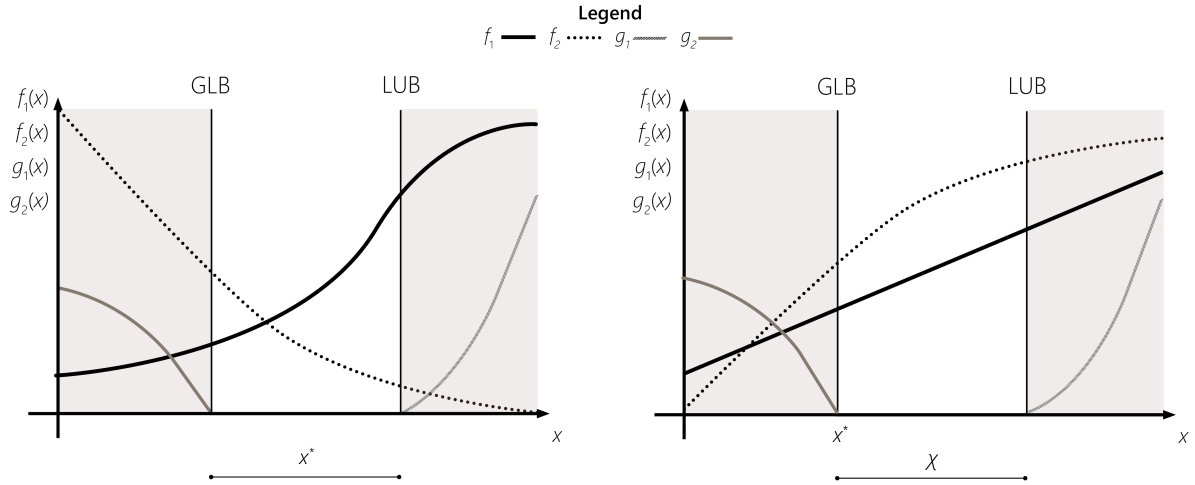


Figure 4.2: *Adapted from Paper A*:The difference between trade-off variables (left) and harmonious variables (right). Notice, that any solution in the feasible domain of the trade-off variable is optimal, whereas the harmonious variable has a single optimal value

Theorem 1 Influence of Monotonic Trade-off Variables
In the presence of monotonic trade-off variables, no dominant minimum exists, resulting in a Pareto set. The proof for this is a corollary to MP1.

Proof. Let f_1 be monotonically increasing w.r.t. $x \in \mathbb{P}$ and f_2 monotonically decreasing, and let x be well bounded from above and below. Then by MP1, $\arg\min \mathrm{f}_1(\mathrm{x}) = \underline{x}$, and $\arg\min \mathrm{f}_2(\mathrm{x}) = \overline{x}$, meaning that the minimizers for the two objectives are defined by the *greatest lower bound* (glb)and the *lowest upper bound* (lub) respectively. Hence any feasible value of x will yield a unique Pareto point. ■

Corollary 1.1 Boundedness of trade-off variables
Following Theorem 1, multiobjective problems can only be said to be well-bounded if all trade-off variables are bounded from above and below.

For instance, if a bound objective, c_i, is critical w.r.t. a monotonic trade-off variable, $\overline{x}_1$, then the multiobjective problem is not well bounded, as $\overline{x}_1 \to \infty$ or $\underline{x}_1 \to 0$ when $\epsilon_i \to \infty$ and f is minimised. This can either be handled by introducing additional constraints, or by selecting suitable limits for the upper-bound problem $\epsilon_{\mathbf{L}}$, $\epsilon_{\mathbf{U}}$.

By extension, a variable x_1 that has a non-monotonic influence on one objective $f_i(x_1^N)$ and a monotonic influence on another, i.e. $f_j(x_1^+) \wedge f_j(x_1^-)$ will also be a trade-off variable, so long as argmin $f_i(x_1) \neq$ argmin $f_j(x_1)$. Thus, we can in such cases employ regional monotonicity analysis to assess whether $x_1$might be a trade-off variable, looking at the monotonicity of the objective close the the glb and lub of x_1, while considering whether an interior optimum of x_1 might exist.

In upper-bound formulations, we treat objectives as additional constraints and iteratively identify Pareto points, exploring $\overline{\mathbf{x}} \in \mathcal{X}$, for different values of ϵ, as illustrated in Figure 4.3. If a bound objective is active, the model is essentially exploring a smaller region of the feasible domain $\mathcal{X}_\epsilon \in \mathcal{X}$. From this, an additional theorem arises:

Theorem 2 Activity of Bound objectives
A bound objective $c_i(\underline{x}; \epsilon_i)$ can either be active, semi-active, dominated, or inconsistent with

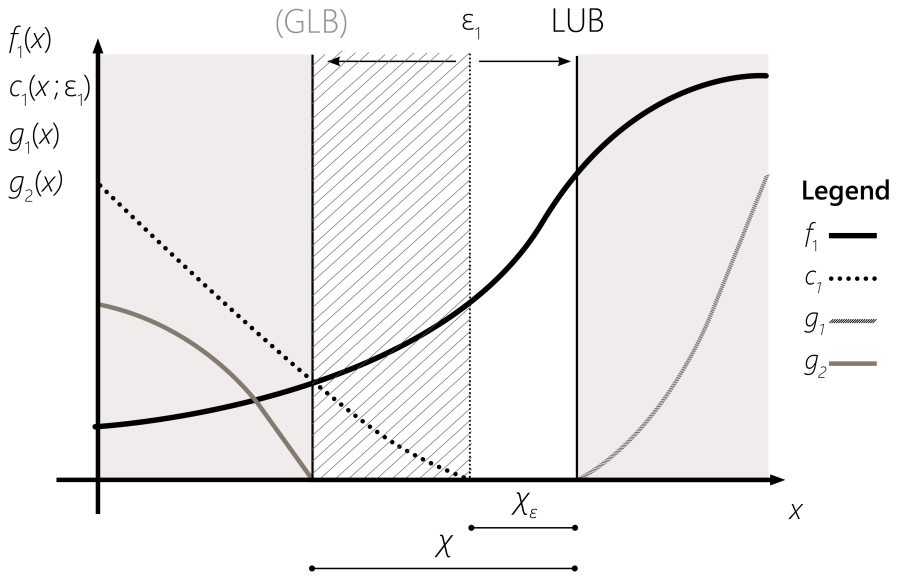


Figure 4.3: *From Paper A*: MOMA allows the partial identification of the Pareto set, by identifying the values of ϵ where the bound objectives are active, semi-active, violated, and inconsistent

other constraints, depending on the value of ϵ_i. The change in activity of $c_i(\underline{x}; \epsilon_i)$ across $\mathcal{A}$ affects the shape of the Pareto set.

Consider an objective pair, $f_1(x_i^+)$ and $c_1(x_i^-, \epsilon_1)$, with the design variable x being bounded from below by $g_1(x_i^-)$ and from above by $g_2(x_i^+)$, where ϵ is the upper bound parameter. Here, the value of ϵ determines constraint activity:

1. For the values of ϵ_1 where $g_1(x_i) < c_1(x_i)$, c_1 is active, and the result will be Pareto-optimal.
2. For the values of ϵ_1 where $c_1(x_i) < g_1(x_i)$, c_1 is inactive, and the result will not be Pareto-optimal
3. For the values of ϵ_1 where $g_2(x_i) < c_1(x_i)$, $\mathcal{X}_\epsilon \in$, and thus these constraints are inconsistent. In this case, g_2 shapes a boundary of the Pareto set.
4. For the value of ϵ_1 where $c_1(x_i) = g_1(x_i)$, the bound objective is semi-active, yielding the single-objective optimum for f_1. Correspondingly, $c_1(x_i) = g_2(x_i)$ yields the single-objective optimum for f_2.

Thus, exploring these changes in the activity of c_1 yields the Pareto set for the objective pair. We can hence utilise MOMA to identify the conditions under which a bound objective is active, dominated, or inconsistent. This can reveal important relationships between the objectives and the constraints g_i that affect the Pareto set. Here, it is important to consider the the influence of ϵ on the activity of $\mathbf{g}(\mathbf{x})$:

Definition 2 Global Activity
In the monotonicity analysis of an upper-bound problem, a constraint $g_i(\mathbf{x})$ is said to be globally active if and only if $f(\mathcal{X}_i) < f(\mathcal{X}_*)$ for any $\{\epsilon \in \mathbb{P} \,|\, \epsilon_L \leq \epsilon \leq \epsilon_U\}$.

Trade-off variables can only be optimized out if an active bound objective is used to eliminate it or if the bound objective can be determined to be dominated w.r.t. said trade-off variable by another globally active constraint. This notion of global activity is central to multiobjective

monotonicity analysis. A reduced model would potentially only identify parts of the Pareto set if we were to optimize variables out with constraints that are not globally active.

The final extension to MA that is necessary in order to deal with multiobjective problems is the question of how to partially minimise several objectives concurrently:

Definition 3 Partial minimisation of bound objectives
In a well-constrained multiobjective, upper-bound minimization problem, any increasing objective variable not in the primary objective, is bounded below by at least one non-increasing active constraint.

Modelling objectives as constraints is merely a route to identifying Pareto points. It is still desirable to identify partial minima for bound objectives. By simply extending MP1 into multiobjective problems, we can reduce multiple objectives, i.e., identify partial minima for f_{i+1} in $c_i(\mathbf{x}, \epsilon_i) = f_{i+1}(\mathbf{x}) - \epsilon_i \leq 0$. Nevertheless, it is necessary to take particular care in this process. Unless it is certain that the optimal value of a given variable is the same for all objectives, i.e., $\arg\min f_i(x) = \arg\min f_j(x)$ for any *i* and *j*, optimizing the variable out would result in a model that does not describe the entire Pareto set. When a globally active constraint can be identified, the bound objectives can always be partially minimized. This is relatively straightforward to do in situations where the condition $\arg\min f_i(x) = \arg\min f_j(x)$ for any *i* and *j* is upheld by definition. Following MP1, harmonious variables and critically constrained variables [12] will always meet this condition. As will variables that are bound by constraints that only depend on harmonious variables or on variables that only influence one objective, because constraint activity will be unaffected by the values of ϵ.

Impact of constraint activity in multiobjective problems

With these definitions, we can apply MA to multiobjective problems and, in doing so, identify trade-off variables that may be *hidden* in constraints. Here, it is beneficial to note the impact on the objective functions. There are two situations of relevance to trade-off analysis; when an objective changes monotonicity w.r.t a variable, or when it becomes dependant on new variables. Consider an example:

$$\begin{aligned}
\text{min.} \quad & f_1(x_1, x_2, x_3) = x_1^2 - x_2 + x_3 && (4.9)\\
& f_2(x_2, x_4, x_5) = \frac{1}{x_2} - x_4^2 + 2x_5 && (4.10)\\
\text{s.j.t} \quad & 2x_4 - x_1 \leq 0 && (4.11)\\
& x_2^2 + 4x_2 - 2x_3 \leq 0 && (4.12)\\
& x_2^3 + 2x_4 \leq P_1 && (4.13)\\
& 10 - 3x_5 \leq x_5^2 && (4.14)\\
& x \in \mathbb{P} && (4.15)
\end{aligned}$$

Without inspection of the influence of the constraints, it would seem there is no trade-off between f_1 and f_2, as they are both monotonically decreasing w.r.t the only shared variable, x_2. Yet, when converted into an upper-bound formulation, monotonicity analysis reveals

hidden dependencies:

$$\text{min.} \quad f_1(x_1^+, x_2^-, x_3^+) = x_1^2 - x_2 + x_3 \tag{4.16}$$

$$\text{s.j.t} \quad c_1(x_2^-, x_4^-, x_5^+; \epsilon_1^-) = \frac{1}{x_2} - x_4^2 + 2x_5 - \epsilon_1 \leq 0 \tag{4.17}$$

$$g_1(x_1^-, x_4^+) = 2x_4 - x_1 \leq 0 \tag{4.18}$$

$$g_2(x_2^+, x_3^-) = x_2^2 + 4x_2 - 2x_3 \leq 0 \tag{4.19}$$

$$g_3(x_2^+, x_4^+) = x_2^3 + 2x_4 - P_1 \leq 0 \tag{4.20}$$

$$g_4(x_5^-) = 10 - x_5^2 - 3x_5 \leq 0 \tag{4.21}$$

where f_2 has been converted into a bound objective $c_1(\mathbf{x}, \epsilon_1)$. Following MP1, it is clear that g_1 and g_2 are critical w.r.t. x_1 and x_3, respectively, for any value of ϵ_1, and are therefore active. Following Definition 3, we also conclude that g_4 is active as it is critical for x_5, meaning we partially minimize f_2 in c_1 by optimizing x_5 out. Solving for the minimizers yields $x_1^* = 2x_4$, $x_3^* = \frac{1}{2}x_2^2 + 2x_2$, and $x_5^* = 2$. With back-substitution, a reduced problem is reached:

$$\text{min.} \quad f_1(x_2^+, x_4^+) = 4x_4^2 + \frac{1}{2}x_2^2 + x_2 \tag{4.22}$$

$$\text{s.j.t} \quad c_1(x_2^-, x_4^-; \epsilon_1^-) = \frac{1}{x_2} - x_4^2 + 4 - \epsilon_1 \leq 0 \tag{4.23}$$

$$g_3(x_2^+, x_4^+) = x_2^3 + 2x_4 - P_1 \leq 0 \tag{4.24}$$

Here, f_1 has changed monotonicity w.r.t. x_2 and now depends on x_4, being oppositely monotonic to the bound objective c_1. Following Theorem 1, both x_2 and x_4 are trade-off variables, meaning that there is no single solution to the optimization problem but rather a Pareto set. Considering Corollary 1.1 the problem is, in fact, asymptotically bounded, as x_2 and x_4 are unbounded from below unless a well defined upper limit is imposed on ϵ_1. Hence, c_1 is globally active.

While this example may seem simplistic, it demonstrates the shifts in dependency between objectives that occur in the presence of active constraints. Such relationships are not necessarily easy to spot in non-reduced optimization models, nor is it given that the designer is aware of them. As such, monotonicity analysis can be used to identify trade-off variables, and in doing so, reveal what constraints in a design cause a lack of objective alignment - in this case, g_1 and g_2, as they introduce trade-off variables into the problem. Such insights may subsequently be used in a targeted redesign approach, aimed at eliminating specific dependencies-, or relaxing the constraints that reduce objective alignment.

4.2.3 ϵ-Monotonicity Analysis

With the theoretical developments introduced so far, one can apply monotonicity analysis to systematically reduce multiobjective models, gradually converging towards an explicit description of the Pareto set while identifying trade-off variables in the process. When all globally active constraints have been identified, one can optimize the active bound objectives out of the model. If one determines that $c_j(\mathbf{x}; \epsilon_j) \equiv 0$, and subsequently optimizes a trade-off variable $\overline{x_i}$ out, then $f(\mathbf{x})$ and $g(\mathbf{x}), c_i(\mathbf{x}; \epsilon) \in D_s(x_i), i \neq j$ become dependent on ϵ_j through back-substitution. A parameter from an eliminated bound objective will be denoted as $\tilde{\epsilon}_j$ and treated as a variable, referred to as the *reduced-objective variable*.

The reasoning behind treating $\tilde{\epsilon_j}$ as a variable is twofold. Firstly, the primary objective function has been transformed into a bi-objective function, $f(\mathbf{x}, \tilde{\epsilon}_j)$, describing the trade-off between the primary objective, $f(\mathbf{x})$ and $\tilde{\epsilon}_j$. Secondly, the feasible values of $\tilde{\epsilon}_j$ are now determined by a set of constraints $\mathbf{g}(\mathbf{x}, \tilde{\epsilon}_j)$. The bi-objective Pareto front between f_1 and

f_{j+1} will thus be defined by $f(\mathbf{x}, \tilde{\epsilon}_j)$ and $\mathbf{g}(\mathbf{x}, \tilde{\epsilon}_j)$.Meanwhile, the trade-offs amongst the eliminated objectives themselves are expressed through $\mathbf{g}(\mathbf{x}, \tilde{\epsilon})$, henceforth referred to as *Pareto-constraints*. This means that if we treat $\tilde{\epsilon}_j$ as a variable, identifying the constraints that bound it can be used to better understand the cause of the shape of the Pareto set.

In principle, all active bound objectives can be eliminated from the model. This will result in a multiobjective expression $f(\mathbf{x}, \tilde{\epsilon})$ describing the trade-off between the primary objective and all others, while all the Pareto-constraints $\mathbf{g}(\mathbf{x}, \tilde{\epsilon})$ describe the trade-offs between the eliminated objectives. However, it may not always be beneficial to do so, for instance, when elimination results in a loss of monotonic properties or when explicit elimination becomes too time-consuming. To allow the furthest reduction of the model, it is beneficial to attempt to eliminate the trade-off variables that are shared between the largest number of constraints.

What remains after objective reduction is:

$$\begin{aligned} \text{min.} \quad & f_1(\mathbf{x}, \tilde{\epsilon_1}, ..., \tilde{\epsilon}_{k-1}) && (4.25)\\ \text{s.j.t} \quad & \mathbf{g}(\mathbf{x}, \tilde{\epsilon}) \leq 0 && (4.26)\\ & \mathbf{h}(\mathbf{x}) = 0 && (4.27) \end{aligned}$$

where $f_1(\mathbf{x}, \tilde{\epsilon_i}^+)$ or when $\tilde{\epsilon_i}$ is a maximisation objective, and $f_1(\mathbf{x}, \tilde{\epsilon_i}^-)$ when $\tilde{\epsilon_i}$ is a minimisation objective. Applying monotonicity analysis to this formulation thus allows the identification of active Pareto-constraints at the single objective optimum, f_1^*. Solving for $\tilde{\epsilon_i}^*$ would then yield an explicit description of the relationship between the remaining design variables, and $\tilde{\epsilon_i}$ at a single Pareto point. Subsequent back-substitution reveals how influential the trade-off with $\tilde{\epsilon_i}$ is upon f_1^*. To study the whole Pareto set, however, a symbolic cost function $U(f_1, \tilde{\epsilon})$ is introduced. $U(f_1, \tilde{\epsilon})$ is monotonically increasing w.r.t. minimisation objectives and decreasing w.r.t. maximisation objectives:

$$\begin{aligned} \text{min.} \quad & \text{U}(f_1^+, \tilde{\epsilon_1}^+, ..., \tilde{\epsilon}_{k-1}^-) && (4.28)\\ & f_1(\mathbf{x}, \tilde{\epsilon_1}^-, ..., \tilde{\epsilon}_{k-1}^+) && (4.29)\\ \text{s.j.t} \quad & \mathbf{g}(\mathbf{x}, \tilde{\epsilon}) \leq 0 && (4.30)\\ & \mathbf{h}(\mathbf{x}) = 0 && (4.31) \end{aligned}$$

In minimising cost, we can exploit its inherent monotonicity w.r.t. the objectives to identify the constraints that *bound* $\tilde{\epsilon}$, and hence affect the topology of the Pareto set. Thus MP1 can be employed to derive the following theorem:

Theorem 3 Boundedness of $\tilde{\epsilon_i}$
In a reduced multiobjective problem, the single objective optimum of a minimisation objective, $\tilde{\epsilon_i}$, is determined by its greatest lower bound. Correspondingly, the lowest upper bound determines the nadir of $\tilde{\epsilon_i}$. As such, the span of the Pareto set is in part determined by $\chi(\tilde{\epsilon})$.

Essentially, each reduced-objective variable is bounded by one or more Pareto-constraints across the objective space. Beyond simple optimization models, they are not necessarily critically constrained. Rather, the optimization of one $\tilde{\epsilon_i}$ will affect the constraints of another, $\tilde{\epsilon_j}$, if their respective glb/lub share variables, or depend on multiple $\tilde{\epsilon}$.

Theorem 4 Conditional Activity of Pareto Constraints
In a set of Pareto-constraints that are conditionally critical for $\tilde{\epsilon_i}$, any constraint, $g_i(\mathbf{x}, \tilde{\epsilon})$, will at least be semi-active w.r.t. $\tilde{\epsilon_i}$ somewhere in the objective space, if it is dependant on $\underline{\overline{x}}$ or more than one reduced-objective variable. That is, unless there exists a Pareto constraint g_j such that $g_i(\mathbf{x}, \tilde{\epsilon}) < g_j(\mathbf{x}, \tilde{\epsilon}) \leq 0$ for any feasible value of $\tilde{\epsilon}$.

The implication here is, that changes in constraint activity can occur across the Pareto set if no $\tilde{\epsilon_i}$ is critically constrained, and no Pareto-constraint is dominant. Identifying these changes in activity reveals how the objectives interact, as exemplified in figure 4.4. Pareto-constraints can take on several forms, that shape the Pareto set in different ways:

- *Bound shift*: A Pareto constraint can for example shift the extremum of a monotonic variable, in effect making it a trade-off variable. Consider a problem where $f_1(x_1^+, x_2^-, \tilde{\epsilon_1}^-)$, and one of the constraints is $g_i(x_2^+, \tilde{\epsilon_1}^-) \equiv 0$. As $\tilde{\epsilon}_1 \rightarrow 0$, the lub of x_1 shifts downward, worsening the optimum of f_1. Thus, g_i makes x_1 a trade-off variable w.r.t. f_1 and $\tilde{\epsilon_1}$, with argmin$\{\tilde{\epsilon_1}, x_2 \in \chi\} = \underline{x_2}$.
- *Inconsistency by* ϵ: Pareto constraints can narrow the feasible domain of design variables that are bounded from above and below. Consider a problem with $U(f_1^+, \tilde{\epsilon_1}^+, \tilde{\epsilon_2}^-)$ where a variable x_1 is bounded from above by $g_1(x_1^+, \tilde{\epsilon_1}^-) \leq 0$ and from below by $g_2(x_1^-, \tilde{\epsilon_2}^+) \leq 0$. As $\tilde{\epsilon} \rightarrow \tilde{\epsilon}^*$, the feasible domain for x is reduced, meaning g_1 and g_2 become inconsistent beyond the Pareto set. Hence, g_1 and g_2 reduce objective alignment between $\tilde{\epsilon_1}$ and $\tilde{\epsilon_2}$, with one becoming semi-active at the resulting bi-objective Pareto frontier.
- *Multiple objectives*: Pareto constraints that depend on multiple $\tilde{\epsilon_i}$ drastically reduce objective alignment, for instance if a constraint takes the form $g_1(\mathbf{x}, \tilde{\epsilon_1}^+, \tilde{\epsilon_2}^-)$.

Hence, trade-offs between the reduced-objectives are apparent in the Pareto-constraints themselves. An objective pair, $\tilde{\epsilon_i}$ and $\tilde{\epsilon_j}$, is in trade-off if they share a constraint of the form $g(\mathbf{x}, \tilde{\epsilon_i}, \tilde{\epsilon_j})$ or if their constraints become inconsistent w.r.t. to a shared variable, x, when $\tilde{\epsilon} \rightarrow \tilde{\epsilon}^*$. Such constraints therefore require special attention.

4.3 Trade-off Root-cause Analysis

With the developments to monotonicity analysis presented in the previous section, we now have a rigorous foundation for the study of the dependencies that exist between design objectives. These dependencies may exist throughout the the attainable set $\mathcal{A}$, be specific to the entire Pareto set $\mathcal{C}$, specific to a region of $\mathcal{C}$, or to single Pareto points.

This allows us to study the topology of the Pareto set; i.e. how it is *shaped* and positioned by shared variables and constraints, thereby revealing the root-causes of the trade-offs in a design problem. Yet, this requires systematic reduction of multi-objective problems down to a point where this information is reached. In the following, an overview of the information gained through model reduction is given, followed by an overall analysis procedure, and perspectives on how to support the mostly manual analysis procedure with data from numerical solution of the optimization problem.

4.3.1 Drivers of Trade-off

The model reductions permitted by MOMA, allow independent variables (i.e. variables that only affect one objective) and harmonious variables to be optimized out of the problem. Thereby, the objectives in the problem are either all partially minimized simultaneously, or they are partially minimized individually without any impact on the others. In this process, the reduced model gradually converges towards being an explicit description of the Pareto set.

A multi-objective model can be reduced continually, so long as additional active constraints can be identified, and harmonious or independent variables still exist in the problem. Trade-off variable cannot be optimized out without eliminating a part of the Pareto set, unless it is used to eliminate a bound objective, following the ϵMA approach.

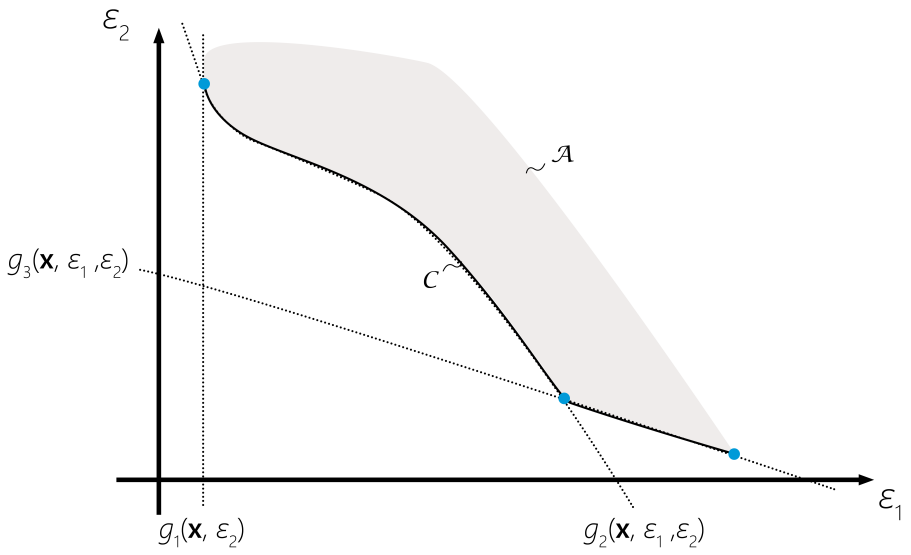


Figure 4.4: *From Paper A*: An example of how the topology of a Pareto set is affected by Pareto constraints. Here the optima of ϵ_1 and ϵ_2 are determined by g_1 and g_3 respectively, with the multiobjective Pareto constraint, g_2 further reducing objective alignment

This can lead to a sequence of back-substitutions introducing new variables or contributions into the objectives and remaining constraints, causing some of the remaining constraints to become active (e.g. due to critically), in turn allowing further back-substitutions. In this process, trade-off variables not present in the original model come to light, being introduced by active constraints throughout the reduction process.

Beyond the benefits in reduced computational cost and model verification, MOMA can hence be used with the aim of identifying these specific variables and active constraints, which in effect cause the trade-offs in the problem. Some of these trade-off variables can be introduced due to the activity of multiple constraints, all of which will in effect be a part of the cause of the resulting trade-off between the objectives. Thus, we are interested in reducing multi-objective models as *far down* as possible, to reveal most or all of these relationships, with each reduction step revealing more information. In principle, the model can be reduced until all of the remaining variables are either trade-off variables, or variables for which an active constraint cannot be identified (e.g. due to non-monotonicity or conditional criticality). This is illustrated in figure 4.5.

This information can then be leveraged by the designer to identify design changes that result in an improved Pareto set; this will be covered in chapter 5. In principle, there further down in reduction a trade-off variable is introduced, the more information the designer gains about approaches to changing the design, to improve the Pareto-set. Not only can they attempt to make the objectives independent of the initial variable which introduced a trade-off variable when it was substituted, they can also attempt to change the design in a way that alters or eliminates one or more of the active constraints that led to the back-substitution. Consider the example from section 4.2.2:

$$\text{min.} \qquad f_1(x_1^+, x_2^-, x_3^+) = x_1^2 - x_2 + x_3 \tag{4.32}$$

$$\text{s.j.t} \qquad c_1(x_2^-, x_4^-, x_5^+; \epsilon_1^-) = \frac{1}{x_2} - x_4^2 + 2x_5 - \epsilon_1 \leq 0 \tag{4.33}$$

$$g_1(x_1^-, x_4^+) = 2x_4 - x_1 \leq 0 \tag{4.34}$$

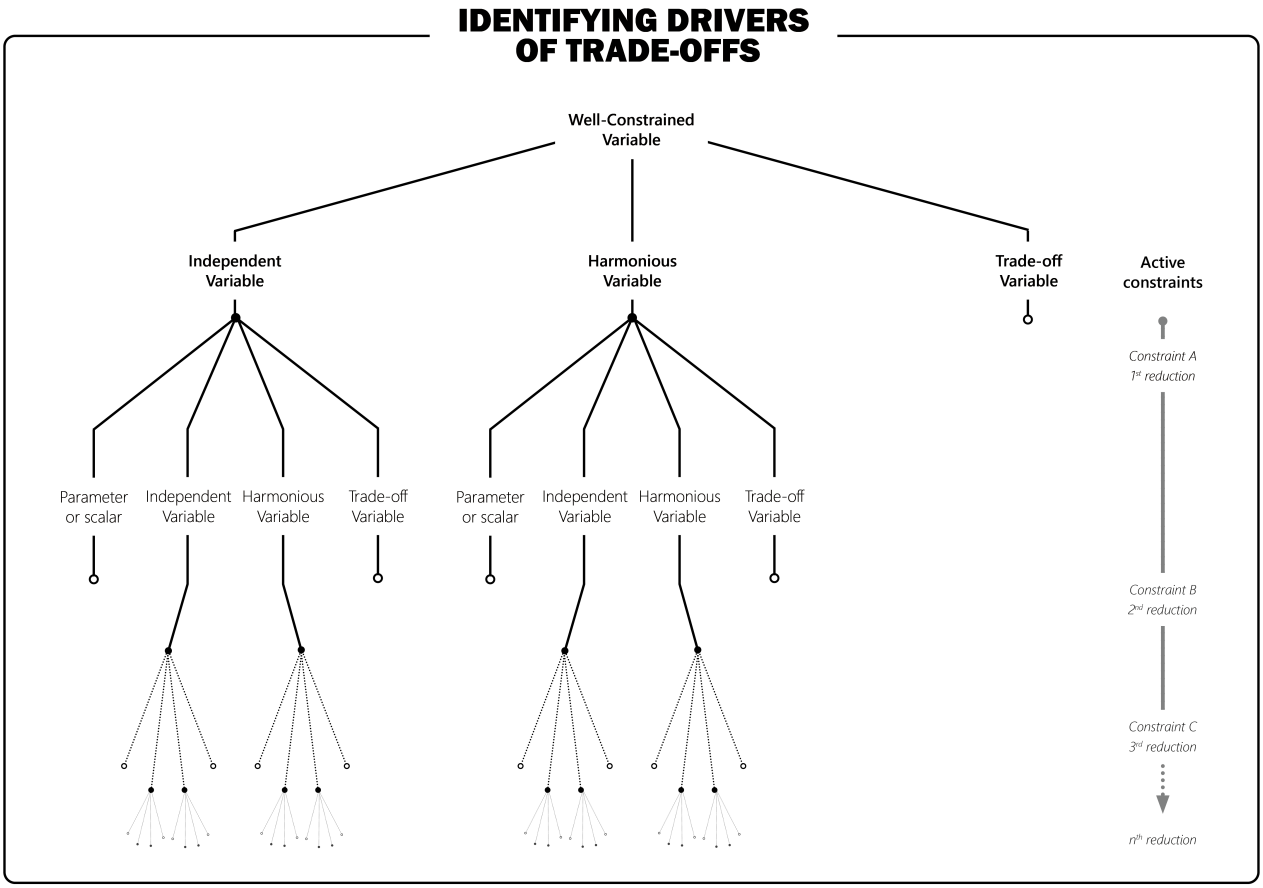


Figure 4.5: Each model reduction can introduce different types of contributions into the objective and constraint functions. If new independent variables or harmonious variables are introduced into the objective(s), further reductions can be performed if active constraints can be identified. As a result, trade-off variables between objectives can hide behind several *layers* of active constraints.

$$g_2(x_2^+, x_3^-) = x_2^2 + 4x_2 - 2x_3 \leq 0 \tag{4.35}$$
$$g_3(x_2^+, x_4^+) = x_2^3 + 2x_4 - P_1 \leq 0 \tag{4.36}$$
$$g_4(x_5^-) = 10 - x_5^2 - 3x_5 \leq 0 \tag{4.37}$$

which took the following form after reduction:

$$\text{min.} \qquad f_1(x_2^+, x_4^+) = 4x_4^2 + \frac{1}{2}x_2^2 + x_2 \tag{4.38}$$
$$\text{s.j.t} \qquad c_1(x_2^-, x_4^-; \epsilon_1^-) = \frac{1}{x_2} - x_4^2 + 4 - \epsilon_1 \leq 0 \tag{4.39}$$
$$g_3(x_2^+, x_4^+) = x_2^3 + 2x_4 - P_1 \leq 0 \tag{4.40}$$

In the unreduced problem, neither x_2 nor x_4 seemed to be trade-off variables. The trade-off between f_1 and c_1 is caused by the activity of g_1 and g_2, which led to x_1 and x_3 being optimized out. If this were a design situation, the designer could thus explore several alternative design changes that would result in beneficial model transformations, to reduce or eliminate the trade-off between f_1 and the bound objective c_1:

1. Make f_1 independent of x_1 and/or x_3
2. Reduce the gradient of f_1 w.r.t x_1 and/or x_3
3. Make g_1 independent of x_4
4. Make g_2 independent of x_2
5. Reduce the derivative of g_1 w.r.t x_4
6. Reduce the derivative of g_2 w.r.t x_2

Altering c_1 would not necessarily result in an improved optimal set, given that c_1 is monotonically decreasing w.r.t. the x_2 and x_4, both before and after reduction. As this demonstrates, there are more potential options for improvement of the Pareto set, than if the two objectives had shared a trade-off variable in their unreduced form.

4.3.2 Analysis and Reduction Procedure

Applying the MOMA and ϵ-monotonicity theorems to multiobjective optimization problems allows systematic reduction down to a point where the dependencies that exist in the Pareto set are revealed. The root causes of these dependencies are, from a design perspective, the constraints and shared variables that create said dependencies. Thus, if we systematically reduce multiobjective problems and make a note of trade-off variables, the constraints that introduce them, and the constraints that bound the Pareto set, we find the relationships that in effect create, shape, and position the Pareto set.

To reach this point of reduction however, it is necessary to follow a systematic analysis procedure, to ensure that the largest degree of reduction is achieved, without accidentally eliminating trade-off variables from the problem. Furthermore, given that regional and local analysis can become necessary, especially in ϵMA, case analysis is an important part of the analysis process. The steps in the suggested analysis process, build upon upon monotonicity analysis as developed by Papalambros and Wilde [12, 45]. These are as follows:

1. Model the multiobjective problem as an upper-bound formulation in negative null form.
2. Set up a monotonicity table, and assess the monotonicity of the objectives and constraints w.r.t. the design variables. Make note of any trade-off variables.
3. Use monotonicity analysis procedures to assess whether the model is well bounded [12], with the addition of the special case of the well-boundedness of trade-off variables. If the model is not well bounded, add constraints.
4. Identify constraints that are active w.r.t the primary objective and use them to reduce the model. Make note of constraints that introduce new trade-off variables. If possible, identify the conditions under which the bound objectives become active, following Theorems 1 and 2.
5. Partially minimise the bound objectives when no further reductions to the primary objective can be made. Take care not to use constraints that potentially bound other variables regionally in the objective space. Make note of constraints that introduce new trade-off variables.
6. When the remaining variables are either trade-off variables, non-monotonic or bounded by a conditionally critical set of constraints, run the optimization model.
7. If the numerical results reveal further globally active constraints, make further model reductions.

8. If any bound objectives are globally active, optimize said objectives out, eliminating trade-off variables in the process. The ϵ parameters will now appear in the remaining constraints and objective functions.

9. Treat ϵ parameters of the eliminated bound objectives as variables and identify the constraints that bound them. In the presence of conditional critical Pareto constraints, decompose the problem into *Pareto-Optimal Activity Cases*. Identify the values of ϵ that cause change in constraint activity or make specific constraints inconsistent. Verify this against the numerical results. Either do this exhaustively, or use the numerical results to identify cases of interest.

Box 1 - Analysis of Pareto-optimal Activity Cases (*From Paper A*)

Step 1 - Case identification

To minimise U$(f_1^+, \tilde{\epsilon})$, identify the conditionally critical set of Pareto constraints for each $\tilde{\epsilon}_i$. For each $g_j(\mathbf{x}, \tilde{\epsilon}_i)$ that is conditionally critical w.r.t. $\tilde{\epsilon}_i$:

1. Assume $g_j(\mathbf{x}, \tilde{\epsilon}_i) \equiv 0$, and solve w.r.t. $\tilde{\epsilon}_i$.
2. Identify the constraints that become active as a consequence of $\tilde{\epsilon}_i \rightarrow \tilde{\epsilon}_i{}^* \wedge g_j(\mathbf{x}, \tilde{\epsilon}_i) \equiv 0$ and use this to reduce the expression $\tilde{\epsilon}_i{}^*(\mathbf{x}, \tilde{\epsilon})$.
3. Back substitute the eliminated variables into the remaining constraints, including the Pareto constraints that bound other reduced objective variables. If possible, identify the glb and lub of $\tilde{\epsilon}_l, \forall l \neq i$ and use it to solve for $\underline{\tilde{\epsilon}_l}$.

Step 2 - Case elimination

Compare the terms for $\tilde{\epsilon}_i{}^*$ from each case:

1. If any case *j* is dominant, i.e. $\tilde{\epsilon_{i,j}}{}^* > \tilde{\epsilon_{i,k}}{}^*$ for any feasible value of $\mathbf{x}$ and $\tilde{\epsilon}_l \, , \forall l \neq i$ then $g_k(\mathbf{x}, \tilde{\epsilon})$ is either inactive or bounds another variable.
2. If any variable is revealed to be unbounded as a consequence of $g_j(\mathbf{x}, \tilde{\epsilon}_i) \equiv 0$, then the problem is either not well-constrained, or g_j is never critical w.r.t. $\tilde{\epsilon}_i$, meaning the case can be disregarded.
3. Identify the conditions under which the remaining cases become active. If feasible values of $\mathbf{x}$ and $\tilde{\epsilon}_l \, , \forall l \neq i$ exist such that two cases become equivalent, i.e. $\tilde{\epsilon_{i,j}}{}^* = \tilde{\epsilon_{i,k}}{}^*$ then g_j and g_k are locally active in the objective space, with a change in activity occurring at $\tilde{\epsilon_{i,j}}{}^* = \tilde{\epsilon_{i,k}}{}^*$. Such points are vertices of the Pareto set.

Step 3 - Case reduction

Reduce the remaining cases further to identify the extrema of the Pareto set:

1. Further minimise $\tilde{\epsilon}_i{}^*(\mathbf{x}, \tilde{\epsilon})$ by optimizing trade-off variables out, letting $\underline{\overline{x}} \rightarrow \{\underline{x}$ if $\tilde{\epsilon}_i(x^+), \overline{x}$ if $\tilde{\epsilon}_i(x^-)\}$. If the glb and lub of $\overline{x}$ cannot be determined, the problem case can be split into sub-cases.
2. If possible, identify the cases that yield utopia and nadir points for each objective

Following Theorem 4, the bounds of $\tilde{\epsilon}$ can be interdependent, meaning that the minimisation of $\tilde{\epsilon}_i$ affects the bounds of the remaining $\tilde{\epsilon_j} \, , \forall j \neq i$, and $\underline{\overline{x}}$, causing changes in activity across the Pareto set. Each change in activity implies local dependencies between the objectives in regions of the Pareto set, as illustrated in Fig. 4.4. Each potential combination of active Pareto constraints hence represents a unique *Pareto Efficient Activity Case*. One can either exhaustively study all cases, or focus the analysis procedure upon cases of interest. The case analysis procedure, demonstrated on a problem with minimisation objectives, is shown in Box 1. It closely resembles the parametric solution procedure developed by Wilde [12], albeit for objectives instead of design parameters. The analysis of Pareto optimal activity case is

demonstrated at the end of this chapter.

Beyond potentially deriving single-objective optima, this procedure can be used to explicitly derive trade-off functions of the forms $f_1(\mathbf{x}, \tilde{\epsilon})$ and $\tilde{\epsilon_i}^*(\mathbf{x}, \tilde{\epsilon})$. As a consequence of Theorem 1, these equations actually describe the Pareto set prior to the elimination of $\underline{\overline{\mathbf{x}}}$, as any feasible value of a monotonic trade-off variable yields a Pareto point. If an objective pair $\tilde{\epsilon_i}$ and $\tilde{\epsilon_j}$ is in trade-off, then these reduction steps will inevitably yield minima of the form $\tilde{\epsilon_i}^*(\mathbf{x}, \tilde{\epsilon_j}^-)$ and $\tilde{\epsilon_j}^*(\mathbf{x}, \tilde{\epsilon_i}^-)$, or of the form $\tilde{\epsilon_i}^*(\mathbf{x})$ and $\tilde{\epsilon_j}^*(\mathbf{x})$, where $\underline{\overline{\mathbf{x}}} \subset \mathbf{x}$.

Pareto-constraints that become inconsistent beyond the Pareto set are revealed as the bound objectives are optimized out. In simple problems, this degree of reduction might be reached through algebraic manipulations alone. For complex problems, however, full reduction might not be worthwhile due to the algebraic effort. Here, one can employ a more pragmatic approach by utilizing numerical results to identify additional active constraints that can be used to reduce the model further post optimality.

4.3.3 Interactive Computation and Analysis

As might be evident, the analysis procedure is mostly manual. In some situations it can be quite arduous to perform all of the analysis manually. Luckily, there are ways in which numerical solution of the optimization problem, can be used to support the identification of additional active constraints. For this reason, the trade-off root cause analysis procedure includes a step involving numerical solution, and subsequent model reduction. Furthermore, numerical solution can also support the interpretation of the results of the trade-off analysis.

Numerical Identification of Active Constraints

If numerical solution reveals constraints that fulfill the Global Activity criterion from Definition 2, then such constraints can essentially be dealt with exactly as with constraints that are found to be active through MA. This would require that a constraint g_i has an associated Lagrange multiplier $\lambda_i > 0$ in all feasible iterations. As outlined in the pseudo-code in section 3.1, this does require that the Lagrange multipliers are stored for each Pareto point.

Any constraint that meets this condition can thus be used to reduce the model further, back-substituting variables, thereby, giving a clearer picture of the relationships that exist at the Pareto set. In principle, one could rely on numerical results to identify all of the globally active constraints, after constructing the model and setting up the monotonicity table. Here, it would be useful to update the numerical model to reflect the reductions made. This would allow the evaluation of the effect of the reductions upon the remaining constraints.

This approach can also be applied to active nonlinear constraints that have no closed form solution. Eliminating these constraints can make subsequent reductions difficult, as the introduction of implicit terms into the remaining constraints can make it impossible to identify the dominant constraint among a set of potentially active constraints. Here, numerical solution of the model can help reveal the subsequent active constraints, allowing further reduction.

Constraint Violation and Pareto constraints

Numerical solution can also help identify which Pareto constraints shape the Pareto set. This can reduce the need to studying numerous Pareto optimal Activity Cases. By definition, one or more of the Pareto constraints will be violated just beyond the Pareto set; the closer to the utopia point, the more inconsistent the constraints become. Identifying which constraints are violated "first" - i.e. closest to the Pareto set, can help identify the Pareto constraints that bound the Pareto set.

This can be done by exploiting an oft discussed limitation of the ϵ-constraint method. Compared to other formulations, the ϵ-constraint method is computationally inefficient[54], as it "wastes" computational iterations, for values of ϵ which lie between the utopia point and

the Pareto set. Furthermore, if the limits of ϵ, $\epsilon_{\mathbf{L}}$ and $\epsilon_{\mathbf{U}}$ are wider than the attainable set $\mathcal{A}$, prior to computation, the model will fail to identify feasible solutions in cases where the enforced objective bounds result in inconsistent constraints, meaning $\chi = \{\}$. The larger the discrepancy the more computational effort is seemingly wasted. Identifying suitable bounds a-priori to computation is not necessarily trivial. Yet, for the purposes of trade-off root cause analysis, and the identification of active Pareto constraint, these iterations are not necessarily wasted. In fact, they can help explain why the region of the objective space between the utopia point and the Pareto set is not attainable.

There are a few prerequisite to doing so. First, it is necessary to verify that the iterations that yielded infeasible results, did so due to inconsistency between constraints for specific values of ϵ, and not due to computational issues. One could for instance run the solve a constraint satisfaction problem for the iterations that yielded infeasible solutions. In many optimization engines, this is equivalent to setting disregarding the primary objective, setting $f(\mathbf{x}) = 0$ and running the optimization problem. One could also utilise global optimization methods such as NSGA-ii, or a multi-start approach, to assess whether a poor initial guess or a discontinuous attainable set the cause of a lack of convergence.

Secondly, the use of a solver that allows exploration of infeasible regions, exhibits global convergence to feasibility when provided an infeasible initial guess, and can recover from infeasibility during iteration, is also a prerequisite. An example of such would be the use of a sequential quadratic programming (SQP) algorithm and a penalty based merit function [119, 120] (e.g. the SQP implementation in MATLABs *fmincon* routine):

$$\phi(\mathbf{x}, \lambda, \mu) = f(\mathbf{x}) + \sum_{j=1}^{l} w_j |h_j| + \sum_{j=1}^{m} w_j |\min(0, -g_j)| \tag{4.41}$$

where w_j are penalty weights used to minimise constraint violation while minimising the objective. This ability to minimise violation, allows exploration of the constraint inconsistencies that occur beyond the Pareto set. If no feasible solution exists, the algorithm will converge towards a minimal constraint violation, providing convergence is achieved before the max. n.o. iterations or the min. step size is reached. For a given computational iteration *i* where $\chi = \{\}$, the minimisation of ϕ will hence result in an infeasible point which is locally as close to feasibility as possible. As such, one can look at the outputs from numerical solution - $\mathbf{F}^*(:, i)$, $\mathbf{G}^*(:, i)$, $\mathbf{H}^*(:, i)$, and $\Lambda(:, i)$ - to assess the relationship between the values of ϵ and the violated constraints for a given solution iteration, *i*.

If done exhaustively across the objective space, one therefore automatically gains data on how the Pareto set is shaped by the violation of certain constraints. This does not result in much additional computational effort, and can help select Pareto optimal activity cases of interest.

In the SOMA case study, the SQP implementation in MATLAB R2019a *fmincon* routine is used, which utilises a globally convergent algorithm for handling inconsistent QP problems in SQP. The algorithm is based on work by Spellucci [120], and involves recovering from poor initial guesses using slack variables, and identifying descent directions for ϕ that restore feasibility by relaxing constraints that are violated in inconsistent QP problems. Alternatively, one could employ the similar SQP implementation developed by Burke [121]. Beyond the identification of local optima, it is aimed at allowing the identification of *external* stationary points, which are candidate optima that are as *close* to feasibility as possible.

4.4 Analysis of the SOMA Device

In order to apply the MOMA and ϵMA methods developed in the preceding sections, we first need to transform the model described in chapter 3, into ϵ-constraint form:

$$\text{min} \quad f_1(\mathbf{x}) = -\frac{\sum_{p=1}^{p=8} m_p \cdot (C_p + Z_p)}{(l_{t1} + l_{t2} + l_{b1}) \cdot \sum_{p=1}^{p=8} m_p} \tag{4.42}$$

$$\text{s.j.t} \quad c_1(\mathbf{x}; \epsilon_1) = d_{t1} - \epsilon_1 \leq 0 \tag{4.43}$$

$$c_2(\mathbf{x}; \epsilon_2) = \epsilon_2 - \rho \frac{\pi}{4} d_{n1}^2 \left(l_{n1} + \frac{1}{3} \cdot l_{n2} \right) \leq 0 \tag{4.44}$$

$$c_3(\mathbf{x}, \epsilon_3) = \epsilon_3 - \sqrt{2 z_{acc} \left(g_C + \frac{z_{pre} G_{st} d_{ps2}^4}{8 d_{ps1}^3 n_a m_{acc}} - \frac{z_{acc} \left(1 + \frac{m_{acc}}{m_{rec}}\right) G_{st} d_{ps2}^4}{16 d_{ps1}^3 n_a m_{acc}} \right)} \leq 0 \tag{4.45}$$

$$\mathbf{g}(\mathbf{x}) \leq 0 \tag{4.46}$$

$$\mathbf{h}(\mathbf{x}) = 0 \tag{4.47}$$

$$\mathbf{x}, \epsilon \in \mathbb{P} \tag{4.48}$$

where $f_i = c_{i-1}$ for $i > 1$. As the device size objective is a minimisation objective, while API payload and impact speed are maximisation objectives, they take different forms, when transformed into bound objectives. This form allows us to apply the theorems and reduction procedures.

4.4.1 Pre-Optimality Analysis and Reductions: MOMA

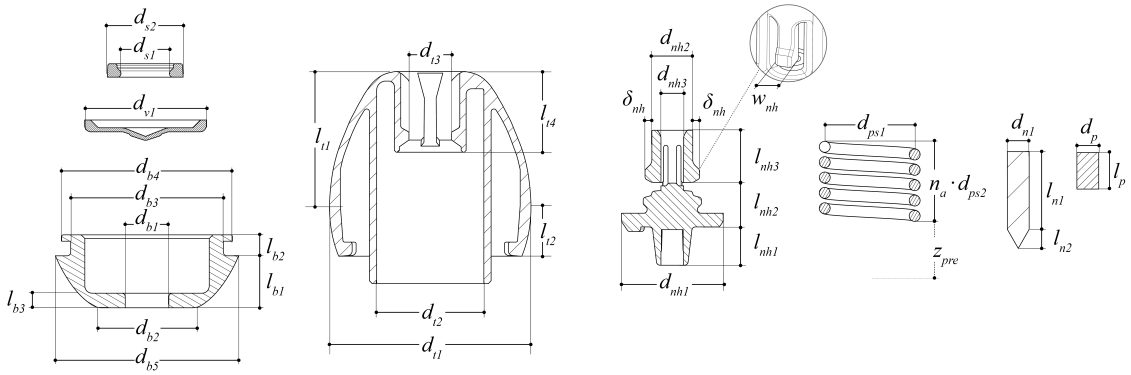


Figure 4.6: *From Paper A*: An overview of the key design variables in the SOMA device, included again in this chapter for the sake of readability

As mentioned, the model is too large to include in this dissertation in its entirety. Correspondingly, much of its reduction through MOMA involves both extensive and somewhat tedious algebraic analysis. For the sake of readability, only the constraints and variables that contribute to the the design issues ultimately revealed by the trade-off root cause analysis, will be included in the demonstration of the methods developed in this chapter. The shear amount of variables and constraints involved in the full model, and all of the reduction steps involved, are too expansive to include here. Furthermore, the analysis the monotonicity of each objective and constraint w.r..t each variable, is left out as this involves well described

analysis procedures. For examples of such, see Papalambros [12, 58, 68], Azarm et al [69], and Williams & Cagan [122].

Using MOMA, the model was reduced prior to computation down to 18 design variables, 28 inequality constraints, 2 equality constraints and 4 objectives, 3 of them bound. The constraints of interest to the trade-off analysis fall in three categories. Firstly, there are the radial fits between the parts, ensuring that the parts fit together radially - i.e., that the plug fits into the hub, the hub into the top housing, the spring around the trigger system, and the top housing in the axial fits between the parts. These are the constraints h_2, h_3 and g_1 to g_{10}. Secondly, there are two stress constraints; a creep stress limit for the load bearing trigger interface (g_{11}), and a spring yield limit (g_{12}) which is handled implicitly below due to the size of the actual equation. Finally, there are the axial constraints h_8, and $g_{20} - g_{25}$ which ensure that the needle has sufficient clearance to the valve before injection, that the needle hub and valve are manufacturable, and that there is a sufficient amount of dead windings in the spring.

$$h_1(l_{t1}^+, d_{t1}^-) = l_{t1} - d_{t1}C_T = 0 \tag{4.49}$$

$$h_2(d_{nh3}^+, d_p^-) = d_{nh3} - d_p - 2R_{cl} = 0 \tag{4.50}$$

$$h_3(d_{b4}^+, d_{b5}^-) = d_{b4} + 2R_{wt} + 2R_{cl} - d_{b5} = 0 \tag{4.51}$$

$$\begin{aligned} &h_8(n_a^+, n_d^+, d_{ps2}^+, l_{nh1}^+, l_{nh2}^+, l_{n1}^+, l_{n2}^+, z_{acc}^+, l_{b1}^-, l_{t2}^-, l_{t1}^-) \\ &= R_{wt} + (n_a + n_d)d_{ps2} + l_{nh2} + l_{nh1} + l_{n1} + l_{n2} + z_{acc} - l_{b1} - l_{t2} - l_{t1} = 0 \end{aligned} \tag{4.52}$$

$$g_1(d_{t1}^-, l_{t1}^-, l_{t2}^+, d_{b5}^+) = d_{b5} - \sqrt{\frac{2(l_{t1}-l_{t2})d_{t1}^2}{l_{t1}} - \frac{(l_{t1}-l_{t2})^2 d_{t1}^2}{l_{t1}^2}} \leq 0 \tag{4.53}$$

$$g_2(d_{b3}^+, d_{b4}^-) = d_{b3} + 2R_{ov} - d_{b4} \leq 0 \tag{4.54}$$

$$g_3(d_{t2}^+, d_{b3}^-) = d_{t2} + 4R_{wt} + 2R_{cl} - d_{b3} \leq 0 \tag{4.55}$$

$$g_4(d_{t2}^-, d_{ps1}^+, d_{ps2}^+) = d_{ps1} + d_{ps2} + 2R_{cl} - d_{t2} \leq 0 \tag{4.56}$$

$$g_5(d_{t3}^+, \delta_{nh}^+, d_{ps2}^+, d_{ps1}^-) = d_{t3} + 2(\delta_{nh} + R_{cl} + R_{wt}) + d_{ps2} - d_{ps1} \leq 0 \tag{4.57}$$

$$g_6(d_{nh2}^+, d_{t3}^-) = d_{nh2} - d_{t3} + 2R_{cl} \leq 0 \tag{4.58}$$

$$g_7(d_{nh2}^-, d_{nh3}^+) = d_{nh3} + 2R_{wt} - d_{nh2} \leq 0 \tag{4.59}$$

$$g_8(\delta_{nh}^-) = 0.3\text{mm} - \delta_{nh} \leq 0 \tag{4.60}$$

$$g_9(d_p^-) = 1\text{mm} - d_p \leq 0 \tag{4.61}$$

$$g_{10}(\delta_{nh}^+, d_{nh3}^-) = 2\delta_{nh} + R_{cl} - d_{nh3} \leq 0 \tag{4.62}$$

$$g_{11}(d_{ps1}^-, d_{ps2}^+, n_a^-, z_{pre}^+, \delta_{nh}^-, w_{nh}^-) = \frac{z_{pre}\cos(\Theta_{nh})G_{st}d_{ps2}^4}{16d_{ps1}^3 n_a \delta_{nh} w_{nh}} - \sigma_c \leq 0 \tag{4.63}$$

$$g_{12}(d_{ps1}^-, d_{ps2}^+, n_a^-, z_{pre}^+; \sigma_y) \leq 0 \tag{4.64}$$

$$g_{20}(l_{b3}^+, l_{v1}^+, z_{acc}^-) = l_{b3} + l_{v1} + Z_{tol} - z_{acc} \leq 0 \tag{4.65}$$

$$g_{21}(l_{nh1}^-) = 1.5\text{mm} - l_{nh1} \leq 0 \tag{4.66}$$

$$g_{22}(l_{nh2}^-) = \frac{5}{2}R_{wt} - l_{nh2} \leq 0 \tag{4.67}$$

$$g_{23}(l_{v1}^-) = R_{wt} - l_{v1} \leq 0 \tag{4.68}$$

$$g_{24}(n_d^-) = 1.5 - n_d \leq 0 \tag{4.69}$$

$$g_{25}(w_{nh}^+, d_p^-) = w_{nh} + R_{cl} + 3/2R_{st} - d_p \leq 0 \tag{4.70}$$

where C_T is the aspect ratio between the diameter and height of the top housing, R_{wt} is the min. wall thickness, R_{ov} the min. radial interface overlap, R_{cl} the min. radial clearance, Θ_{nh} the contact angle in the trigger interface, G_{st} the shear modulus of the spring steel, and σ_c

the allowable stress in the trigger interface, σ_y is the allowable stress in the spring, and n_d is the number of dead windings in the spring. The variables are illustrated in fig. 4.6. The monotonicity of the objectives is shown in the partial monotonicity table of the optimization problem, seen in table 4.1. As the table is partial, some variables may seem to be critically constrained without actually being so.

The only trade-off variables that are visible so far, are the device diameter, d_{t1}, the spring wire diameter d_{ps2}, the length of the needle and its tip, l_{n1} and l_{n2}, as exhibited by their opposite monotonicity in the objectives. To begin reducing the problem, we first use h_1 to eliminate l_{t1}, h_2 to eliminate d_{nh3}, h_3 to eliminate d_{b5}, and h_8 to eliminate z_{acc}. Furthermore, several inequality constraints are critical; g_2 is critical wrt. d_{b4}, g_3 w.r.t d_{b3}, g_4 w.r.t d_{t2}, g_6 w.r.t. d_{t3}, g_{21} w.r.t. l_{nh1}, g_{22} w.r.t. l_{nh2}, g_{23} w.r.t. l_{v1}, and g_{24} w.r.t. n_a. Following Definition 1 and Theorem 1, none of these are trade-off variables, meaning we can apply MP1 to eliminate these variables. After back-substitution, the objectives and constraints have changed. This results in the following changes, with some functions being shown implicitly, due to their length:

$$\text{min.} \quad f(l_{t2}^-, d_{t1}^-, l_{b1}^-, d_{t3}^+, d_{ps1}^+, d_{ps2}^+, d_{nh2}^+, \delta_{nh}^+, d_p^+, w_{nh}^+, l_{n1}^+, l_{n2}^+) \tag{4.71}$$

$$\text{s.j.t.} \quad c_1(d_{t1}^+; \epsilon_1) = d_{t1} - \epsilon_1 \leq 0 \tag{4.72}$$

$$c_2(d_n^-, l_{n1}^-, l_{n2}^-; \epsilon_2) = \epsilon_2 - \rho\frac{\pi}{4}d_{n1}^2\left(l_{n1} + \frac{1}{3}\cdot l_{n2}\right) \leq 0 \tag{4.73}$$

$$c_3(d_{t1}^-, l_{t2}^-, l_{b1}^-, d_{ps1}^+, d_{ps2}^-, d_{nh2}^+, \delta_{nh}^+, l_{n1}^+, l_{n2}^+; \epsilon_3) \tag{4.74}$$

$$g_1(d_{t1}^-, l_{t2}^+, d_{ps1}^+, d_{ps2}^+) = d_{ps1} + d_{ps2} + 6R_{wt} + 6R_{cl} + 2R_{ov} - \sqrt{\frac{2(C_T d_{t1} - l_{t2})d_{t1}}{C_T} - \frac{(C_T d_{t1} - l_{t2})^2}{C_T^2}} \leq 0 \tag{4.75}$$

$$g_5(d_{nh2}^+, \delta_{nh}^+, d_{ps2}^+, d_{ps1}^-) = d_{nh2} + 2\delta_{nh} + d_{ps2} - d_{ps1} + 4R_{cl} + 2R_{wt} \leq 0 \tag{4.76}$$

$$g_7(d_{nh2}^-, d_p^+) = d_p + 2R_{wt} + 2R_{cl} - d_{nh2} \leq 0 \tag{4.77}$$

$$g_8(\delta_{nh}^-) = 0.3\text{mm} - \delta_{nh} \leq 0 \tag{4.78}$$

$$g_9(d_p^-) = 1\text{mm} - d_p \leq 0 \tag{4.79}$$

$$g_{10}(\delta_{nh}^+, d_p^-) = 2\delta_{nh} - R_{cl} - d_p \leq 0 \tag{4.80}$$

$$g_{11}(d_{ps1}^-, d_{ps2}^+, n_a^-, z_{pre}^+, \delta_{nh}^-, w_{nh}^-) = \frac{z_{pre}\cos(\Theta_{nh})G_{st}d_{ps2}^4}{16d_{ps1}^3 n_a \delta_{nh} w_{nh}} - \sigma_c \leq 0 \tag{4.81}$$

$$g_{12}(d_{ps1}^-, d_{ps2}^+, n_a^-, z_{pre}^+; \sigma_y) \leq 0 \tag{4.82}$$

$$g_{20}(l_{b3}^+, n_a^+, d_{ps2}^+, l_{n1}^+, l_{nh2}^+, d_{t1}^-, l_{t2}^-, l_{b1}^-) = l_{b3} + (n_a + 1.5)d_{ps2} + l_{n1} + l_{n2}7/2R_{wt} + Z_{tol} + 1.5\text{mm} - C_t d_{t1} - l_{t2} - l_{b1} \leq 0 \tag{4.83}$$

$$g_{25}(w_{nh}^+, d_p^-) = w_{nh} + R_{cl} + 3/2R_{wt} - d_p \leq 0 \tag{4.84}$$

So far, trade-offs have been revealed between size, c_1, and both impact velocity, c_3, and self-orientation, f_1, through d_{t1}, and between impact velocity and self-orientation through the spring wire diameter d_{ps2}. Increasing the wire diameter increases spring force and hence velocity, but it also increases the spring mass, shifting the system centre of mass upward. Here it is also worth noticing, that the elimination of z_{acc} and l_{t1} has introduced trade-off

	l_{t1}	l_{t2}	d_{t1}	d_{t2}	d_{t3}	l_{b1}	l_{b3}	d_{b3}	d_{b4}	d_{b5}	d_{ps1}	d_{ps2}	n_a	n_d	z_{pre}	l_{n1}	l_{n2}	z_{acc}	l_{nh1}	l_{nh2}	d_{nh2}	d_{nh3}	δ_{nh}	w_{nh}	d_p	l_{v1}
f_1 - Center of mass	-	-	-	+	+	-	-	+	+		+	+	+	+		+	+		+	+	+	-	+	+	+	+
c_1 - Device diameter			+																							
c_2 - API payload																-	-									
c_4 - Impact velocity											+	-	+		-			-	+	+	+	-	+	+		
h_1 - Device Height/Width	+		+																							
h_2 - Plug fit																						+			-	
h_3 - Base cylinder width									+	-																
h_8 - Acceleration stroke	-	-				-						+	+	+		+	+	+	+	+						
g_1 - Radial top-base fit	-	+	-							+																
g_2 - Base snap size								+	-																	
g_3 - Base-top cylinder fit				+				-																		
g_4 - Spring-cylinder fit				-							+	+														
g_5 - Spring-trigger fit					+						-	+											+			
g_6 - Trigger arm fit					-																+					
g_7 - Hub wall thickness																					-	+				
g_8 - Min. trigger overlap																							-			
g_9 - Min plug dia.																									-	
g_{10} - Arm collision																						-	+			
g_{11} - Creep stress											-	+	-		+								-	-		
g_{12} - Spring stress											-	+	-		+											
g_{20} - Needle-valve clear.							+											-								+
g_{21} - Needle IF length																			-							
g_{22} - Plate thickness																				-						
g_{23} - Valve Thickness																										-
g_{24} - N.o. dead windings														-												
g_{25} - Arm width																								+	-	

Table 4.1: A partial monotonicity table for the SOMA device

variables into c_3, namely d_{t1}, l_{n1}, and l_{n2}, meaning impact velocity is in trade-off with device size API payload.

Following Theorems 1 and 2, the boundedness of d_{t1} reveals important information about the SOMA device. Firstly, $c_1(\mathbf{x};\epsilon_1) \equiv 0$ for any $\epsilon_L(1) < \epsilon_1 < \epsilon_U(1)$. Secondly, the only *non-objective lower bound* for d_{t1} is g_1, the constraint that ensures that the top and base housings fit together radially. This means that g_1 will be active at the single-objective minimum. Looking at eq. 4.75, it is evident that all the objectives cannot be minimised simultaneously, without reaching a point where $c_1(d_{t1}^+) < g_1(d_{t1}^-)$ meaning that $\mathcal{X}(d_{t1}) =$. Hence, g_1 is at least semi-active in any bi-objective Pareto-front involving the size objective, c_1. As a consequence, l_{t2} is a trade-off variable when g_1 is active, and d_{ps2} also becomes a trade-off variable w.r.t. size. The implication for design is, that the further the mating surface between top and base is moved downward, the less space there is available for the spring mechanism. The only harmonious variable left in g_1, is d_{ps1}; identifying its' glb may reveal additional variables that contribute to the trade-offs between f_1, c_1 and c_3.

The remaining variables, including $d_{ps1}, n_a,$ and w_{nh} have a conditionally critical set of constraints. Specifically, the spring, hub, and plug variables are potentially bound by inequality constraints relating to the yield stress of different parts, while the top housing variables are also involved in remaining axial fit constraints not shown above. Hence they cannot be eliminated without substantial algebraic manipulation to identify the *glb* or *lub* of each variable for any values of ϵ, meaning it is more efficient to identify the remaining active constraints numerically.

4.4.2 Numerical Results

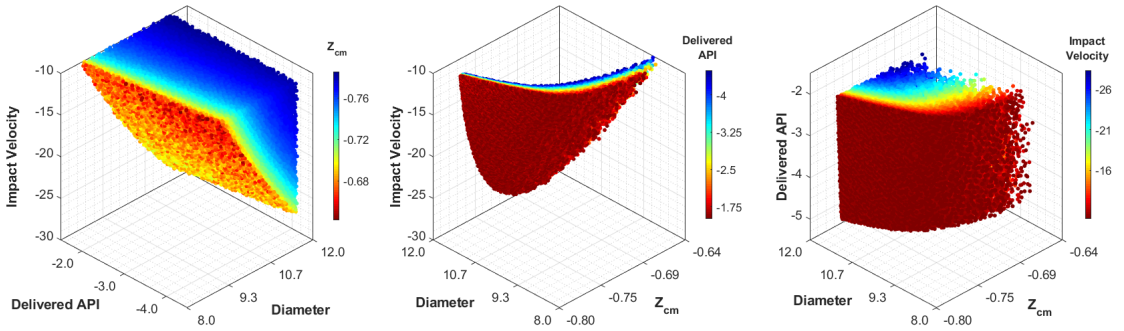
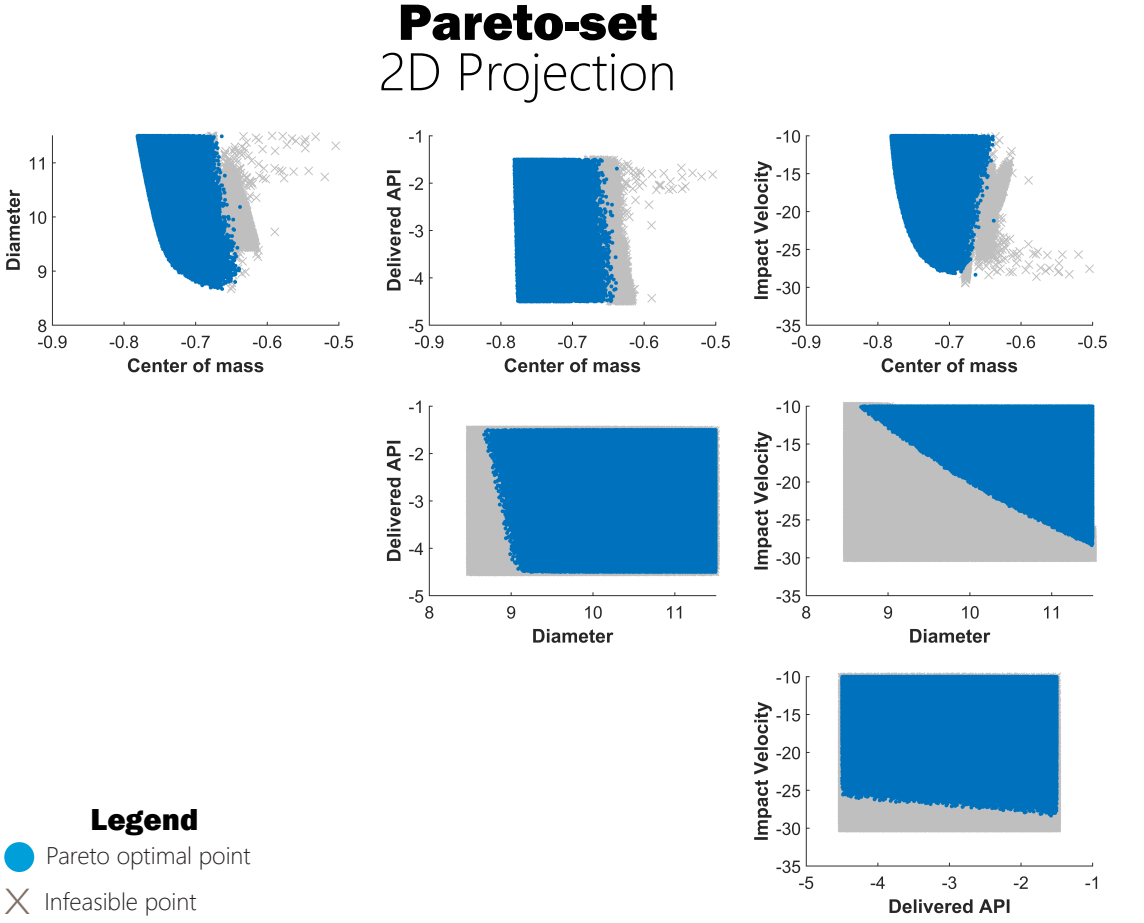


Figure 4.7: *From Paper A*: Different projections of the 4D-Pareto set, where the 4th objective is visualised with a color map

The upper bound problem was solved 200,000 times using the SQP fmincon routine in MATLAB2019R [123] for different values of ϵ sampled from a uniformly distributed quasi-random set (a leaped Halton set) between $\epsilon_{\mathbf{L}} = [8.5\text{mm}; 1.5\text{mg}; 10\text{m/s}]$ and $\epsilon_{\mathbf{U}} = [11.5\text{mm}; 4.5\text{mg}; 30\text{m/s}]$. These values were set based on input from the SOMA team in Novo Nordisk. The results are shown in Figs. 4.7-4.4.2 and Table 1.
As the minimum λ values of each bound objectives are positive, they are active in the entire feasible sampling region and all feasible solutions are Pareto-optimal. As seen in figures 4.7-4.8, all four objectives are in trade-off with each other.

While 42% of the iterations yielded feasible, optimal solutions, the other 58% failed to identify a feasible solution. To verify the model that led to these results, a few measures were taken. Firstly, the validity of the MA was assessed by running the original unreduced model over a narrower range of ϵ values (due to the increased computational cost). This led to the



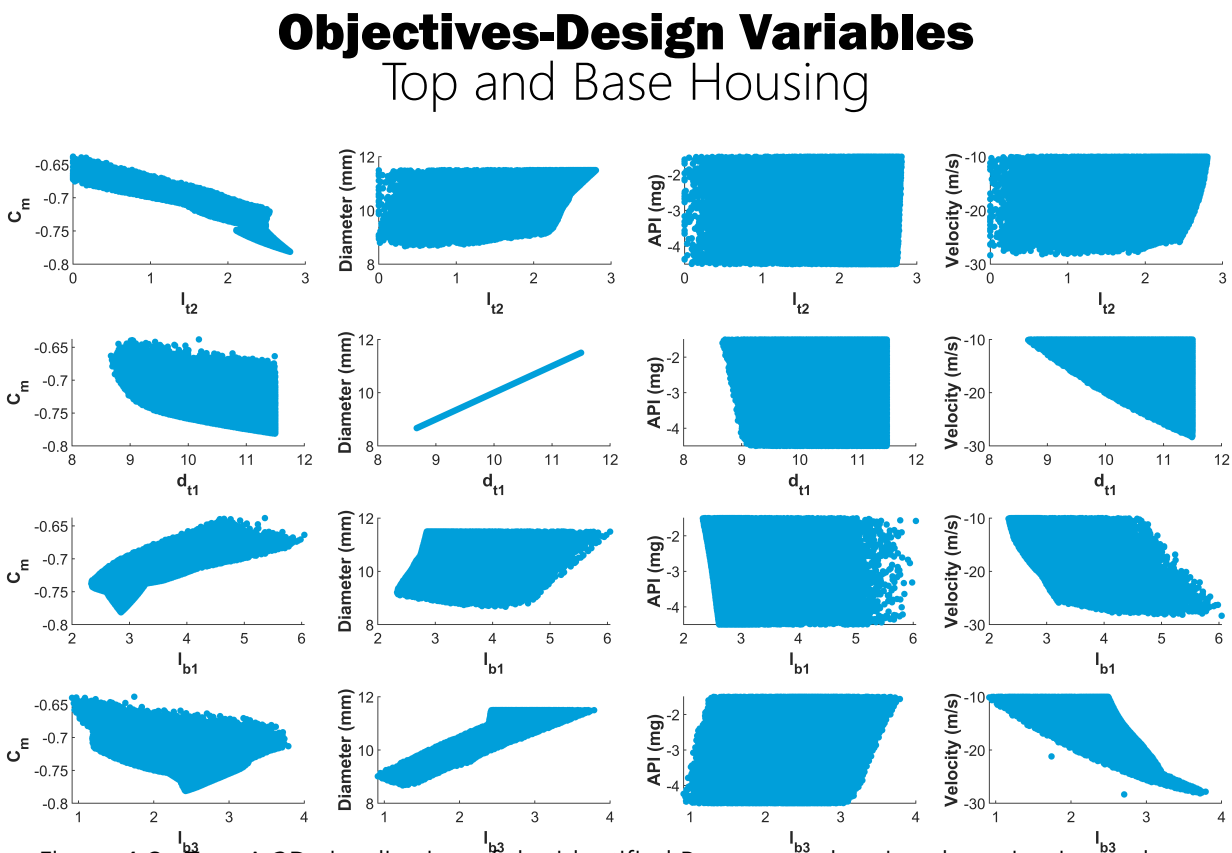
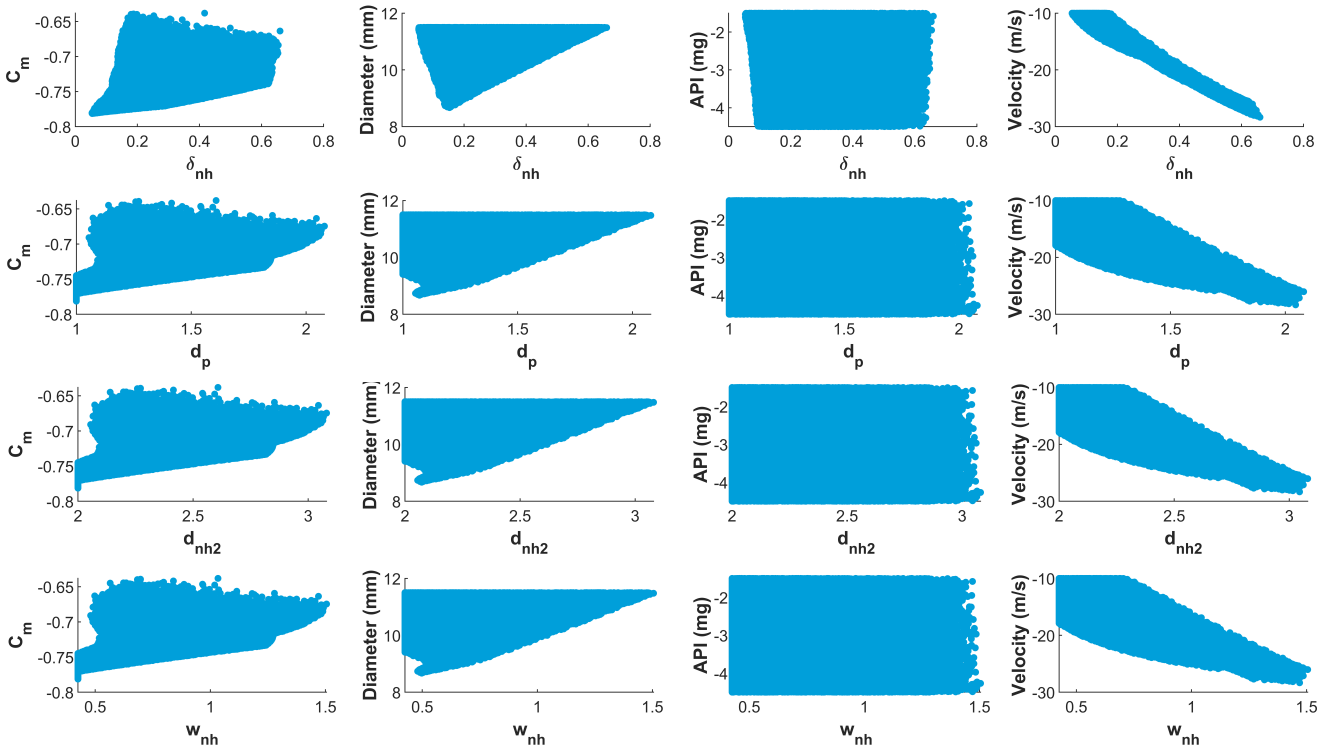


Figure 4.8: *Top*: A 2D visualisation of the identified Pareto set, showing the pair-wise trade-offs *Bottom:* The objective-variable relationships that exist at the optimum, for the design variables left in the reduced model describing the top and base housings.

Objectives-Design Variables
Trigger System



Objectives-Design Variables
Power Spring

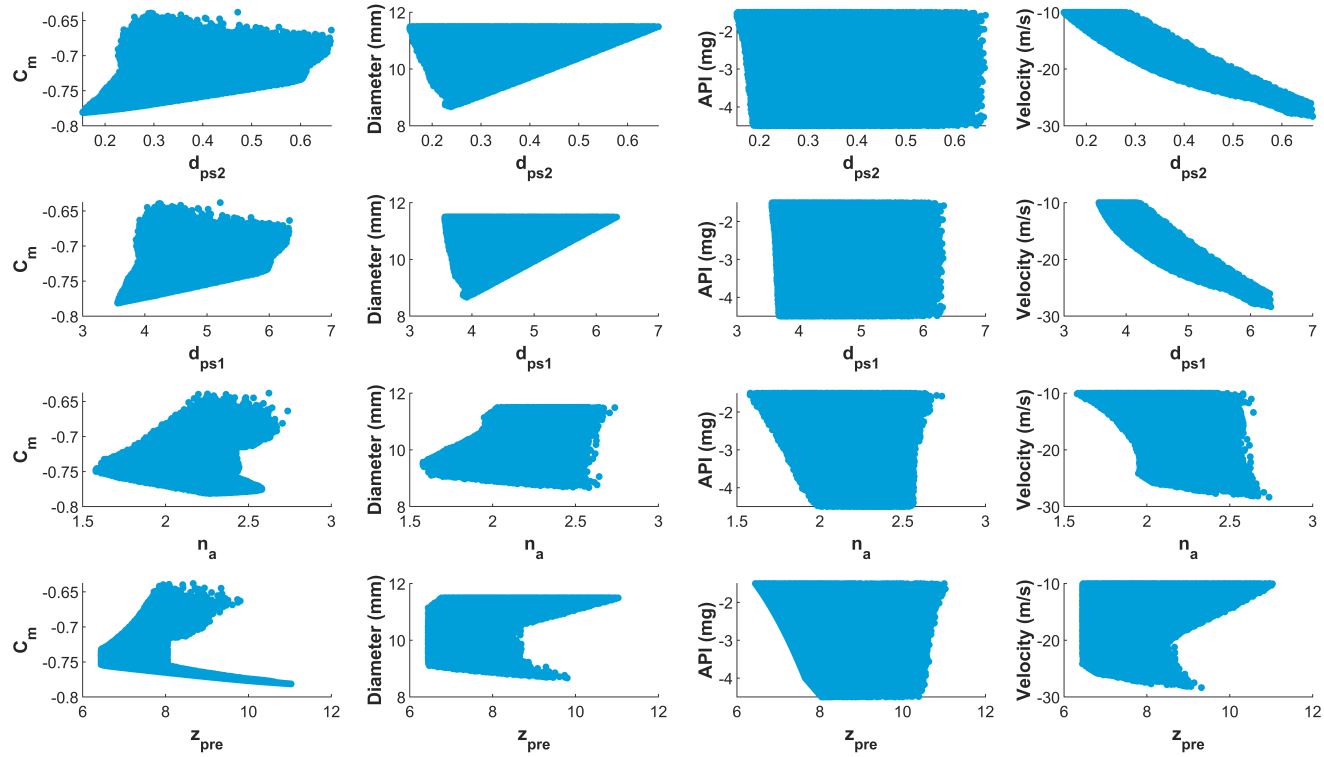


Figure 4.9: *Top*: The objective-variable relationships that exist at the optimum, for the design variables left in the reduced model describing the trigger system *Bottom:* The objective-variable relationships that exist at the optimum, for the design variables left in the reduced model describing the spring.

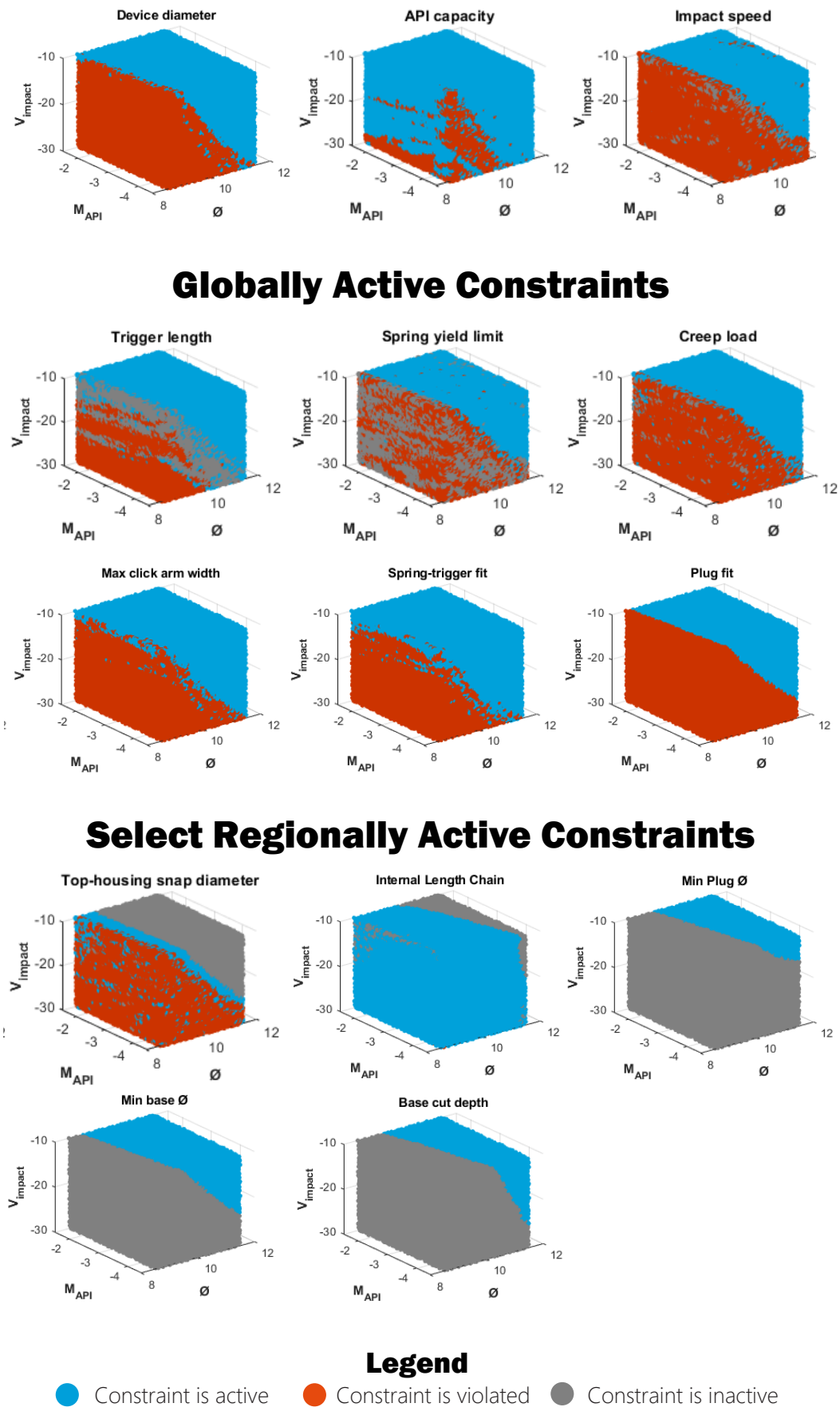


Figure 4.10: Constraint activity plotted across the sampled range of ϵ values. Notice how the *"Top-housing snap diameter"* becomes active, and then violated in the transition from the Pareto set to the infeasible region.

Objective	Optimum	Nadir	λ_{min}	λ_{max}
f_1	-0.78h	-0.64h	-	-
ϵ_1	8.67 mm	11.50 mm	0.0131	2.7708
ϵ_2	4.50 mg	1.50 mg	0.0016	0.4355
ϵ_3	28.34 m/s	10 m/s	0.0008	0.3795

Table 4.2: Numerical results

same results as with the reduced model. Secondly, a constraint satisfaction problem was run for all the failed iterations. This was used to search for a feasible solution to use as a new initial guess in a re-run with the same values of ϵ. Only 1.8% of these cases identified a new feasible initial guess, and these all yielded a Pareto point subsequently. This indicates that the remaining iterations indeed failed due to a lack of feasible domain caused by inconsistent constraints, meaning that the approximate Pareto-frontier of the sampled objectives had indeed been identified. Finally, the non-reduced problem was also run with the NSGA-ii routine built into MATLAB (*gamultiobj*), to rule out the possibility of a discontinuous attainable set. No solutions were identified that dominated any of the Pareto points found using the ϵ-constraint model. It should be mentioned however, that this suffered from computational issues (a substantial computation time) meaning the model was run with a smaller *population* than the size of the results from the ϵ-constraint model.

Furthermore, several Pareto optimal designs in the set were identified that were similar in proportions to the existing design developed by the SOMA project. This indicates that the project has reached a design through iterative design that is at least Pareto *adjacent*, which also indicates that the optimization model reflects reality.

Looking at the results, it is worth recalling that the US-FDA generally recommends pills and capsules stay below a standard 00-size[111], which has a 8.35mm diameter, while the largest standard size, 000 capsules, are 9.91mm in diameter. Complications from swallowing pills start at about 8mm and grows substantially beyond a diameter of 11mm [110]. Thus, a substantial loss of utility occurs beyond 9.91 mm, beyond which a sharp increase in swallowing complications is likely, all the while the SOMA would become more expensive, having to be supplied inside non-standard capsule. Initial work in the SOMA project has revealed that the impact velocity is critically important to the bioavailability of the delivered API (the % of the administered drug that reaches systemic circulation). It is also critical to the robustness and cost of the shaping of the needle geometry, as a low velocity results in a need for a sharper tip. Thus, the trade-off between size and velocity ultimately affects the amount of drug that can be delivered in a swallowable device, and the cost of treatment. Beyond this, the impact velocity objective also seems to have the largest degree of trade-off with the other objectives.

In inspecting the results further, an one might get an indication of what drives these trade-offs. First, looking at the objective-variable relationships for each of the Pareto points, does reveal additional information. Figures 4.8-4.9 show these relationships for select variables. Inspecting the shapes of these plots, indicates that there might be some linear dependencies between some of the design variables (due to active constraints) meaning more variables can be optimized out, post-optimality. There are:

- Between the plug diameter (d_p), plug hole (d_{nh2}), and trigger arm width (w_{nh}). It is worth noting that all three have opposite trends with the device diameter and the impact velocity, despite them being harmonious variables in the pre-optimality MOMA. This points to that the active constraints that create the linear dependency, probably worsen the trade-off between the two objectives.

- Between the spring coiling diameter (d_{ps1}), spring wire diameter (d_{ps2}), and the trigger overlap (δ_{nh}). Here, it is again worth noticing the trend for d_{ps1} and δ_{nh} in the device diameter and impact velocity objectives; both are harmonious variables, again pointing to that the active constraints determining their value, are worsening the trade-off between the two objectives.
- Between spring pretension (z_{pre}) and number of active coils (n_a), which is perhaps unsurprising given that they are both involved in the spring yield stress constraint, having an opposite monotonic influence on the constraint.

Inspecting the activity of the constraints across the entire sampled range of ϵ yields further information, which can be used in the subsequent trade-off root cause analysis. As shown in figure 4.10, there are six constraints that are globally active - i.e. in each Pareto point (c.f. definition 2). The constraints in question are g_5 (spring-trigger radial fit), g_7 (plug fit), g_{11} (trigger creep load), g_{12} (spring yield limit), g_{25} (a molding constraint), and the axial spring-trigger fit (they are equally long).

Furthermore, a few constraints that are regionally active constraints are shown at the bottom of figure 4.10. Note how g_1 becomes active around the same region of the sampled domain, as the solutions start becoming infeasible. Interestingly, one of either g_1 and g_5, and g_7 were violated in every infeasible iteration, pointing to inconsistent constraints beyond the Pareto set, indicating that the design of the spring and trigger might be determining the 3D Pareto frontier between the three bound objectives. This might indicate, that the sudden activity of g_1 around the 3D frontier, is causing some constraint inconsistencies beyond the Pareto set. Yet, without further analysis, the optimization results cannot directly reveal what causes these trade-off.

This numerical data gives additional insights that can be used in further MOMA and ϵMA, to reveal the root causes of the trade-offs in the SOMA design. These subsequent model reductions might reveal the explicit relationships at play in and beyond the Pareto set, and the trade-off variables therein.

4.4.3 Post-Optimality Analysis: ϵMA and Trade-off Root causes

Given the numerical results, we now have a set of globally active constraints, that can be used to reduce the problem further. Furthermore, it is of special interest to identify the active constraints that cause the aforementioned linear dependencies, as these seem to drive trade-off between the size and impact velocity.

The global activity seen in the numerical results is used to eliminate several variables, which are eliminated in the following sequence:

1. The activity of g_7 reveals that $d^*_{nh2} = d_p + 2(R_{wt} + R_{cl})$
2. The activity of g_5 reveals that $d^*_{ps1} = d_p + 2\delta_{nh} + d_{ps2} + 6R_{cl} + 4R_{wt}$
3. The activity of g_{25} reveals that $w^*_{nh} = d_p - 3/2R_{wt} - R_{cl}$
4. The activity of g_{12} reveals that $n^*_a = n_a(d^-_p, d^+_{ps2}, z^+_{pre})$

This allows further reduction of equations 4.71-4.84, with the sequence of elimination meaning that z_{pre} disappears from g_{11}. Along with the elimination of d_{ps1} and w_{nh}, this leaves $g_{11}(d^+_{ps2}, \delta^-_{nh}; \sigma_c)$. For the sake of comprehension, g_{11} will not be eliminated yet.

After these reductions, we can move on to ϵMA, by using the globally active bound size objective $c_1(d^+_{t1}; \epsilon_1)$ to eliminate $\overline{d_{t1}}$, and $c_1(d^-_n, l^-_{n1}, l^-_{n2}; \epsilon_2)$ to eliminate $\overline{l_{n1}}$. This introduces $\tilde{\epsilon_1}$ into g_1, and $\tilde{\epsilon_2}$ into g_{20} , and both of them into the two objectives, f_1 and c_3,:

$$\text{min.} \quad f(l_{b1}^-, l_{t2}^-, d_{t3}^+, d_{ps2}^+, \delta_{nh}^+, d_p^+, \tilde{\epsilon_1}^-, \tilde{\epsilon_2}^+) \tag{4.85}$$

$$\text{s.j.t.} \quad c_3(l_{b1}^-, l_{t2}^-, d_{ps2}^-, d_p^+, \delta_{nh}^+, \tilde{\epsilon_1}^-, \tilde{\epsilon_2}^+; \epsilon_3) \tag{4.86}$$

$$g_1(\tilde{\epsilon_1}^-, l_{t2}^+, d_{ps2}^+, d_p^+, \delta_{nh}^+) = 2d_{ps2} + 2\delta_{nh} + d_p + 10R_{wt} + 12R_{cl}$$
$$+ 2R_{ov} - \sqrt{\frac{2(C_T\tilde{\epsilon_1} - l_{t2})\tilde{\epsilon_1}}{C_T} - \frac{(C_T\tilde{\epsilon_1} - l_{t2})^2}{C_T^2}} \leq 0 \tag{4.87}$$

$$g_8(\delta_{nh}^-) = 0.3\text{mm} - \delta_{nh} \leq 0 \tag{4.88}$$

$$g_9(d_p^-) = 1\text{mm} - d_p \leq 0 \tag{4.89}$$

$$g_{10}(\delta_{nh}^+, d_p^-) = 2\delta_{nh} - R_{cl} - d_p \leq 0 \tag{4.90}$$

$$g_{11}(d_{ps2}^+, d_p^-, \delta_{nh}^-) \leq 0 \tag{4.91}$$

$$g_{20}(l_{b3}^+, d_{ps2}^+, d_{n1}^+, l_{n2}^+, l_{t2}^-, l_{b1}^-, \tilde{\epsilon_1}^-, \tilde{\epsilon_2}^+) = l_{b3} + (n_a(d_p^-, d_{ps2}^+, z_{pre}^+) + 1.5)d_{ps2}$$
$$+ \frac{4\tilde{\epsilon_2}}{\rho_{api}\pi d_{n1}^2} + l_{n2}/3 + 7/2R_{wt} + Z_{tol} + 1.5\text{mm} - C_t\tilde{\epsilon_1} - l_{t2} - l_{b1} \leq 0 \tag{4.92}$$

In this form, two Pareto constraints are revealed, namely g_1 and g_{20}. As expected, g_1 makes l_{t2} a trade-off variable, as $\overline{l_{t2}} \to 0$ as $\tilde{\epsilon_1} \to 0$ when $g_1 \equiv 0$,and given that $f(l_{t2}^-)$ and $c_3(l_{t2}^-)$. The velocity objective c_3 has not been optimized out, as there is no closed form solution to $c_3(\mathbf{x}, \tilde{\epsilon_1}; \epsilon_3) \equiv 0$ w.r.t any $\underline{x}$. Its elimination would involve solving for d_{ps2}, as it is critically constrained from below by c_3 and is shared with the largest number of constraint functions that remain in the model. This would make g_1 a multiobjective Pareto constraint. Therefore, g_1 is involved in three Pareto-optimal activity cases; when g_1 bounds $\tilde{\epsilon_1}$, d_{ps2}, and l_{t2}. We will use these cases to demonstrate the application of ϵMA, using the *Analysis of Pareto-optimal Activity Cases* procedure described in section 4.3. This reveals the root cause of the shape and position of the bi-objective Pareto front between size and velocity.

Activity case 1: Smallest Possible Device, U($\tilde{\epsilon_1}^+$)
Here g_1 determines $\underline{\tilde{\epsilon_1}}^*$ and yields the optimal size. Eliminating l_{t2}, allows a closed form solution for $\tilde{\epsilon_1}$ using Eq. 4.87. Letting $l_{t2} \to 0$, implying that the mating surface between top and base is located at the widest point of the device, allows the smallest $\tilde{\epsilon_1}$. Inserting this, and the parameter values, $C_t = 0.68$, $R_{wt} = 0.45$mm, $R_{cl} = 0.1$mm, $R_{ov} = 0.6$mm , and letting $\delta_{nh} \to \underline{\delta_{nh}}$ and $d_p \to \underline{d_p}$ yields a reduced expression:

$$g_1(\tilde{\epsilon_1}^-, d_{ps2}^+, d_p^+, \delta_{nh}^+) = 2(d_{ps2} + \delta_{nh}) + d_p + 7\text{mm} - \tilde{\epsilon_1} \leq 0 \tag{4.93}$$

$$\Rightarrow \underline{\tilde{\epsilon_1}}(d_{ps2}^+, \delta_{nh}^+, d_p^+) = 2d_{ps2} + 2\delta_{nh} + d_p + 7\text{mm} \tag{4.94}$$

$$\Rightarrow g_8(\delta_{nh^-}) \equiv 0 \wedge g_9(d_p^-) \equiv 0 \tag{4.95}$$

$$\Leftrightarrow \underline{\delta_{nh}} = 0.3\text{mm}, \underline{d_p} = 1\text{mm} \tag{4.96}$$

$$\Leftrightarrow \underline{\tilde{\epsilon_1}}^* = 2d_{ps2} + 8.6\text{mm} \tag{4.97}$$

As d_{ps2} is a trade-off variable, minimising $\tilde{\epsilon_1}$ will lead to a point where $g_{11} < g_8$ and $g_{11} < g_9$ w.r.t. δ_{nh} and d_p. This results in g_8 and g_9 becoming active, leading to the back-substitution performed in eqs. 4.94-4.97. As we know, g_{11} is globally active, meaning that $\mathcal{X}(d_{p2})$ is narrowed at the Pareto frontier between size and velocity, given that the reductions made in this activity case leave $g_{11}(d_{ps2}^+; \sigma_c)$ and $c_3(d_{ps2}^-, \tilde{\epsilon_1}^-; \epsilon_3)$. Further, c_3 is critical w.r.t. bounding d_{ps2} from below (as $\tilde{\epsilon_1}^*(d_{ps2}^+)$) , meaning that $\epsilon_L(3)$ ultimately determines the lowest feasible value of d_{ps2}, and hence $\underline{\tilde{\epsilon_1}}^*$.

Activity case 2: Maximum Impact Velocity, U($\tilde{\epsilon_3}^-$)
Here g_1 determines $\overline{d_{ps2}}$. As c_3 is monotonically decreasing w.r.t. d_{ps2} to the power of 4, its supremum yields the single-objective optimal impact velocity. Thus, the same parameter values and value of l_{t2} can be inserted as in Case 1. Furthermore, we utilise that we know that g_{11} is globally active. It is a interface stress criterion, and because the spring force grows with the wire diameter, the dimensions that determine the area - d_p and δ_{nh} increase correspondingly. This in turn makes $g_{10}(\delta_{nh}^+, d_p^-)$ active, as $g_9 < g_{10}$ for any $\delta_{nh} > 0.45$mm.. Assuming the spring force is increased to make g_{10} active, and subsequently using g_{11} to eliminate δ_{nh}, yields:

$$g_8 \equiv 0 \Rightarrow d_p^* = 2\delta_{nh} - R_{cl} \tag{4.98}$$

$$g_{11} \equiv 0 \Rightarrow \delta_{nh} = \delta_{nh}(d_{ps2}^+; \sigma_c, \sigma_y) \tag{4.99}$$

$$\Rightarrow g_1(\tilde{\epsilon_1}^-, d_{ps2}^+) = 2d_{ps2} + 4\delta_{nh}(d_{ps2}^+; \sigma_c, \sigma_y) - R_{cl} + 7\text{mm} - \tilde{\epsilon_1} \tag{4.100}$$

$$\Rightarrow \overline{d_{ps2}} = 0.5(\tilde{\epsilon_1} - 4\delta_{nh}(d_{ps2}^+; \sigma_c, \sigma_y) - 6.9mm) \tag{4.101}$$

As no closed form solution exists, g_{11} has been used to implicitly eliminate the trigger overlap δ_{nh} which determines the size of the load bearing trigger interface ; σ_y is the yield stress of the spring, and σ_c is the allowable static stress in the trigger interface. This substitution reveals a feedback coupling; as the wire diameter d_{ps2} is increased, so does the required load bearing area, reducing the space available for the spring wire in a device of a given size, $\tilde{\epsilon_1}$. Correspondingly, as $d_{ps1}^* = d_p + 2\delta_{nh} + d_{ps2} + 4R_{cl} + 2R_{wt}$, stiffening the spring by reducing the coiling diameter, also reduces the available load bearing area, and the allowable plug size. In other words, the Pareto frontier between the device diameter and spring, is caused by the spring needing to fit around the trigger.

Activity case 3: Lowest Possible Center of Mass, $f_1(l_{t2}^-)$
Here g_1 determines $\overline{l_{t2}}$. As $f_1(l_{t2}^-)$, this case occurs at the single objective optimal self-orientation. Given the non-linearity of Eq. 4.87 w.r.t. l_{t2}, the variable is best eliminated implicitly, yielding $\overline{l_{t2}} = l_{t2}(\tilde{\epsilon_1}^+, d_p^-, d_{ps2}^-, \delta_{nh}^-)$. As a consequence d_{ps2} is bounded by c_3, d_p by either g_9 or g_{10}, and δ_{nh} by either g_8 or g_{11}. Furthermore, $\tilde{\epsilon_1} = \epsilon_U(1)$, as no constraint bounds $\tilde{\epsilon_1}$ from above.

4.4.4 Design Implications

Through the application of MOMA, numerical solution of the optimization, and subsequent MOMA and ϵMA, we have gained valuable information about the SOMA design. Through opportunistic, yet rigorous analysis, we have found the dependencies and constraints that cause the trade-offs between the four modelled objectives.

First, the application of multi-objective monotonicity analysis before and after computation, revealed several trade-off variables, some of which are not directly apparent in the initial mode. The key trade-off variables and the constraints that introduce them or worsen their effect, are shown in table 4.3. Of the variables not already discussed in the preceding sections, l_{b3} is involved in a regional trade-off between API payload and self orientation, caused by the regional activity of g_{20}. In essence, the longer the API needle, the less space remains for a thick layer of material in the lowest area of the base housing, and correspondingly, the larger l_{b3} is, the less space there is for the needle. This is driven my the fact that g_{20} ensures that the needle does not protrude through the valve, before the device is triggered. Furthermore, d_{n1}, the diameter of the needle is a trade-off variable, as it increasing it increases the API payload, but it also increases the size of the hole in the base that it passes through during

	$\overline{d_{t1}}$	$\overline{l_{t2}}$	$\overline{l_{b3}}$	$\overline{d_{ps2}}$	$\overline{l_{n1}}$	$\overline{d_{n1}}$
f_1 - Self Orientation	-	-	-	+	+	+
c_1 - Diameter	+	(+)		+	(+)	
c_2 - API Payload	(-)		(+)	(+)	-	-
c_3 - Impact Velocity	-	(-/N)		-	+	+
Caused or worsened by	h_1,h_8, g_1, g_{20}	h_8,g_1,g_{20}	g_{20}	h_8, g_1, g_5, g_8, g_{10}, g_{11}, g_{12}, g_{20}	h_8,g_{20}	g_{31}

Table 4.3: An overview of the key trade-off variables, and the constraints that either cause their introduction into the objective functions, or increase their influence. A parentheses denotes a regional dependency, caused by a regionally active constraint. Of note here, is that c_3 becomes non-monotonic w.r.t. l_{t2}, if g_1 is active and is used to eliminate d_{ps2}.

injection. This removes material from the lowest part of the device, worsening the position of the center of mass.

Importantly, some active constraints do not directly introduce trade-off variables, but rather increase the influence of variables already present in the optimization problem. An example of such, is the back-substitution of n_a, which increases the influence of d_{ps2} upon self orientation, as the spring yield limit constraint (g_{12}) in effect means that the number of spring coils grows monotonically as the spring wire diameter is increased.

Other constraints might not be globally active in the numerical results, but when reduced further, are revealed to further worsen trade-offs between certain objective pairs. Such tendencies were revealed through ϵMA, giving a better understanding of what relationships define bi-objective Pareto frontiers or the single-objective optima. An example of such, is g_1, which ultimately defines the shape and position of the frontier between impact velocity and device diameter , as seen in activity cases 1-3.

Yet, the most important question here , is what these results imply about the current SOMA design. A revealing approach, is to look at the active constraints and (regional) trade-off variables present at different vertices in the Pareto set. Examples of important vertices (e.g. the single objective optimum of self orientation and size) are visualised in figure 4.11.

Quite tellingly, the spring coiling diameter, d_{ps1} is at its lower bound in each one of these vertices, but as seen in the objective-variable plots in figure 4.9, its optimal value grows monotonically with the impact velocity. Meanwhile, l_{t2} is on opposite ends of its feasible domain, when comparing the center of mass objective with device diameter and impact velocity. Despite g_1 being a regionally active constraint, it does seem to cause a quite a bit of conflict between the objectives. Based on this, the identified trade-off variable,s the constraints that cause them, and the regional relationships that exist between the objectives we can, seen from a design perspective, surmise that there are (at least) four primary causes of trade-off

1. **Radial Spring Fit** The spring fits around the trigger, meaning that , the spring coil diameter is given by $d^*_{ps1} = \underline{d_{ps1}} = d_{ps2} + 2\delta_{nh} + d_p + 6R_{cl} + 4R_{wt}$. The coiling diameter influences the spring force to the third power, yet cannot be reduced without reducing the available load bearing surface in the trigger surface. Looking at the spring index (ratio between the coiling diameter and wire diameter) of each Pareto optimal solution, we see that the configuration itself can be argued to be far from optimal. Figure 4.12 shows a histogram of the spring index, for the set of Pareto optimal solutions identified in section 4.4.3.

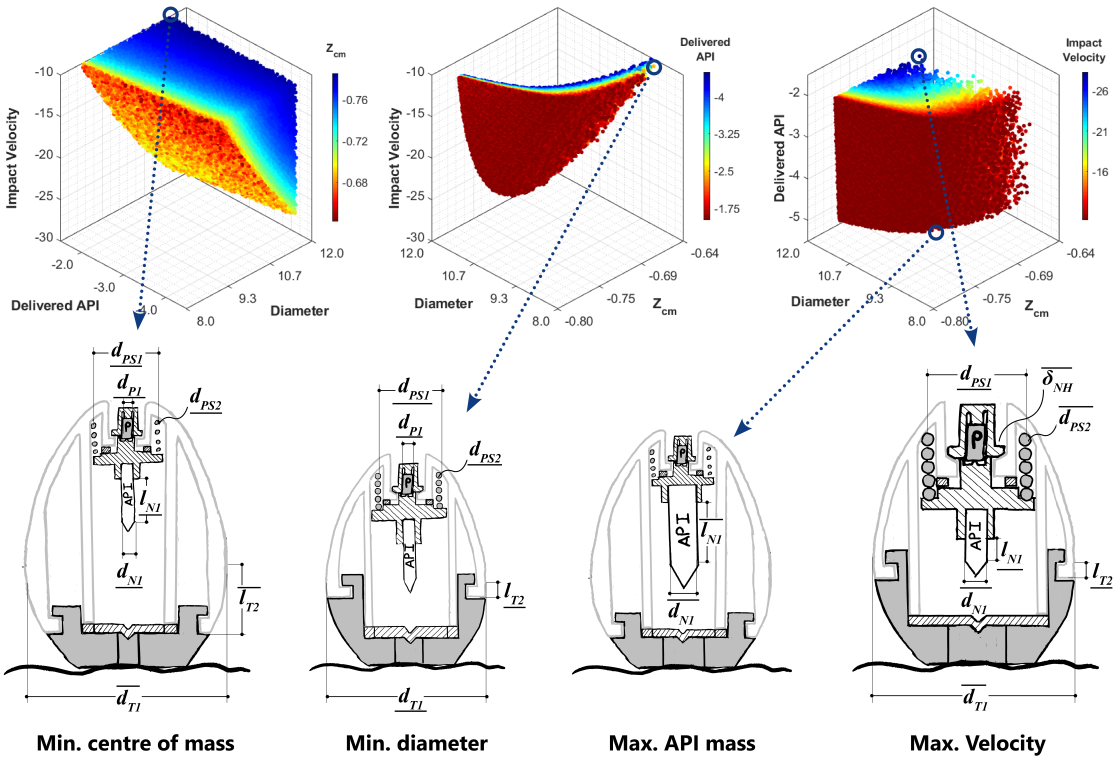


Figure 4.11: Looking at the changes in constraint activity at important vertices in the Pareto set, reveals important knowledge about the drivers of the trade-offs between the modelled objectives. The effect of these activities is illustrated in the sketches of the vertex designs of the SOMA

As found by Wahl [112], the volumetric efficiency of compression springs decreases monotonically with the spring indices beyond 4. Yet, given that diameter of the trigger grows with the spring force, the volumetric efficiency of the SOMA is reduced as the impact velocity is increased. This worsens the achievable combination of device size and impact velocity.

The result of this relationship, is that $g_1$takes the form:

$$g_1(\tilde{\epsilon_1}^-, l_{t2}^+, d_{ps2}^+, d_p^+, \delta_{nh}^+) = 2d_{ps2} + 2\delta_{nh} + d_p + 7mm - \sqrt{\frac{2(C_T\tilde{\epsilon_1} - l_{t2})\tilde{\epsilon_1}}{C_T} - \frac{(C_T\tilde{\epsilon_1} - l_{t2})^2}{C_T^2}} \leq 0 \tag{4.102}$$

In its non-reduced form, g_1 was a fit constraint describing the fit between upper and lower housing. When g_1 is active, any increase in spring wire diameter results in an increase in device size, a reduction in trigger overlap, or a reduction of l_{t2}. The result, is a three dimensional trade-off between velocity, size, and self orientation.

2. **Boundedness of the trigger**: In continuation of the above, the trigger geometry itself presents its own challenges. Prior to computation, the harmonious variables d_p and δ_{nh} were bound by a conditionally critical set of constraints. Looking at constraint activity at the bi-objective Pareto frontier between size and velocity, revealed a locally active glb of d_p, $g_{10} = 2\delta_{nh} - d_p - R_{cl} \leq 0$, which prevents the trigger arms from colliding with each other thereby jamming the device. Further, a stress criterion for the trigger interface, $g_{11}(d_{ps2}^+, \delta_{nh}^-)$, is globally active, in effect locking the relationship between the spring wire diameter, d_{ps2} and the trigger overlap δ_{nh}.

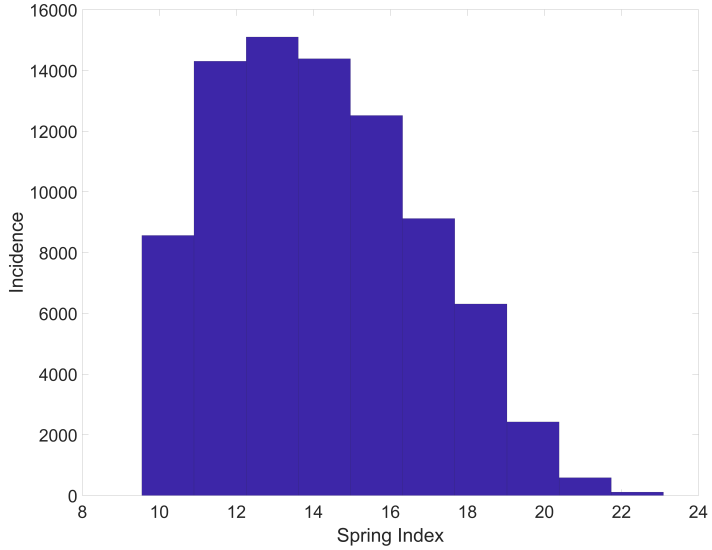


Figure 4.12: The distribution of spring indices in the set of Pareto-optimal solutions

The implication of the activity of g_{10} and g_{11}, is that the influence of d_{ps2} upon the trade-off of velocity against size and self-orientation is multiplied, as any increase in d_{ps2} results in an increase in both d_p and δ_{nh}, all of which contribute to size and mass. These dependencies, specific to the Pareto-set, mean that the spring force can only be increased to a certain point for a given device size. Beyond this point, the device would fail due to creep fracture driven by high static interface stress or simply not fit together radially.

3. **Axial fit of internal components**: The impact velocity is determined by the force profile exerted by the spring, the frictional resistance in the system, and the acceleration stroke distance between the tip and tissue (z_{acc}). Given that the internal parts in the SOMA device are mounted in a vertical series, z_{acc} is involved in the following constraints:

$$\begin{aligned} h_8 = R_{wt} + (n_a + n_d)d_{ps2} + l_{nh2} + l_{nh1} + l_{n1} + l_{n2} \\ +z_{acc} - l_{b1} - l_{t2} - l_{t1} = 0 \\ g_{20} = l_{b3} + l_{v1} + Z_{tol} - z_{acc} \leq 0 \end{aligned} \tag{4.103}$$

Through MOMA and computation, several constraints were found to be active, resulting in the following term describing z_{acc} in the optimal set:

$$z_{acc} = C_t d_{t1} + l_{t2} + l_{b1} - (n_a(d^+_{ps2}, z^+_{pre}) + 1.5)d_{ps2} - l_{n1} - l_{n2} - 1.5\text{mm} - 7/2R_{wt} \tag{4.104}$$

This relationship, which is specific to the optimal set, reveals that any increase in the spring length comes at the cost of needle length or device size. Furthermore, any increase in the needle length also moves the spring (and its high mass) upward, resulting in a worse position of the center of mass. Thus, we can surmise that the serial arrangement of spring and needle detrimentally influences the achievable combination of API payload, impact velocity, and position of center of mass.

Furthermore, this serial arrangement also means that g_{20}, a locally active constraint that prevents the needle from protruding through the valve before injection, contributes to the trade-offs involving all four objectives, as seen through its reduced form, shown in equation 4.92.

4. **Assembly Features - Joining of the top and base housings** The housing snap ($d_{b4} - d_{b3} = 2R_{ov} = 1.2$mm), the cylindrical geometry on the top housing which seals the valve against the base, and needle attachment ($l_{nh1} = 1.5mm$) on the needle hub, are assembly features that result in parametric contributions that detrimentally affect the Pareto set, as seen through the radial and axial constraints, through $\overline{z_{acc}}$, and the parametric contributions in the the reduced form of g_1. Given the activity of these constraints, any design change that eliminates these contributions would improve the Pareto-set. The housing snap and cylindrical geometry have an especially negative impact on the trade-off between self orientation and device diameter, given that they affect the relationship between the achievable device size, and the position of the housing split l_{t2} (through g_1), which is highly influential upon the position of center of mass, as seen in the objectives-variables plots in figure 4.8.

How the knowledge of these four issues can be leveraged in redesign, will be demonstrated at the end of the next chapter. This will involve systematic design changes which result in changes to the Pareto set, resulting in improved overall performance and reduced trade-offs.

5 Trade-off Mitigation Through Redesign

This chapter presents the prescriptive research which builds upon the results of the previous chapter, to define an opportunistic configuration redesign methodology where the designer uses the MOMA and ϵMA methods to identify directions for configuration design changes which may yield reduced trade-offs or improved performance. Hence, the chapter presents answers to RQ3. Most of this work was originally described and submitted for publication in Paper B. The chapter starts with formal definition of design improvement, followed by the a formal treatment the outputs of Pareto set dependency analysis (MOMA and ϵMA), which gives rise to a set of design principles. The chapter then continues with the presentation of a redesign methodology as the systematic application of these principles to eliminate dependencies and relax the constraints that limit optimality. This methodology is then applied to the SOMA case, to derive 11 novel redesigns. This redesign process, and four of the resulting redesigns are described in Paper B, and in (until now) three patent applications submitted by the case company in the authors name.

5.1 Defining Design Improvement

In the preceding chapter, the Pareto-set Dependency Analysis method was developed, with the aim of allowing systematic root cause analysis of the trade-offs in a design. This method consists of a set of mathematically founded procedures for multi-objective monotonicity analysis, model reduction, and analysis of regions or vertices of the Pareto set. In the presence of monotonic design properties, the systematic application of these procedures allow the identification trade-off variables, harmonious variables, active constraints, and Pareto constraints of a design problem, both globally and regionally in the Pareto set.

As these dependencies and constraints affect the Pareto set, it stands to reason, that introducing configuration design changes which eliminate or reduce some of these dependencies and relax some of the constraints, would result in new Pareto set. Thus, we can potentially use the outputs of Pareto set dependency analysis to identify configuration design improvements.

Yet, before considering how to identify design improvements, we first need a formal definition of what *design improvements* even constitute. Here, it is important to consider that configuration design changes imply changes to the optimization model. Thus, we need a formal approach to comparing the results of multiple optimization models, describing different configuration designs, in order allow the evaluation of whether a redesigned configuration is in fact an improvement. Here, we can employ the notion of *meta-Pareto optimality* [55] which involves comparison of different solutions to the same design problem:

Definition 4 Meta-Pareto Set
Given Pareto sets $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_p$ for p configuration solutions for a given design problem, the meta-Pareto set $\breve{\mathcal{C}}$ consists of points within the union of these sets, $\mathcal{C}_\mathcal{U} = \mathcal{C}_1 \cup C_2, \cup ... \cup \mathcal{C}_p$, that are Pareto-optimal with respect to the set $\breve{C}$. A point $\mathbf{f}_*$ is meta-Pareto-optimal if and only if there exists no point $\mathbf{f} \in \mathcal{C}_U$ such that $f_i \leq f_i*$ for all i and that $f_i < f_i*$ for at least one i.

The use of the meta-Pareto set essentially allows the comparison of Pareto points from different optimization models, so long as they involve the same objective measures. Assuming that the configuration design changes identified with the support of Pareto-dependency analysis, do not result in new or changed objectives, we can thus introduce the following definition:

Definition 5 Design Improvement Criterion
If a configuration with Pareto set $\mathcal{C}_0$ is redesigned, resulting in a new Pareto set $\mathcal{C}_1$, the redesign is said to be an improvement, if and only if the meta-Pareto set of $\mathcal{C}_0$ and $\mathcal{C}_1$ is identical to $\mathcal{C}_1$, namely, $\breve{\mathcal{C}} = \mathcal{C}_1$ for any weighting vector **W**, *which implies that all of the Pareto points of the original design are at least weakly dominated by the Pareto points of the redesign.*

Under this definition, the redesigned configuration(s) can still involve changed objective and constraint functions, so long as the objectives themselves (e.g. *minimise system mass*) are the same, allowing comparison in the same *k*-dimensional objective space.

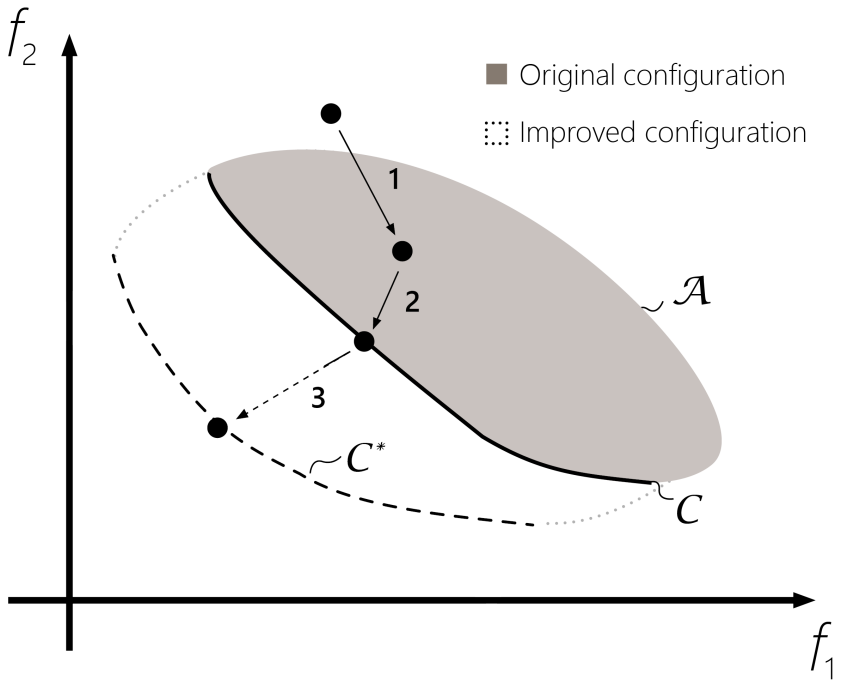


Figure 5.1: *From Paper B:* The difference between constraint satisfaction (1), design optimization (2), and configuration design improvement (3).

The definition implies that the achievable performance in the new design is at least equal to or better than that of the previous design, w.r.t. all criteria, exemplified in Fig. 5.1. This formal definition is independent of the design context and the relative importance of the objectives, and it uses quantifiable properties we can employ in deriving rigorous redesign principles. Since optimality is defined only in the context of the particular optimization model [12], there is an implicit assumption that such comparisons are made using models of similar fidelity. To account for the situation in which the condition $\breve{\mathcal{C}} = \mathcal{C}_1$is not strictly met, we introduce a further definition:

Definition 6 Potential Design Improvement
If a configuration with Pareto set $\mathcal{C}_0$ is redesigned, resulting in a new Pareto set $\mathcal{C}_1$, the redesign is said to be a partial improvement, if and only if the meta-Pareto set $\breve{\mathcal{C}}$, consists of Pareto points both from $\mathcal{C}_0$ and $\mathcal{C}_1$.

Here, we cannot say for sure whether the redesign is a design improvement, as the meta Pareto set contains solutions from both sets. This is exemplified by figure 5.2. Yet, the relative weighting of the design objectives might still mean that the redesign can be viewed as an improvement over the original configuration. In such situations, whether the redesign is *better* is a matter of the subjective opinion of the designer/decision maker.

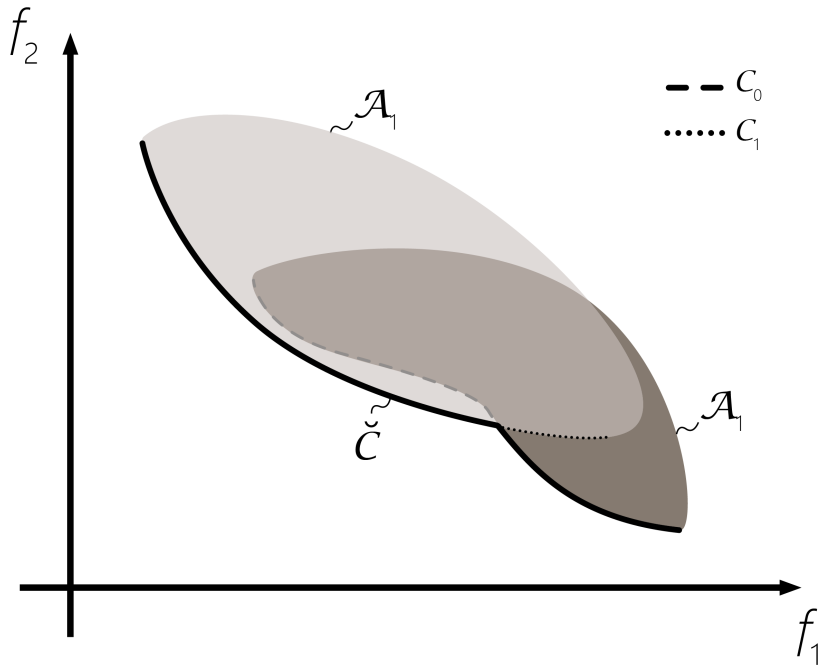


Figure 5.2: In this example, which is also shown in Chapter 3, the redesign $\mathcal{C}_1$ can only be said to be an improvement over $\mathcal{C}_0$ if the weighting of f_2 is sufficiently large compared to the weighting of f_1.

5.2 On the Implications of Pareto Set Dependency Analysis

With this definition of design improvement, we can look into what the outputs of the analysis methods imply about the Pareto set and the configuration design itself. As this will reveal, certain changes to the optimization model, will result in an improved Pareto set.

As discussed, the application of MOMA and ϵMA uncovers the relationships in the Pareto set that limit optimality and drive trade-offs. Looking deeper, we can build upon the theorems and definitions introduced in the previous chapter, to uncover the relationship between the shape and position of Pareto and the key outputs of Pareto-set dependency analysis, as illustrated in Fig.5.3.

We start by noting that the trade-off variables defined earlier stem either from an inherent dependency between the objective functions or from a dependency that exists at the optimum due to active constraints. They have a substantial influence on the Pareto set, which can be understood by considering the effect of their absence in a design problem:

Theorem 5 Existence of Pareto set
If no trade-off variables exist (globally or regionally) after back-substitution of active constraints, then the optimum is a point $F*$, rather than a set. Therefore, a Pareto set cannot exist without trade-off variables.

Proof. If $x_i \notin \overline{x}$, for any design variable *i*, then following Theorem 1 and MP1, argmin $f_i(\mathbf{x})$

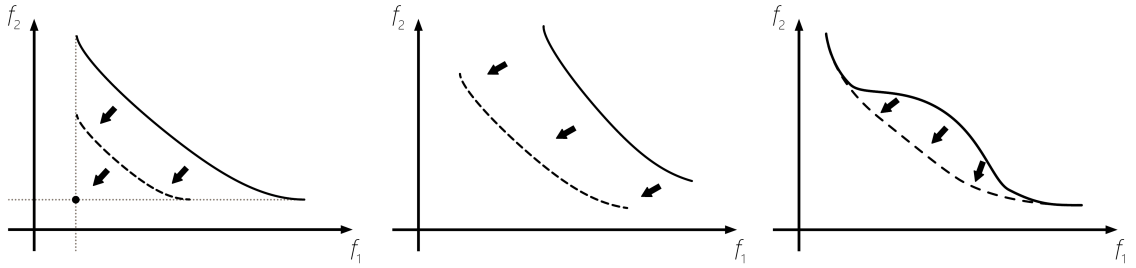


Figure 5.3: *From Paper B*: The relation between analysis outputs and the Pareto set. (*Left*): Trade-off variables result in an optimum that is a set rather than a point; eliminating the underlying dependency brings the set closer to utopia. (*Center*): Harmonious Variables affect the position of the Pareto set; relaxing their constraints shift the set. (*Right:*) Pareto constraints introduce regional relationships that affect the Pareto set; eliminating or relaxing them changes the shape of the set

= argmin$f_j(\mathbf{x})$ for $i \wedge j = 1..k, i \neq j$. Hence, dim($\mathbf{x}^*$) $= [n, 1]$, meaning a single dominant optimum exists. ■

Thus, the more trade-off variables exist in a design problem, the larger the distance between the utopia point F^0, and $\mathcal{C}$. Furthermore, recall from Theorem 1 that the optimum of each objective dependent on a trade-off variable $\underline{\overline{x_i}}$, exists at either $\underline{x_i}$ or $\overline{x_i}$ depending on the objective's monotonicity, with any other feasible value of $\underline{\overline{x_i}}$ yielding a Pareto point. As a result, size of the feasible domain of trade-off variables contributes to the size of the Pareto set.

Finally, being oppositely monotonic, the partial derivatives of an objective pair in trade-off w.r.t $\underline{\overline{x_i}}$ will have an opposite sign across the entire Pareto-set. The larger the difference between these, the larger the slope of the frontier. Thus, the impact of $\underline{\overline{x_i}}$ on the trade-offs between objectives can be worsened by multipliers and divisors.

In summary, the trade-off variables in an optimization problem cause the existence of the Pareto set, affect its span, and to some extent its shape. From this, we can derive several corollaries which reveal design principles that can guide the generation of an improved Pareto set.

Corollary 5.1 Separation of Trade-off Variables
If the trade-off variable $\underline{\overline{x_1}}$ affecting the objectives $f_i(x_1^+)$ and $f_j(x_1^-)$, is substituted through design change in one objective by a new variable x_2, such that $\hat{f}_i(x_2^+), x_2 \notin f_j$, and min $\hat{f}_i \leq$ min f_i, *then* argmin $\hat{\mathbf{f}}(\mathbf{x}) = \{\overline{x_1}, \underline{x_2}\}$ *while* argmin $\mathbf{f}(x_1) = x_1 \in \mathcal{X}$. *As a result* $\hat{\mathbf{f}}(\mathbf{x}) < \mathbf{f}(\underline{\overline{x}_1})$ *for any value of* x_1.

The same would also apply if x_2 was instead substituted into the problem such that $\hat{f}_j(x_2^-)$, or if the influence of x_1 upon one of the objectives in a multivariate problem was simply eliminated without the introduction of another variable. In such problems, it follows that such changes would result in a new Pareto set $\mathcal{C}_2$ that at least weakly-dominates the original Pareto set, $\mathcal{C}_1$. The term separation here is used in the same spirit as in TRiZ [16], reflecting the removal of a dependency.

Correspondingly, the same would occur if a design change is introduced that makes x_1 a harmonious variable, without otherwise affecting the objective functions. If fact, any modification of the design problem which eliminates or reduces the influence of the trade-off

variable on one objective, may result in a new, weakly dominant Pareto set. This is stated formally with the following corollaries.

Corollary 5.2 Flipping Trade-off Variables
If the monotonicity of a trade-off variable $\overline{x_n}$ *affecting the objectives* $f_i(x_n^+)$ *and* $f_j(x_n^-)$*, is flipped in one objective through design change, such that* $\hat{f}_i(x_n^-)$*, and* $\min \hat{f}_i \leq \min f_i$ *, then* $\operatorname{argmin} \hat{\mathbf{f}}(\mathbf{x}) = \overline{x_n}$ *whereas* $\operatorname{argmin} \mathbf{f}(x_n) = x_n \in \mathcal{X}$*. As a result* $\hat{\mathbf{f}}(x_n) < \mathbf{f}(\underline{x}_n)$.

Corollary 5.3 Scaling Trade-off Variables
If the influence of a trade-off variable $\overline{x_n}$ *affecting the objectives* $f_i(x_n^+)$ *and* $f_j(x_n^-)$ *is scaled through the introduction of an independent variable* x_{n1} *in* f_i *such that* $\partial \hat{f}_i/\partial x_n < \partial f_i/\partial x_n$ *then* $\min \hat{f}_i < \min f_i$ *for any value of* x_n*, reducing the trade-off between* f_i *and* f_j*. Correspondingly, if* $\partial \hat{f}_j/\partial x_n > \partial f_j/\partial x_n$ *then* $\min \hat{f}_j < \min f_j$.

As defined earlier, harmonious variables are variables shared between objectives, all being of like monotonicity, denoted $\overline{\overline{x}}$ when the objectives are monotonically decreasing, and $\underline{\underline{x}}$ when they are increasing. For such variables, the glb (for $\underline{\underline{x}}$) or lub (for $\overline{\overline{x}}$) is active at all Pareto points. They might be optimized out in the MOMA process if a globally active constraint can be identified, but may also remain in the model when there are regionally active constraints. Whereas trade-off variables create the Pareto set, the identification of harmonious variables reveals other useful information:

Theorem 6 Position of $\mathcal{C}$
Harmonious variables, $\underline{\underline{x}}$ *and* $\overline{\overline{x}}$*, affect the position of the Pareto set* $\mathcal{C}$ *relative to the origin. Thus design changes that widen their feasible domains in an improving direction, yield a new strongly dominant Pareto set,* $\mathcal{C}_{i+1} < \mathcal{C}_i$.

The reason why, is best understood by examining what happens to the Pareto set when the glb/lub is changed:

Proof. Let f_i and f_j depend on $\overline{\overline{x}}_1$, $\underline{\underline{x}}_2$, $\overline{x_3}$, i.e., $f_i(x_1^-, x_2^+, x_3^+)$, $f_j(x_1^-, x_2^+, x_3^-)$ where $x_1, x_2, x_3 \in \mathcal{P}$. If the problem is well bounded, then by MP1 and Theorem 1, $\arg\min f_i(\mathbf{x}) = \{\overline{x_1}, \underline{x_2}, \underline{x_3}\}$ and $\arg\min f_j(\mathbf{x}) = \{\overline{x_1}, \underline{x_2}, \overline{x_3}\}$. If the active constraints are modified or relaxed, such that $\overline{x_1} < \overline{\hat{x}_1}$ and/or $\underline{\hat{x}_2} < \underline{x_2}$, then $\hat{\mathbf{f}}^*(\mathbf{x}) < \mathbf{f}^*(\mathbf{x})$ for any value of $\overline{x_3}$, given the monotonicity of f_i and f_j. Hence, $\breve{\mathcal{C}} = \mathcal{C}^*$. The reverse is true if the active constraints are tightened, such that $\overline{x_1} > \overline{\hat{x}_1}$ and/or $\underline{\hat{x}_1} > \underline{x_1}$. Hence, the harmonious variables and their bounds influence the position of $\mathcal{C}$. ■

Whereas changing the bounds of trade-off variables only enlarges the Pareto set and moves its utopia point, relaxing the constraints of harmonious variables results in an improved Pareto set. Furthermore the slope of $\mathcal{C}$ will be affected as well, unless x_1 and x_2 influence f_i and f_j equally.

Lastly, Pareto-constraints are a consequence of the systematic reduction of multiobjective problems modelled in an upper bound formulation, which is used as it has several benefits w.r.t. monotonicity analysis. When globally active - i.e. for any $\epsilon_L \leq \tilde{\epsilon} \leq \epsilon_U$ - Pareto constraints allow the derivation of terms of the form $\tilde{\epsilon}_i(\mathbf{x}, \tilde{\epsilon})$, revealing additional trade-off variables while describing the relationship that exist between the objectives at the Pareto-set. In this case, they are merely a representation of trade-off variables - albeit one which may have a large impact on the Pareto set.

When they are regionally active however, Pareto constraints reveal regional trade-off variables. This occurs when some of the constraints in the non-reduced model become active

for specific values of ϵ, causing discontinuous trade-offs. For instance, imagine a situation where the size of a mechanism is reduced while the system load is increased. At a certain point, the components will start to yield, causing a trade-off between the two. If the system is being designed with other objectives in mind as well (e.g. cost, weight, output, efficiency), this trade-off might only be relevant to the designer, if size and system load are of a large importance compared to the others.

In higher dimensional problems, $k \geq 3$, regionally active Pareto constraints might cause a Pareto frontier between an objective pair. Such situations can be studied through the *Analysis of Pareto efficient activity cases* procedure described in chapter 4. Thus, Pareto constraints may, when studied, reveal additional trade-off variables, or discontinuous relationships such as variables that are in trade-off in specific regions of the Pareto-set, thereby effecting its shape.

In summary, Pareto-set dependency analysis helps explain the relationship between the design problem and the shape of the Pareto set. Thus, we can utilise these theorems and corollaries to derive a set of redesign principles.

5.3 Configuration Redesign Principles

Insights into the relationship between the design problem and the shape of the Pareto set is of substantial value in the synthesis and improvement of configuration designs. As discussed in chapter 4, the Pareto set is created by variables and constraints that are shared between the objectives. Some shared variables can be used to improve upon several objectives simultaneously, while others cannot. To a large extent, these relationships are determined by the decisions made in conceptual and configuration design. Designers hence need to identify and manage global (i.e. shared variables) and regional dependencies (i.e. shared active constraints) at an early stage to reach *good* configuration designs.

Pareto set dependency analysis can help designers reason about changes in configuration to discover improved designs. Here, it is posited that experienced designers apply tacit knowledge of constraints [59] and trade-offs [33], to synthesise and improve configurations. Designers will typically use this knowledge to attempt to configure components of a system in a way that leverages harmonious variables to achieve a high performance, e.g. placing rotating components as far inside an assembly as possible and load-bearing components as far outside. Similarly, designers will attempt to obviate trade-offs through design change or identify acceptable compromises early on, as argued by Howard and Andreassen [124], Ahmed et al [33], and Althuller [16]. Given that trade-off variables cause the existence of the Pareto set, designers are thus either attempting to eliminate the underlying dependency or active constraint that introduces the trade-off variable, or attempting to identify design proportions that yield an acceptable compromise.

Reaching the required insights is not trivial, especially in highly interdependent systems. Pareto set dependency analysis bridges this gap, providing a causal link between the optimal result and the limitations of the configuration design. This allows more informed and deliberate identification, prioritisation, and handling of the dependencies that cause trade-offs. The introduced theorems, proofs, and corollaries demonstrate how certain types of model transformation based on the results of this analysis lead to an improved Pareto set. Translating these transformations into specific design changes would mitigate the dependencies that create the Pareto set and relax the constraints that position it, just as experienced designers do through tacit knowledge.

Based on the theorems and proofs in the previous section, it is apparent that from a mathematical perspective, there are a limited number of model changes we can introduce, which

result in an improved Pareto set. Each of these transformations have associated forms of design changes. This means that we can introduce a set of reconfiguration design principles.

When employed in configuration redesign, the following principles lead to an associated improvement of the Pareto set:

1. **Align Trade-off Variables**. Reduce or eliminate the effect of trade-off variables on the objectives without impacting their single-objective optima, thereby improving their alignment and the Pareto set, per Theorem 2. This involves eliminating the dependency, making the variable harmonious, or scaling it, per the corollaries to Theorem 2. These strategies may apply to both objective and constraint functions, as trade-off variables can be introduced through an active constraint. Such constraints are especially important, as they not only limit optimality of certain objectives overall, as their activity also introduces a trade-off with other objectives.

2. **Leverage Harmonious Variables**.Widen the feasible domains of harmonious variables in the improving direction, as per Theorem 3. This involves design changes that modify or delete the constraints that bound harmonious variables, striving towards letting $\overline{\overline{x}} \rightarrow \infty \wedge \underline{\underline{x}} \rightarrow 0$.

3. **Relax Pareto Constraints** - Relax globally active Pareto constraints, thereby aligning trade-off variables. Relax regionally active Pareto constraints, or eliminate inconsistencies that might exist between them beyond the Pareto set (i.e., in the infeasible region). This might change or eliminate the Pareto frontiers between certain objectives.

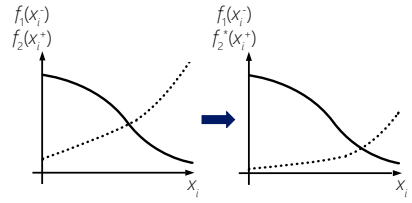
4. **Eliminate Parasitic Contributors**. Consider situations where it is not possible to widen the feasible domain of harmonious and independent variables; e.g. when their bounds represent unmodelled objectives or physical phenomena that cannot be circumvented. Such situations can introduce parametric or scalar contributions to the objectives that worsen their optima. Therefore, it may be better to eliminate the influence of these variables on the objectives rather than relax the constraint. The underlying constraints may represent important modelled objectives (e.g. cost driving manufacturing constraints), and it may hence be better to attempt to eliminate their influence on the modelled objectives entirely, than attempt to relax the constraint itself. A common example i variables bound from below, $\underline{\underline{\mathbf{x}}}$ by manufacturing constraints (e.g. minimum feature sizes, assembly feature dimension, and wall thickness. When such contributions then end up in the objectives or Pareto constraints, they can have a substantial impact on the optimum, which should be avoided when possible.

Within each principle, there are a number of more specific strategies stemming from basic model transformations that result in an improved Pareto set. Each represents an alternative way of implementing the principles and corresponds to certain forms of design change. Figures 5.4-5.6 illustrate each of the four principles and the available strategies within each principle.

These principles and associated strategies relate to specific variables and constraints, and they can be applied recursively to improve a configuration after its initial optimality. Essentially, they comprise the set of distinct forms of design change (i.e. transformations to the optimization model) one can introduce based on the outputs of Pareto-set dependency analysis, which would result in design improvement following definition . Each of the underlying strategies are alternative ways of implementing the principles, stemming from basic model transformations that result in an improved Pareto set, which correspond to certain forms of design change. Hence, given an overview of the trade-off variables, harmonious variables, active constraints, and Pareto constraints in a design problem, the designer can apply

Align Trade-off Variables

SCALE

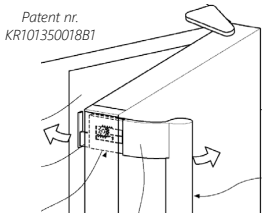


Mathematical transformation
Any design change adding a multiplier or divisor (a parameter or variable) to a trade-off variable in one objective allows scaling of the trade-off. Examples of such transformations include:

$$f_1(x_i^+) = x_i^2 \quad f_1^*(x_i^+, x_j^-) = x_i^2/x_j$$
$$f_2(x_i^-) = -x_i \rightarrow \quad f_2(x_i^-) = -x_i$$
$$f_1(x_i^-) = -x_i \quad f_1^*(x_i^-, x_j^-) = -x_i x_j$$
$$f_2(x_i^+, x_j^-) = x_i - x_j \rightarrow \quad f_2(x_i^+, x_j^-) = x_i - x_j$$

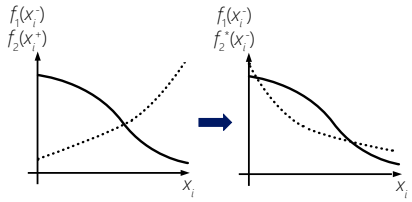
Typical design changes
Trade-offs can be reduced through a wide range of design changes that scale one objective but not the other, ranging from the addition of lubrication to new subsystems, features, and interfaces. The introduction of gearing and mechanical leverage in general, load balancing, lubrication, and intermittent kinematic constraints (e.g. for nonlinear stiffness), are all examples of scaling solutions.

Related heuristics: Amplification and filtering [32], manage friction [3,5,13], local quality in TRIZ [12], decoupling [11], and leverage/gearing [2,5].



Example: *Fridge door mechanisms*
To ensure efficient cooling, refrigerator doors are tightly sealed when closed, which is achieved with pretension of the door with a rubber seal. Combined with negative pressure inside the fridge due to cooling, this results in a high opening force. Several designs scale down this *efficiency vs. opening force* trade-off e.g. with auxiliary opening mechanisms and pivoting lever handles.

FLIP MONOTONICITY



Mathematical transformation

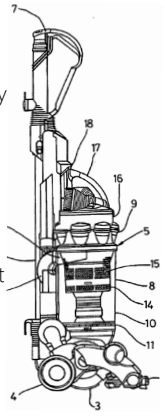
$$\begin{matrix} f_1(x_i^-) \\ f_2(x_i^+) \end{matrix} \longrightarrow \begin{matrix} f_1(x_i^-) \\ f_2^*(x_i^-) \end{matrix}$$

Any design change that inverts the monotonicity of one objective w.r.t. a trade-off variable, while the rest are unchanged, effectively makes the variable harmonious. In nonlinear terms this might be achieved by changing the bounds of other variables that act as multipliers or divisors to the trade-off variable.

Typical design changes
While somewhat challenging, making a trade-off variable harmonious can be achieved in certain circumstances, especially if the dependency stems from an active constraint. Changes such as the inversion of components and interfaces, *"self-helping"* systems, redistribution of sub-functions, changes in working directions and load paths, or the use of a different working principle, can result in a change in monotonicity.

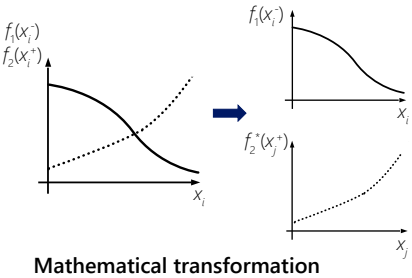
Related heuristics: "The other way round", nested doll, and self-help [12] Principles of self-help and force transmission [2],

Example: *Dyson Vacuum*
Bag-based vacuum cleaners are generally affected by a trade-off between filtration quality and suction pressure; the tighter the filter the larger the pressure loss. Vacuums that rely on cyclonic separation where filtration increases with the pressure, get around this issue. While the example is conceptual, as it relates to a change in filtration principle, it illustrates the general idea.



Patent nr.
EP 1786568 B1

SEPARATE



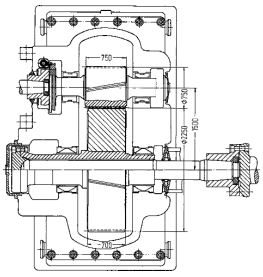
Mathematical transformation

$$\begin{matrix} f_1(x_i^-) \\ f_2(x_i^+) \end{matrix} \longrightarrow \begin{matrix} f_1(x_i^-) \\ f_2^*(x_j^+) \end{matrix}$$

Any design change that makes an objective independent of a trade-off variable - either through substitution or elimination, mitigates the trade-off, unless the objectives share additional trade-off variables.

Typical design changes
Separation is a widely used principle, involving changes such as the splitting parts, change in working axis and load direction, parallel subsystems, asymmetry, or the avoidance of "unintended" dependencies through exact constraint design. It may result in an increased number of parts, but can also involve the redistribution of functionality amongst the parts of the system. Unlike in other frameworks, the approach here is to only apply separation to trade-off variables.

Related heuristics: Independence axiom [11], division of tasks [2], separation in space, time, or condition [12].

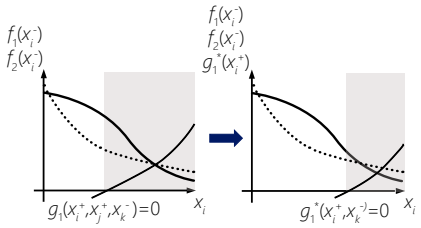


Example: *Siemens-Maag Gearbox* [2]
This gearbox drive shaft design described by Pahl and Beitz [2], is a prime example of this principle. The drive shaft has been split in two to eliminate a trade-off between efficiency and wear; the stiff outer shaft transmits the torque from the gears, while the flexible inner shaft is free to absorb oscillations, protecting the gears.

Figure 5.4: *From Paper B:* The strategies within Principle 1: Reduce or eliminate the impact of a trade-off variable upon an objective pair

Leverage Harmonious Variables

SHIFTED BOUNDS



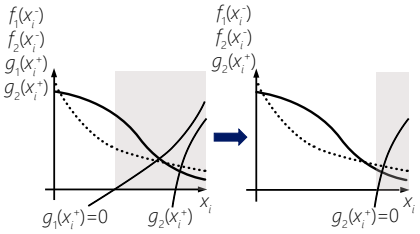
Mathematical transformation
Any design change that shifts the *glb/lub* of a harmonious variable, improves the optimum of its dependant objectives. This involves eliminating increasing contributors, introducing additional decreasing contributors, or scaling parts of the constraint. This can also scale a trade-off, when the harmonious variable is a multiplier or divisor of a trade-off variable.

Typical design change
In configuration design terms, these changes are specific to the type of constraint. Generally speaking this is oft matter of positioning components in an assembly in the most beneficial way - e.g. locating parts with decreasing variables as far *inside* an assembly as possible and increasing variables on the outside. Further, it involves designing to avoid unnecessary contributors to the active constraint, e.g. stress concentrations and associated loads in structural constraints.

Related heuristics: Reduce information content [11], Principle of balanced forces [2], Minimise tolerance paths [5]

Example: *Mazda Skyactiv-G* [33]
In the design of combustion engines, thermal efficiency increases with the compression ratio. Yet, this ratio cannot be increased beyond a point where knocking occurs, which is in part driven by residual gas after combustion. Most petrol engines hence have a ratio between 8:1-12:1. In the Skyactiv engine, Mazda pushed this ratio 14:1, using a longer exhaust manifold, increasing gas scavenging, and shifting the knocking constraint.

CONSTRAINT RELAXATION

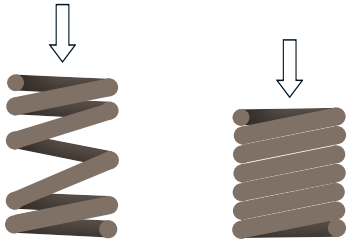


Mathematical transformation
When a harmonious variable is actively (but not critically) constrained, we might try to change the configuration design in a way that eliminates the active constraint. This shifts the *glb* or *lub* of the variable to the next constraint, improving the optimum of all its dependent objectives.

Typical design changes
As with shifted bounds, the changes required to eliminate a constraint, are contextual. Examples include changes aimed at redirection of force paths to eliminate a load case, a new part structure to avoid certain parts being bound by limiting geometric constrains (e.g. one part inside another), a change in assembly sequence to avoid some alignment constraint.

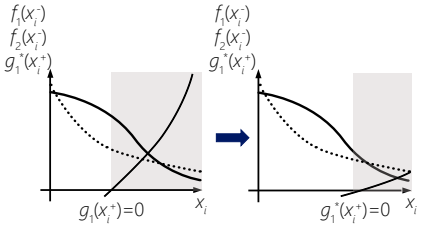
Related heuristics: Vary the structure of main elements [14,], redirect load path [5], merge parts [5,13], shielding [33].

Example: *Spring strength at block*
A well known example of load path redirection, compression springs are self-



reinforcing when deformed to their block length. In design applications where a maximum load resistance is desired, a spring design that is deformed to its' block length rather than to its' elastic yield limit is far stronger. Introducing such a change to a design, is equivalent to eliminating the yield constraint driven by shear stress.

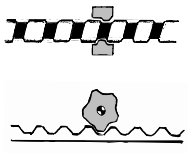
NEW FUNCTIONAL FORM



Mathematical transformation
A complete change in functional form of an active constraint, may yield a widened feasible domain. This is distinct from *shifted bounds*, in that it involves the entire function, and may hence result in changed constraints, monotonicity, exponents, and so on.

Typical design changes
Such a drastic model change will most likely probably require substantial design change, e.g. a change in components, working principles, and/or the physics of the problem. Examples of such include a change in production process, the separation or combination of parts, a change in the realisation of a given sub-function, a change in load type and distribution, and so on.

Related heuristics: Design for pure compression and tension [5], Select rotary over linear motion [3,5] Self-help [3,12]

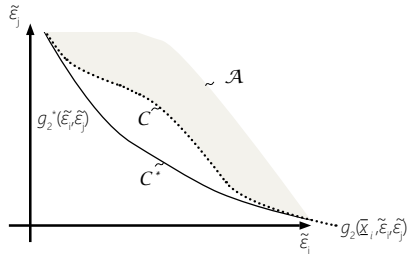


Example: *Rotary to linear movement*
A rack and pinion and a lead screw fundamentally meet the same functional purpose - to convert rotation into linear motion, or vice versa. Yet, what is superior, depends on the objectives, primarily due to quite different constraints involved in their design. For instance, the rack for instance only slides, and as a result the mechanical stress expressions are quite different, compared to the rotating screw, which is why they are commonly used for high load applications.

Figure 5.5: *From Paper B:* The strategies within Principle 2: Increase the influence of harmonious variables

Relax Pareto Constraints

SHIFTED PARETO CONSTRAINT



Mathematical transformation
The elimination of increasing contributions to Pareto constraints shifts the frontier regionally or globally. Transformations include the elimination of parametric-, variable-, or objective contributions, e.g.:

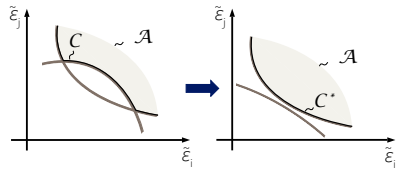
$$g(\mathbf{x}, \tilde{\epsilon}_i^+, \tilde{\epsilon}_j^-) \longrightarrow g^*(\mathbf{x}, \tilde{\epsilon}_j^-)$$
$$g(x_1^+, x_2^-, \tilde{\epsilon}_i^+) \longrightarrow g^*(x_2^-, \tilde{\epsilon}_i^+)$$
$$g(\mathbf{x}, \tilde{\epsilon}_i^+; P^+) \longrightarrow g^*(\mathbf{x}, \tilde{\epsilon}_i^+)$$

Typical design changes
In many ways, the design changes and model transformations involved here, resemble those of *Leverage Harmonious variables*. Shifting Pareto constraints is tantamount to reducing the equilibrium that exists between the objectives in certain (or all) regions of the Pareto set, due to the activity of constraints. Thus, this might involve rearrangement of parts, change in load distribution, and so forth.

Related heuristics: Reduce information content [11], principle of balanced forces [3], vary the structure of elements [14], redirect load path [5]

Example - *Additive Manufacturing and Topology Optimization (TO)*
In industrial practice, TO efforts are usually actively constrained by material and the manufacturing constraints. In this context, the utility of additive manufacturing is broadly cited, as it essentially shifts several manufacturing constraints, e.g. allowing hollow geometry and undercuts, and shaping not being limited by tooling directions. This has allows increasingly light load bearing components, reducing the trade-off between stiffness and mass. While this is more a process change than a design change, it serves to illustrate the model transformation.

DEPENDENCY REDUCTION



Mathematical transformation
Reduction of the dependencies between Pareto constraints, that either result in trade-off variables, variables with empty feasible domains beyond the Pareto set (i.e. *two-sided failure*), or regional bounds for $\tilde{\boldsymbol{\varepsilon}}$, will change the optimal set. An example of such a transformation is:

$$\begin{matrix} g_1(x_1^+, \tilde{\epsilon}_i^-) \\ g_2(x_1^-, \tilde{\epsilon}_j^+) \end{matrix} \longrightarrow \begin{matrix} g_1(x_1^+, \tilde{\epsilon}_i^+) \\ g_2^*(x_2^-, \tilde{\epsilon}_j^-) \end{matrix}$$

Typical design changes
The design changes and model transformations involved here, resemble those of *Align trade-off variables*. The difference is that these might be regional trade-off variables. Hence, it is equally impactful to introduce changes to the eliminated active constraints that contribute to the Pareto constraint, creating the dependencies. Examples include inverting components and interfaces, eliminating load cases, change in working axis and load direction.

Example: *FlexTouch Safety Mechanism*
Insulin pens cannot be dialled to a dose setting beyond the amount of insulin left. An "*end of content*" locking mechanism prevents the user from receiving a smaller dose than has been set. Such locks need

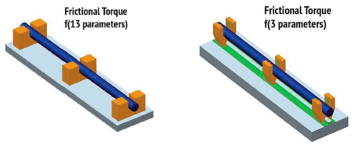
to withstand substantial loads when users unknowingly attempt to get beyond this limit. Ultimately, this affects the achievable combination of device size and dose setting torque (which is important to users with limited dexterity). In the FlexTouch^R device, the dial is connected to the dose setting mechanism via a flexible spline connection, which disengages if the user attempts to set a dose beyond what is left. No load is transferred, protecting the pen from overloading, and eliminating a dependency between the size and torque caused by the yield constraint.

Eliminate Parasitic Contributions

Mathematical transformation
This category is too broad to provide a universal mathematical transformation, but all the sub-types of transformations are well known. These are the removal of parametric and scalar contributions which increase the optimal value of one or more objectives (without decreasing any), and the elimination of harmonious variables that are bound in such a way that they cannot be *leveraged*. These may be unintended contributions (i.e. from design error), or contributions involving active constraints with little room to introduce further relaxation (e.g. a wall thickness constraint).

Typical design changes
There are many types of contributions that are parasitic. Examples the negative impact of undesired vibrations, electromagnetic fields, heat, parasitic loads, friction, unintended contact points, and manufacturing and assembly features. Design changes mostly involve efforts to remove these effects from performance critical part geometries or locations in the assembly. E.g relocating assembly features to another cross section.

Related heuristics: Exact constraint design [3], reduce information content [11], avoid associated loads [13], shielding [34].



Example: *Overconstrained axle* [34]
An oft cited example in kinematic design, over-constrained axles cause major issues w.r.t. production and efficiency. The typical design error is to increase system stiffness by introducing more radial bearings, resulting in statical indeterminacy. This issue can be reduced or resolved entirely by using fewer or different bearings.

Figure 5.6: *From Paper B:* The strategies within Principles 3 & 4: Reduce regional trade-offs and eliminate parasitic contributions

these principles and underlying strategies to identify potential routes to design improvement through configuration change.

These redesign principles and underlying strategies are very general in that they are applicable to any optimization model that has been studied through Pareto set dependency analysis. Yet, their specificity becomes evident when considering a specific design problem. Just as design 'goodness' is contextual to the objectives at hand, so are the design improvements. The principles are thus not intended for use in initial configuration design; rather, they motivate designers to identify potential design improvements after careful quantitative analysis. Thus, the process of optimization becomes a driver for configuration design improvement. It is worth noting that Principles 3 and 4 are essentially special cases of Principles 1 and 2. Globally active Pareto constraints can be seen as representations of trade-off variables stemming from the selected model formulation. Parasitic contributors meanwhile, can be viewed as harmonious or single objective variables that are bound in such a way that they cannot be leveraged to move the Pareto-set.

As the examples in Figs.5.4-5.6 illustrate, the forms of design change involved are quite typical in product design. The redesign principles are also related to well-known design heuristics, albeit with key differences. First, they are opportunistic but have a rigorous foundation and are therefore valid independent of context. Second, they are applied following optimisation and Pareto-set dependency analysis, letting designers rely on analysis results rather than intuition to identify which heuristic to apply where. For example, heuristics such as *separation* [16] and *independence* [13] or *division of tasks* [6], prescribe avoidance of dependency. Through the theorems and proofs in sections 4.2. and 5.2. we see this is actually only relevant for trade-off variables, if the aim is to improve performance and avoid trade-offs.

Admittedly, these principles and strategies are prescribed at a certain level of abstraction. One might dig deeper, and regard the more specific forms of changes one might make to the optimization model using each of the strategies. The lower the level of abstraction the more related these modes of change become to existing qualitatively founded and context specific heuristics.

5.3.1 Relationship with existing heuristics

As discussed in chapter 3, engineering design literature is rife with prescriptive heuristics, which collate different best practices in design. These exist on different levels of generality, ranging from abstract inductive guidelines aimed at supporting synthesis such as French' "*clarity of function*" [98], to more reductive and specific guidelines such as "*avoid press-fits*" as prescribed by Skagoon [96].

As found by Fu et al. in a review of existing design principles and guidelines [128], most heuristics have been developed through the analysis of existing designs (i.e. identifying "good" or "bad" solutions), through the author's own experience, or based on observations in design practice. Hence, most design heuristics are based on limited data sets and are inherently contextual given their extraction from specific existing designs and development contexts.

Among the prescriptive design heuristics that focus on the end result rather than the design process itself, there is a general notion that some approaches and solutions result in *good* designs. Viewed from the design optimization perspective, this implies that heuristics, by their very nature, presuppose a set of underlying design objectives. While some design objectives (e.g. low cost) might be universal, others are specific to the purpose of the system being designed and how it has been embodied.

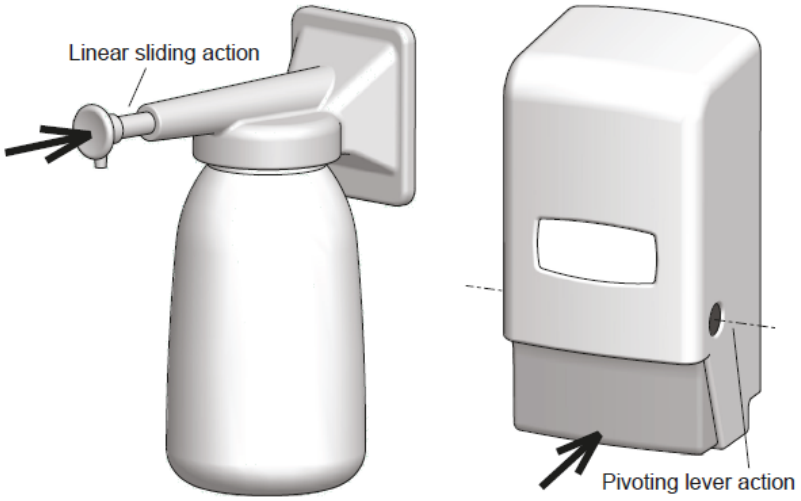


Figure 5.7: *Figure from [96]:* Skagoon uses this design of soap dispensers[96] to illustrate why rotational movement can be preferable over linear movement. In the linear soap dispenser, the friction and risk of self locking is reduced as the length of the guides of the linear slider is increased. In the embodiment with the rotational joint, friction is reduced as the diameter of the joint is reduced.

Example - *Rotary over linear motion*
In the context of kinematic design, it is widely prescribed that rotary joint are preferable to linear joints. Examples of such heuristics include French' "Prefer pivots to slides and flexures to either" [98], and Skagoon's "*avoid sliding friction*" and "*Rotary motion over linear motion*".

French and Skagoon both argue that rotational joints allow for less friction and reduce the risk of self-locking behaviour (often referred to as the "sticky drawer effect"[98]). As a result, both also prescribe that if sliding friction cannot be avoided, one should at least ensure that the linear joint has "long guidance bases" [98]. This set of heuristics has an underlying set of presumed objectives; that a "good" end-design is a *small* system which is *mechanically efficient*.

It is certainly hard to argue that these are uncommon objectives in mechanical design. However, other objectives might be at play in many situations, which render these heuristics irrelevant. Consider two common machine elements; the lead screw and the rack-and-pinion (illustrated in figure 5.8), and variants thereof (ball screws, helical rack and pinions, etc.). Both serve the same basic purpose; converting rotational movement into a linear movement with a degree of positional accuracy. Yet, depending on the application, one will be preferable to the other.

The mechanics of lead screws and rack-and-pinions are inherently different. The mechanical efficiency in rack-and-pinion is dependent on the distance between the pinion and the joint that the rack slides on (l_1 in fig. 5.8), the pitch diameter of the pinion (d_o), and the bearing diameter of the pinion (d_i). In the lead screw, meanwhile, friction decreases with the reduction of the contact diameter between the screw and the nut and the screw and its end bearing. Using a ball interface instead of a key also eliminates the sliding friction that occurs between key and screw.

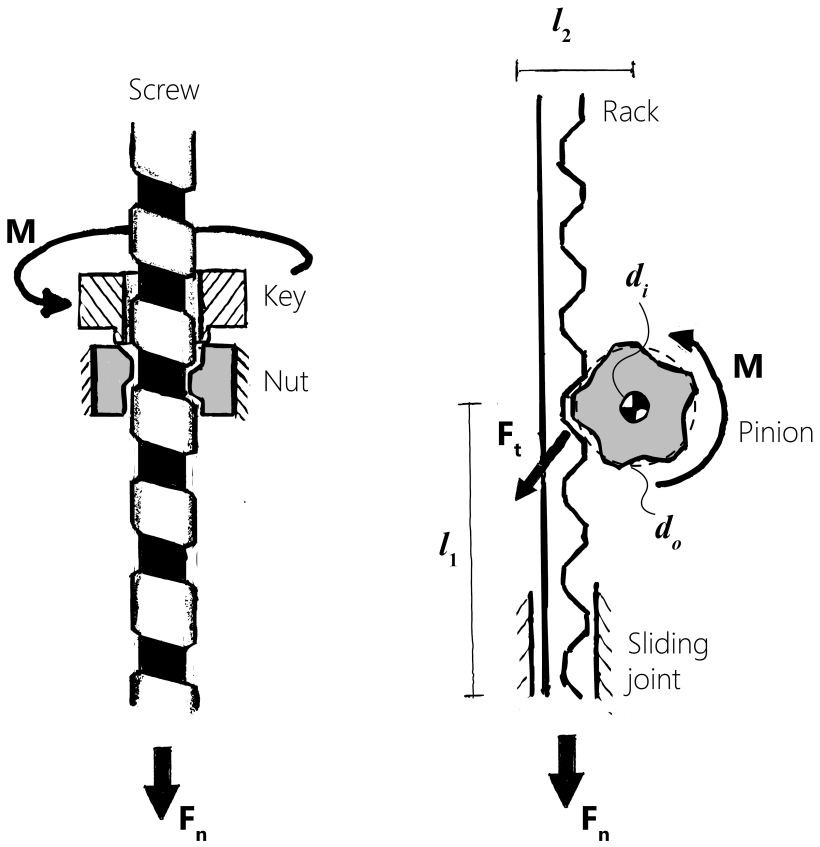


Figure 5.8: Lead screws (*left*) and rack-and-pinions (*right*) have several key differences

This means that the efficiency of the rack-and-pinion ultimately increases with the overall size, whereas the efficiency of lead screws increases as their size is decreased. Hence, if one were designing a system where the only objectives are to minimise size and maximise efficiency, then the heuristics by French and Skagoon are indeed relevant. However, inspecting how supplier catalogues and trade magazines describe the two alternative solutions reveals that if we are designing with load resistance or serviceability in mind, then the rack and pinion can be argued to be superior [129, 130].

The purpose of the inclusion of this example is not to claim that one of the two is superior. Rather, the example serves to illustrate that many heuristics are only valid and applicable in the face of certain design objectives. Alternatively, they are so general that they do not prescribe or reveal concrete embodiment design changes but rather act as a general ambition or attribute that designers can keep in mind during synthesis and redesign. Examples of such include the principles of *Self Help* and *Division of Tasks* prescribed by Pahl and Beitz.

This reveals a somewhat unique characteristic of the *Configuration Redesign Principles* suggested in this chapter. As they rely on the outputs of the generic analysis of the root causes of trade-offs between the objectives modelled in an optimization model, they are independent of context. They can be applied to trade-offs between any pair of design objectives, so long as they are described through an explicit or numerical model, which can be analysed and reduced through (global- or regional-) monotonicity analysis.

That is not to say that the redesign methodology prescribed in this chapter is completely

unrelated to existing heuristics. As alluded to in the one-pagers, the principles and strategies are abstractions made at a certain level of decomposition. Looking at the strategies, for instance, many of them can be achieved through different means. Exploring these from a mathematical perspective reveals that the overarching strategies (*separate*, *flip monotonicity*, *shift bounds*, etc.) are somewhat anecdotally related to existing design heuristics.

If we, for instance, look at *Separation*, there are, in fact, different forms of mathematical transformations that result in separation. When employed in a design problem with the trade-off variable $\underline{\overline{x_i}}$ that influences an objective pair, the transformation could involve:

- *Substitution with an existing variable:*
  E.g $f_1(x_i^-, x_j^+), f_2(x_i^+) \longrightarrow f_1(x_i^-, x_j^+), f_2(x_j^+)$. Here, the trade-off variable x_i is substituted with an existing variable in f_2, making f_2 independent of x_i. In design, this may for instance correspond to reallocating functionality to different components or changing the axis of operation of a given function. This is somewhat analogous to *Another Axis* from TRiZ [16], and *division of tasks for distinct functions* by Pahl and Beitz [6]

- *Substitution with a new variable:*
  $f_1(x_i^-), f_2(x_i^+) \longrightarrow f_1(x_i^-), f_2(x_j^+)$. Here, the trade-off is eliminated by substituting x_i with x_j, an entirely new variable. In the context of design change, this might involve the introduction of new parts or subsystems; related heuristics include *Differentiation* [29], *Segmentation* from TRiZ, *decoupling* or *uncoupling* from Axiomatic Design [13].

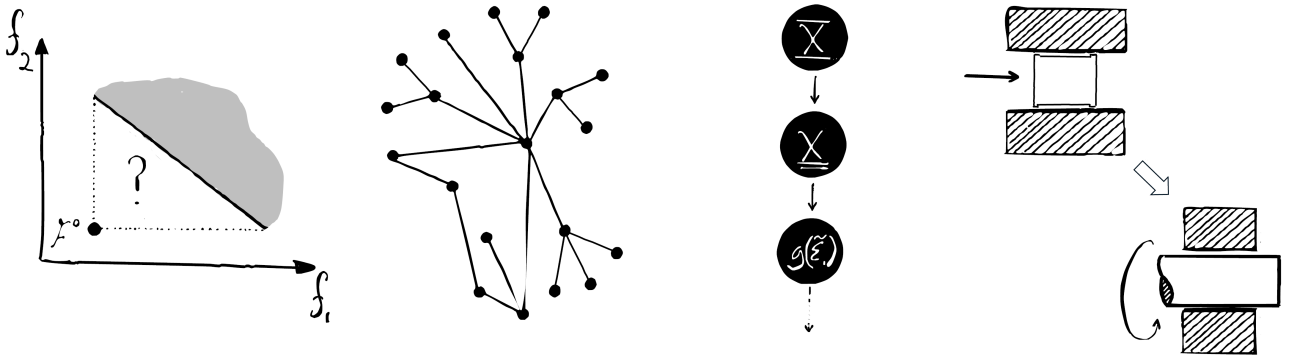
- *Elimination*:
  $f_1(x_i^-), f_2(x_i^+, x_j^-) \longrightarrow f_1(x_i^-), f_2(x_j^-)$. Here, the influence of the trade-off variable x_i is simply eliminated from f_2. This is only possible when the influence of x_i upon f_2 is avoidable, meaning it is analogous to "*avoid dependence on irrelevant variables*" prescribed by French [98], and to numerous heuristics which aim to avoid common errors in design such as the *Principle of Balanced Forces* by Pahl and Beitz [6] and exact constraint design [14].

The same goes for the remainder of the strategies under *align trade-off variables* and *leverage harmonious variables*; at a greater level of decomposition, they can be achieved through different transformations. *Shifted bounds* can, for instance, involve the addition of monotonically decreasing variables or the removal of increasing variables to an active constraint function. In a constraint describing a load case, this might involve making changes that increase the achievable size of load-bearing areas or reduce the contributions to the load being distributed.

One might hence claim that numerous heuristics - such as *Division of Tasks for Identical Functions* [6] and *Avoid cut-outs* [98] - are simply contextual representations of the overall idea behind shifted bounds; to widen the feasible domain in an optimizing direction through design change. Therefore, the configuration redesign principles are in effect the general forms of design change one can introduce in order to reduce trade-offs and improve optimality overall. This addresses a limitation of context-specific heuristics; that it can be difficult to know which heuristics to apply and when to actually apply them in order to achieve improved performance.

5.4 Systematic Configuration Design Improvement

The previous sections may seem excessively formal compared to many design frameworks. However, without the insights MOMA and ϵMA provide, one might introduce changes to eliminate dependencies or relax constraints that have no bearing on the Pareto set, or even



1 PARETO SET DEPENDENCY ANALYSIS
Apply the MOMA and εMA procedures to the optimization problem to identify the underlying trade-off variables, harmonious variables, and Pareto constraints.

2 MAP OUT THE REDESIGN OPTIONS
Consider which optimality principles can be applied where. Rule out any option that requires impossible changes, e.g. making mass and stiffness independent.

3 IDENTIFY REDESIGN SEQUENCE
Develop a redesign sequence for the implementation of the remaining redesign options, following the analysis driven prioritisation scheme.

4 ITERATIVE SOLUTION EXPLORATION
If a selected principle seems infeasible to apply, try an alternative sequence or principle. Stop and evaulate the redesign when new trade-off variables are introduced.

Figure 5.9: *From Paper B:* Configuration Redesign Process - A redesign procedure supported rigorous analysis

accidentally worsen the set. Ultimately, the above principles hence come down to a more targeted approach for dependency reduction and constraint relaxation - design practices widely advocated [6, 13, 16, 24, 59, 131]. Thus, the aim of the optimality improvement principles derived here, is to guide design change towards reducing trade-offs, and grow the design space in a beneficial direction, while eliminating regional issues in the Pareto set. Here, rigorous analysis is necessary, as we need to know which shared variables and constraints create problems in the first place, if we wish to drive improvement through configuration design changes.

As alluded to, the strategies within the same principle are mutually exclusive. For example, we cannot make an objective function independent of a trade-off variable through separation while also scaling the same variable. Depending on the problem, some design changes are also more influential or easier to implement than others. Thus, it is beneficial to map out all options for improvement after analysis and select the most promising ones, rather than randomly applying the principle. While the strategies and underlying principles have a quantitative foundation, the designer must still determine which principle to apply to each variable and constraint, and in which sequence. As summarized in fig. 5.9, this thesis thus proposes a systematic configuration redesign procedure involving said mapping and prioritization steps between analysis and design change

The first step in this procedure, is to perform Pareto Set Dependency Analysis. This is followed by a step involving the mapping the redesign options; i.e. the assessment of which design strategies can be applied to which variables and constraints.

In many cases, some of the strategies will not be possible to implement. Using the SOMA device as an example, the spring will always have a mass and a position in space, meaning it will influence the center of mass of the device. As we cannot make the position of the center of mass independent of the spring, *separation* is impossible, unless there is no spring, which would involve a change of a more conceptual nature. We can, however, look into introducing design changes that *scale* the impact of the spring's mass, or that relax the constraints

on harmonious variables that contribute to its mass or position. Correspondingly, it is not possible to eliminate the spring yield constraint entirely, as this would require a material that does not yield or break. As such, *constraint relaxation* is not possible, but we could explore *shifting bounds*, e.g. by changing the design in a manner renders the constraint inactive.

This mapping is followed by a critical element; the use of a prioritisation scheme in Step 3 to identify a redesign sequence. Prioritisation is necessary, given that the principles are mutually exclusive, and given that some changes might be more influential than others, depending on the results of Pareto set dependency analysis. Hence the following scheme is suggested, which utilises the outputs of analysis and computation. This scheme is determined by two factors; the magnitude of the potential influence of the change and the ease of implementation:

1. Eliminate parasitic influences.
2. Leverage the harmonious variables, attempting *relaxation* rather than *shifted bounds* when possible. Only leverage the variables that:
    - influence multiple objectives,
    - have a multiplying effect on a trade-off variable.
    - are bound by a constraint with a comparatively high Lagrange multiplier.
    - are actively constrained in a manner that introduces a new trade-off variable or a contribution to a Pareto constraint
3. Relax Pareto constraints that depend on more than one $\tilde{\epsilon}$ variable and/or are globally active
4. Align trade-off variables in an order based on the number of influenced objectives and on the relationship between F^* and $\underline{\overline{\mathbf{x}}}^*$. To avoid increasing system complexity, apply flipped monotonicity over the other strategies, and separate over scale unless separation only is possible through the introduction of new variables.
5. Leverage remaining harmonious variables and relax remaining Pareto constraints

The logic behind this prioritization is to ensure that independent issues and obvious design errors are handled first (i.e. parasitic contributions) and then look for changes that result in the largest improvement to the entire set. The step order is defined based on the observation that the globally active Pareto constraints and the harmonious variables in Step 2 will, in most of the outlines conditions, exceed the influence of single trade-off variables. Alternatively, one could base the sequence on objective weighting.

Using this sequence, one can subsequently explore the design changes required in order to achieve the desired change in dependency or constraint activity. Given that each of the strategies has associated forms of design change, they provide a starting point the exploration of changes.

This redesign procedure can be used iteratively and requires a combination of analysis, qualitative reasoning, engineering judgement, and creativity. The given mapping, prioritization, and solution exploration are suggested to increase the likelihood of successful redesigns. Ultimately, the required form of change is context specific, but as shall be shown in the SOMA case at the end of this chapter, the actual changes involved can be relatively simple.

5.5 Redesign of the SOMA Device

Using the results of the Pareto set dependency analysis performed on the SOMA device in the previous chapter, we can explore the application of the redesign methodology. This is also described in Paper B, albeit in less detail and in more condensed form.

5.5.1 Redesign Mapping and Sequence

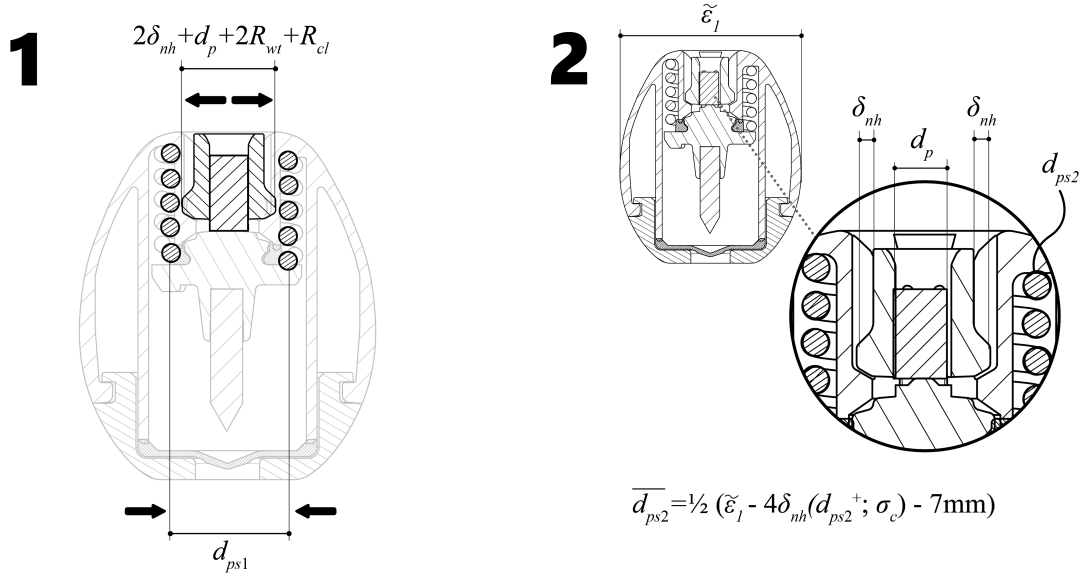


Figure 5.10: *Adapted from Paper B:* The relationships in the SOMA device that introduce trade-offs. 1) As the diameter of the trigger system increases with the spring force, the coiling diameter can not be used to increase the spring force, as this reduces the space available for the trigger. 2) When the trigger arms collapse inward as the device is triggered, they cannot collide with each other. As a result, the diameter of the plug d_p is given by the trigger overlap δ_{nh}, which grows with the spring load.

In the previous chapter, we applied MOMA and ϵMA to identify four key issues in the configuration design of the SOMA device, which cause or worsen the trade-offs between the modelled design objectives; the position of the centre of mass, device diameter, API payload, and injection impact velocity. These are illustrated in figures 5.10-5.11. Most notable among these issues is that the spring fits around the trigger (nr. 1 in figure 5.10). The spring force (and by extension impact velocity) cannot be increased by decreasing the spring coiling diameter, d_{ps1} as this reduces the space available for the trigger. Furthermore, any increase in spring force via an increased spring wire diameter, d_{ps2}, results in a larger trigger overlap δ_{nh} in order to distribute the increased load over a large area. This, in turn, increases the size of the plug d_p to ensure that the trigger arms do not collide, in effect causing a larger coiling diameter d_{ps1}. These two issues greatly limit the volumetric efficiency of the actuator system, affecting the achievable velocity for any given device size and worsening the influence of d_{ps2} on size and self-orientation. As the optimal size, $\underline{\tilde{\epsilon_1}}$, is determined by a radial fit constraint, it generally seems inopportune that the spring force is absorbed over an area in the radial direction. As such, a trade-off will always exist between velocity and size unless we find a more space-efficient way of distributing the load while making the d_{ps1} less dependant on the trigger design and vice versa.

These issues actually reveal a trade-off with an objective that is included in the optimization model as a constraint - *shelf life*. The SOMA device would, like any other pill, need to be storable for a certain amount of time - both in the manufacturer's warehouse, in transit, at

the pharmacy, and finally with the end-user. The spring exerts a static load on the trigger system, which consists of injection-moulded polymer parts, which are prone to creep and, in some cases, creep fracture. Hence, the longer the system is under static load, the larger the risk of creep fracture. This is included through constraint g_{11} in the model. In principle, we could change the allowable static stress σ_c to shift the Pareto frontier of size and velocity, but this would come at the cost of the shelf life of the device.

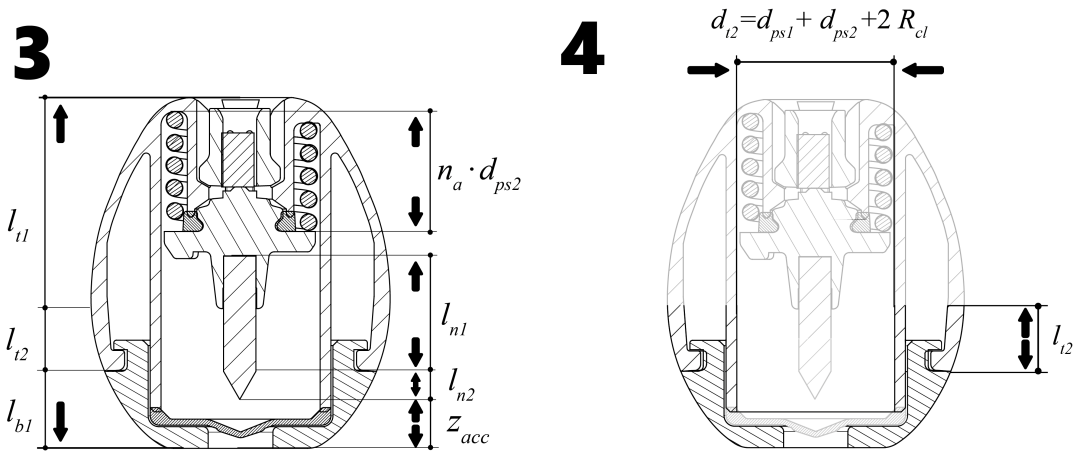


Figure 5.11: *Adapted from Paper B and continued from the previous figure:* 3) The serial arrangement of components in the axial direction introduces several trade-off variables, as any increase in length of the internal components comes at the cost of an increased device size. 4) The design of the snap between the housings ultimately results in less space if the actuator and trigger.

Furthermore, vertical serial arrangement of the internal components results in several trade-off variables (nr. 3 in figure 5.11). In the Pareto set, any increase in needle length either comes at the cost of the acceleration stroke, z_{acc}, spring length, or the height of the device, which increases the diameter (given that the outer shape is predefined). This arrangement also causes a trade-off between velocity and API payload. Further, the amount of spring coils, n_a is given by the spring yield constraint, meaning $n_a^* = n_a(d_{ps2}^+, z_{pre}^+; \sigma_{ps})$. Thus, it multiplies the negative influence of d_{ps2} on self-orientation, as the spring-mass is mounted at the top of the device, and the number of coils increases with the wire diameter.

Finally, the parasitic contributions introduced by the assembly features in crucial cross-sections detrimentally affects the Pareto set. Specifically, the design of the snap-fits between top and base housing and the guiding cylinder in the top housing mean that letting $l_{t2} \rightarrow \overline{l_{t2}}$, to shift the centre of mass is shifted downward, reduced the radial space available for the actuator and trigger.

These issues greatly affect the Pareto set. To mitigate them, we can apply the configuration redesign principles to reduce the trade-offs and improve optimality overall. Yet, some of the underlying strategies might not be applicable in this context, meaning we need to map out the options before proceeding.

Looking back at the trade-off variables identified through Pareto set dependency analysis (shown in table 5.1), some of the trade-off variables in the problem cannot be avoided through *separation* or *flipped monotonicity*. An example of such is d_{ps2}; the spring will always have a wire diameter that contributes to its diameter and mass, thereby affecting the device diameter and position of the centre of mass. Thus, we cannot apply the aforemen-

tioned strategies. What we can do, however, is to explore how we might *scale* its influence. This might, for instance, involve the attempt to move the spring further down in the assembly to reduce its negative impact on the position of the centre of mass. As seen in the table, numerous constraints also worsen its influence on the trade-offs - these stem from issues 2 and 3 (shown in Figs. 5.10-5.11). Thus, exploring changes to the axial arrangement of the components, and a redesign of the trigger system, might relax these constraints, reducing the detrimental impact of d_{ps2}.

Similarly, d_{n1} is an intrinsic trade-off variable. As it needs to pass through the hole in the base, any increase in this diameter results in a larger hole in the base, worsening the mass distribution. As opposed to d_{ps2}, we have no way of scaling its influence or relaxing the constraints that determine the size of the hole. Thus, this trade-off variable is out of the scope of the redesign efforts.

Looking at the harmonious variables, d_{ps1} is especially influential, c.f. the impact velocity objective function derived in chapter 3, its influence on spring mass, and its contribution to the size of the device. As determined in chapter 4, the spring index in all of the identified Pareto points is higher than the ideal, meaning that less energy is stored pr. unit of volume than what is possible. In principle, all three of the *leverage harmonious variables* strategies are applicable, given that $\underline{d_{ps1}}$ is determined by the trigger-spring fit constraints in the entire Pareto set. This is a constraint caused by the configuration of parts and not one that is intrinsic to the design of springs.

	$\overline{\underline{d_{t1}}}$	$\overline{\underline{l_{t2}}}$	$\overline{\underline{l_{b3}}}$	$\overline{\underline{d_{ps2}}}$	$\overline{\underline{l_{n1}}}$	$\overline{\underline{d_{n1}}}$
f_1 - Self Orientation	-	-	-	+	+	+
c_1 - Diameter	+	(+)		+	(+)	
c_2 - API Payload	(-)		(+)	(+)	-	-
c_3 - Impact Velocity	-	(-/N)		-	+	+
Caused or worsened by	h_1,h_8, g_1, g_{20}	h_8,g_1,g_{20}	g_{20}	h_8, g_1, g_5, g_8, g_{10}, g_{11}, g_{12}, g_{20}	h_8,g_{20}	g_{31}

Table 5.1: From chapter 4: An overview of the key trade-off variables, and the constraints that either cause their introduction into the objective functions, or increase their influence.

This line of thinking was applied to all of the trade-off variables, harmonious variables, and Pareto constraints. The result was a set of potential design changes driven by the redesign principles. To determine a suitable redesign sequence, we use the conditions described in the previous section, along with the Lagrange multipliers of the constraints and variable-objective plots. The resulting sequence is:

1. *Eliminate parasitic contributions*: Reduce or eliminate the parasitic contributions of the housing snap and needle-hub interface upon the radial and axial fits, and upon the objective functions
2. *Shifted bounds*: Shift $\underline{\underline{d_{ps1}}}$, leveraging its harmonious influence, which is to the third power w.r.t. velocity.
3. *Pareto constraint dependency reduction*: In some activity cases, the trigger interface stress becomes a Pareto constraint of the form $g_{11}(\tilde{\epsilon_1}^-, \tilde{\epsilon_3}^+; \sigma_{IF})$. Reduce the geometric dependency between the spring and trigger, making the radial fit constraint $g_1(\tilde{\epsilon_1}^-, \overline{\underline{d_{ps2}}}^+, \overline{\underline{l_{t2}}}^+)$, and the g_{11} less interdependent. Attempt this by changing the

working direction of the trigger interface, to add additional degrees of freedom, resulting in a new interface stress criterion (which is currently globally active).

4. *Scale Trade-off Variable*: Reduce the influence of $\overline{d_{ps2}}$ upon size and self-orientation by moving the spring closer to the centre of mass - e.g. using a tension spring.
5. *Shifted bounds*: Eliminate the contributors to h_8 that reduce $\overline{\overline{z_{acc}}}$.
6. *Eliminate Parasitic Contributions*: Reduce the volume/mass of the plastic components when possible. Explore alternative linear guides for the needle hub and sealing principles for the valve component.

5.5.2 Redesign Exploration

This redesign procedure led to an iterative exploration of redesigned configurations, all designed with the aim of implementing the outlined strategies. This resulted in a succession of 11 redesign iterations, illustrated in Figs. 5.12 -5.15, with a sketch of the original design included as iteration 0 for reference.

Each of these iterations represents a gradual implementation of the aforementioned redesign sequence. As such, many of the iterations shift the same bounds and scale the same trade-off variables; they just do so to an increasing extent. Some of the iterations lead to new trade-offs or violate existing constraints, meaning alternative changes were explored in the subsequent iterations. As these redesign iterations illustrate, some of the strategies can only be applied to a certain limit. One cannot necessarily scale a trade-off variable or shift the bounds of harmonious variables infinitely, as other (or new) constraints may become active in the process.

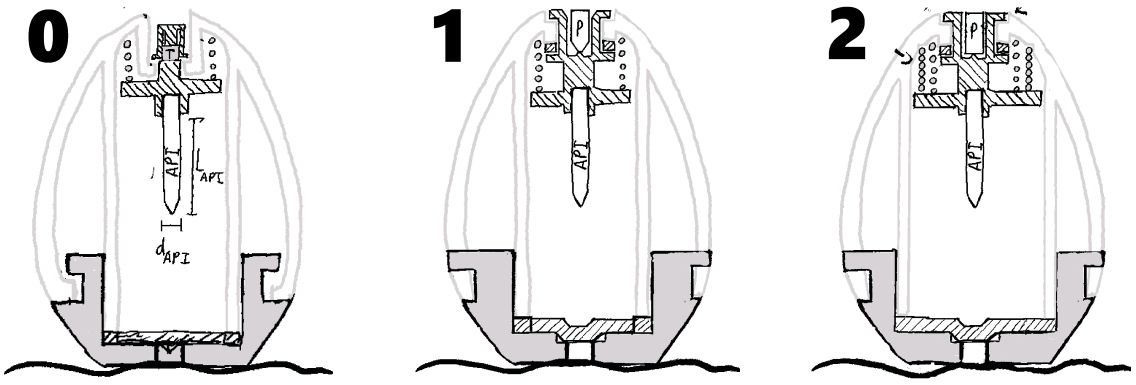


Figure 5.12: The redesign iterations that resulted from the exploration of the redesign sequence. The initial SOMA device is included for reference as iteration 0

In iteration 1, the snap between the housings is redesigned, to eliminate parametric contributions to the radial fit constraint (g_1) which ultimately determine the size of the device ($\underline{\tilde{\epsilon_1}}$). In turn, this allows the vertical position of the mating surface between the top and base to be lowered for a given device size, i.e. increasing l_{t2}, which lowers the centre of mass. This is achieved by changing the snap design of the top housing from an undercut to a cutout and changing the base snap feature correspondingly. Furthermore, the trigger arms have been inverted to work in tension. This allows the plug to move upward, meaning the spring coil ($\underline{d_{ps1}}$) can in part be shaped independently of the plug (d_p), given that the trigger arms are $\overline{\overline{\text{flexible}}}$. Hence, $\underline{d_{ps1}}$ is reduced. This also eliminates a mould tool constraint that limits the achievable width $\overline{\overline{\text{of the}}}$ trigger arms ($\overline{w_{nh}}$), eliminates a locally active buckling constraint and reduces the amount of material at the top of the device. In other words, the

spring is stiffened, the load-bearing area is increased, the centre of mass is improved, and the contributors to device size are reduced, yielding more room inside the device.

The utilization of this room was explored in **iteration 2**. Here, nested springs are introduced, which increases the achievable spring force, introducing new harmonious variables (the diameter- and number of spring coils). Unfortunately, this also introduces a new trade-off variable; the spring wire diameter of the new spring. This ultimately increases the total spring mass, worsening the trade-off between velocity and self-orientation.

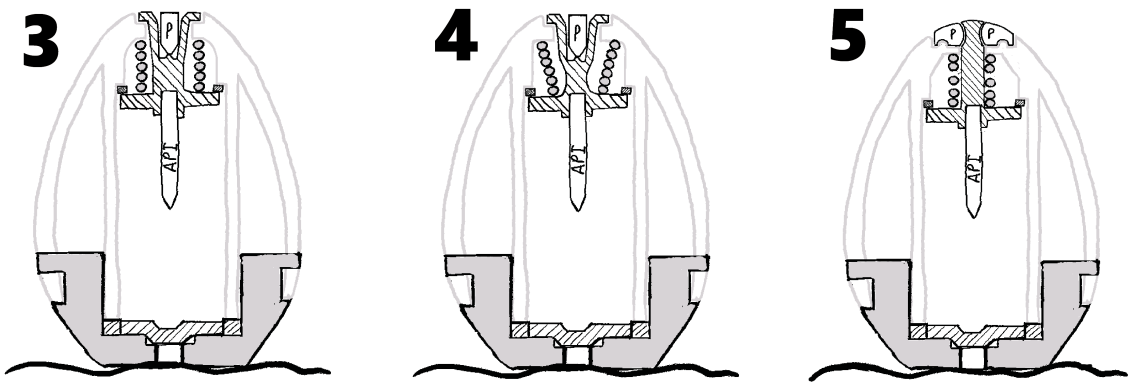


Figure 5.13: *cont.* The redesign iterations that resulted from the exploration of the redesign sequence. These iterations focus on shifting the bound of d_{ps1}, an impactful harmonious variable.

Hence, **iteration 3** involved removing additional contributors to the glb of $\underline{\underline{d_{ps1}}}$ instead, to the benefit of self-orientation, device size, and impact velocity, as more spring force can be achieved within a smaller volume. This is achieved by inverting the o-ring, which helps create a seal to protect the API from the harsh environment of the stomach. Instead of fitting the seal inside the spring, it has been moved outside it. Hence, the spring coiling diameter is now only defined by the trigger geometry on the needle hub component. Compared to the original design, we have have transformed the glb of $\underline{\underline{d_{ps1}}}$:

$$\underline{d_{ps1}} = d_p + 2\delta_{nh} + d_{ps2} + 6R_{cl} + 4R_{wt} \tag{5.1}$$

$$\Rightarrow \underline{d^*_{ps1}} = d_p + 2\delta_{nh} + 2R_{wt} + 2R_{cl} \tag{5.2}$$

As the trigger arms are flexible, one might allow for the removal of the radial clearances, $4R_{cl}$, at the cost of a slight frictional loss. In **iteration 4**, the reduction of the spring coil diameter is taken further. Given the height of the needle hub, the trigger arms do not need to pass through all of the windings during injection, meaning a conical shape is feasible. Correspondingly, a part of the spring can be narrower than the trigger arms, given that they can flex inward during assembly before the plug is mounted. The change to a conical spring also allows for more active windings pr. unit of spring length. Furthermore, this change potentially permits more load-bearing area in the trigger system, as δ_{nh} and w_{nh} can be increased without necessarily increasing the diameter of the entire spring coil.

Iteration 5 is an attempt to shift the glb of $\underline{\underline{d_{ps1}}}$ even further, by inverting the trigger system, replacing the trigger arms with an interface that resembles a ball-lock. However, the new shape of the dissolvable plug component was deemed too challenging to mould due to the characteristics of the material. Inserting the minimum feasible values of each variable in Eq. 5.2 yields $d_{ps1} = 2.7$mm, meaning the dimension is approaching other, previously inactive

inequality constraints, such as minimum spring index, minimum spring diameter for handling in assembly, etc. Thus, the benefit in shifting $\underline{\underline{d_{ps1}}}$ as far as iteration 5 was deemed to be negligible.

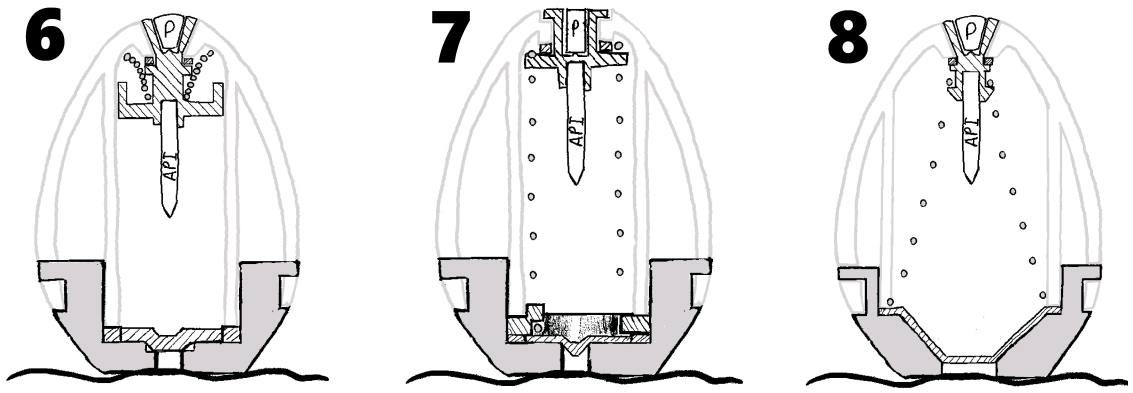


Figure 5.14: *cont.* The redesign iterations that resulted from the exploration of the redesign sequence.

Hence, the subsequent redesign iterations moved on to exploring redesign steps 4-6 in the aforementioned sequence. **Iteration 6** (see Fig. 5.14) increases load-bearing area in the trigger, by changing its working direction from being radial, to being a mix of radial and axial. This results in the trigger changing into a conical wedge-like interface. In turn, this adds a degree of freedom to the trigger interface design (its length in the vertical direction), meaning the spring can be stiffened without reducing the load-bearing interface. The radial fit constraint is now independent of δ_{nh} δ_{nh} which is critically constrained by the creep load constraint $g_{11}(d^{+}_{ps2}, \delta^{-}_{nh})$ at velocity Pareto frontiers. In effect, g_{11} was a Pareto constraint, as it depends on d_{ps2}. Unless another radial fit constraint actively bounds the outer diameter, d_{t1}, the spring force and the load-bearing surface in the trigger can now be increased, without increasing the size of the device beyond the contribution of $\overline{d_{ps2}}$. Up to this point in the redesign process, we have shifted the bounds of $\underline{\underline{d_{ps1}}}$ considerably, relaxed a Pareto constraint (g_{11}), and eliminated several parasitic contributors from g_1 which defines the smallest achievable device size, $\underline{\tilde{\epsilon_1}}$ and is involved in the activity cases analysed in Chapter 4.

Iterations 7-11 attempt to build upon these improvements by exploring the use of a tension spring, which poses issues w.r.t. achieving proper linear guidance of the needle, which ensure that it actually passes through the hole in the base, despite the influence of tolerances and dynamic effects. In the first attempt at this, **iteration 7**, the spring windings return to a block position upon injection, limiting the achievable stroke in the device, affecting the achievable velocity. **Iteration 8** solves this using a telescopic conical spring coil, where the windings of the spring pass through each other during the injection. Yet, this comes at the cost of linear guidance, which is solved in **iteration 9** by(see Fig. 5.15) utilizing a rectangular-wire telescopic tension spring, which is in part self-guiding, at the cost of a higher hysteresis.

Yet, this change to a telescopic conical tension spring introduces new trade-off variables that affect the trade-off between device size (c_1) and impact velocity (c_3). The new shape means that the amount of spring material, and thus the system's energy storage capacity, is defined by the springs minor and major coiling diameter. This makes both potential trade-off variables, as the spring interfaces with the needle hub and housings, at locations that ultimately affect the achievable amount of API and device size. To reduce the effect of this,

iteration 10 involves removing the guiding cylinder, allowing the wider end of the spring against a new load-bearing surface on the top housing. In turn, this increases the amount of steel in the base, lowering the centre of mass. Yet as in iteration 8, this comes at the cost of linear guidance, as the tip of the needle is able to title once the needle hub disengages from the trigger surface. To solve this, one could either rely on a rectangular wire spring at the cost of increased spring hysteresis and a more expensive spring or repurpose the valve to steer the needle tip. This is implemented in **iteration 11**, where the valve geometry has been changed to allow it to act as a radial steer which is relatively stiff in the radial direction while remaining soft in the axial one, as the spring will act against it upon injection.

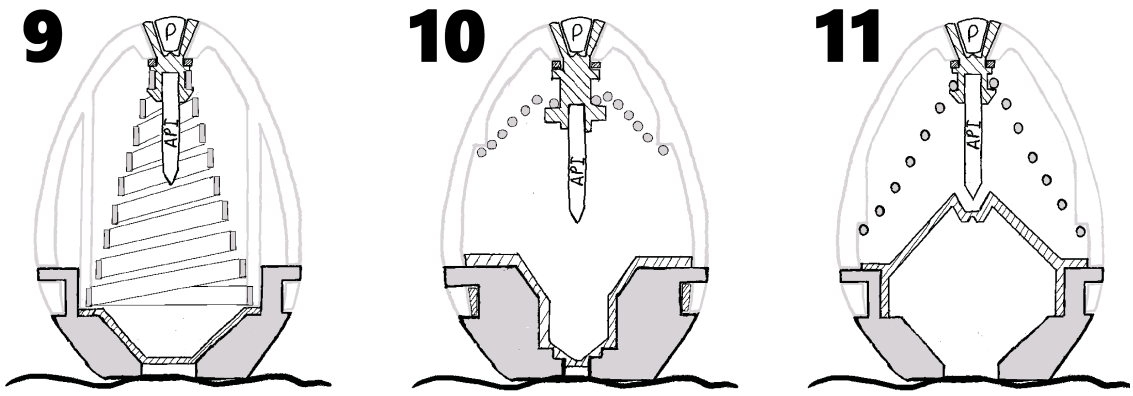


Figure 5.15: *cont.* The redesign iterations that resulted from the exploration of the redesign sequence.

Related to the results of the analysis performed in Chapter 4, the successful implementation of the tension spring has several benefits. The negative influence of the spring wire on self orientation (f_1) in the original design, is multiplied by $\underline{\underline{n_a}}$ and $\underline{\underline{d_p s1}}$ and its mounting height. Inverting the spring scales its negative influence and makes it geometrically independent of the trigger. This inversion also allows the elimination of the contribution of $\underline{\underline{n_a}} \cdot \overline{d_{ps2}}$ from h_8, the equality constraint which determines the achievable acceleration stroke, $\overline{z_{acc}}$, and relaxes $g_{20}(l_{b3}^+, d_{ps2}^+, d_{n1}^+, l_{n2}^+, l_{t2}^-, l_{b1}^-, \tilde{\epsilon_1}^-, \tilde{\epsilon_2}^+)$ which is a regionally active Pareto constraint. As z_{acc} only affects the impact velocity objective (c_3), its optimum might be improved without affecting the other objectives, thereby shifting the Pareto frontier.

In summary, iterations 1-5 shift the bound of d_{ps1} while eliminating parasitic influences stemming from the design of the top-base housing interface. Iteration 6 allowed a substantial increase in the size of the load-bearing surface, shifting the globally active creep constraint g_{11}, and reducing its multiplying effect on trade-off variable d_{ps2}. Iterations 7-11 explore the implementation of a tension-based trigger, with the telescopic spring design eliminating n_a from h_8, greatly improving $\overline{z_{acc}}$, while eliminating the parasitic influence of the mass of the guiding cylinder upon the mass distribution. Iterations 2, 5, 7, and 10 all came with feasibility related issues or worsened trade-offs, which affected the design changes made in the subsequent iteration. While the previous iterations primarily leveraged harmonious variables, iteration 11 successfully aligns a trade-off variable by shifting the spring mass downward. Yet, this comes at the cost of a new dependency between the amount of spring material and the device diameter. Hence we cannot say for sure whether the tension spring redesigns actually improve upon the preceding redesign iterations (e.g. iteration 4 or 6) without further analysis.

Interestingly, there were no drastic changes, such as the introduction of new components or a change in working principle. Rather, simple changes are introduced, which will have a substantial impact on the optimization model and hence probably affect the Pareto set. This analysis-guided redesign procedure potentially results in improved performance without necessarily compromising feasibility or relying on parametric change, which may come at a cost in many cases (e.g. higher quality materials or different production processes). Most of the design changes involved are essentially analogous to well known redesign heuristics, namely, *inversion* as already prescribed in numerous sources [6, 9, 14, 16] *change in working direction/load path* [6, 14, 16], and *contributor reduction* [6, 13]. As such, identifying these redesigns was relatively straightforward, given that the Pareto set dependency analysis had already allowed the identification of what needed to be changed in order to improve the configuration. The influence of these changes will be evaluated in the next chapter.

6 Trade-off Management in Early Design

In the previous chapters, systematic approaches to identifying and mitigating the root causes of trade-offs have been developed and applied to the SOMA case. This chapter expands these developments into applications beyond analysis and improvement of a single embodied system. The chapter begins with the development of a set of design guidelines that arise as a consequence of the theorems and corollaries developed in the previous chapter. These guide the synthesis of the first embodiment towards achieving the best possible optima. This is followed by perspectives on how the avoidance, analysis, and mitigation of trade-offs can be used to steer the iterative design process towards achieving a "good" final embodiment in section 6.2. These developments include content from Paper B but also include further and broader developments. The chapter is concluded with a section that is in part adapted from Paper B, describing novel developments for the evaluation and selection of redesigns using, which are exemplified using the SOMA case.

6.1 Synthesis of the Ideal Design

Until now, this thesis has mostly focused on analysing and improving an existing design. Yet, there will inevitably be a degree of path dependency involved in doing so. What if the limitations of a system are inherent to the overall concept? Or arise due to decisions made in the synthesis of the first embodiment? Suppose the dependencies and constraints that cause trade-offs or limit optimality are inherent to the working principles used in the system (e.g. the use of a spring-driven actuator in the SOMA device). In that case, the methods discussed until now do not necessarily suffice. The challenge in this largely comes down to the intrinsic difficulty in describing a concept through analysis without an embodied design:

> *In the formation of principle solutions (for example working structures), data about the physical relationships may be insufficient, since the geometrical relationships may have a limiting effect and hence may, in certain circumstances, lead to incompatibilities. In that case, physical equations and geometrical structure must first be matched mathematically, and this is not generally possible except for systems of low complexity.*
>
> - Gerhard Pahl & Wolfgang Beitz
> Engineering Design - A Systematic Approach [6]

As the above quote illustrates, the limitations and challenges involved in the design of a system largely become clear at the embodiment level, in the geometric realisation of components/structures. Hence, to a large extent, the synthesis of the initial embodiment is interlinked with conceptual design, just as we can only truly describe the limitations of a concept quantitatively by studying an embodied design. Hence, it is relevant to question whether the design principles and analysis methodology from the previous chapters have implications in embodiment synthesis and conceptual design?

Recall from section 4.1.3. that trade-offs can both be caused by dependency between design objectives through shared design variables or through active constraints. We cannot prescribe an overall synthesis methodology, as trade-offs and their root causes are contextual to the objectives and constraints at hand and the manner in which the system has been embodied.

Yet, we might consider what steps one can take to avoid some of the issues this analysis and redesign methodology concerns itself with. Given that it is difficult (approaching the impossible) to build optimization models that identify the optimal concept of configuration, we can instead consider the following. What decisions can we make in synthesis that reduce the likelihood of trade-off and yield *good* proportional optima in the end product?

6.1.1 Conjectures and Conditions of Ideal Design

In order to define such decisions, we must first define what to strive for. Several heuristic frameworks exist, which prescribe different notions of what *good design* constitutes. Suh [13] argues independence between functional requirements is the key to *good design* while Pahl and Beitz [6] argue for *clarity, simplicity,* and *safety*. These, however, are not necessarily characteristics we can describe in a rigorous manner.

Viewed from an optimization perspective, *good* design is perhaps not as ambiguous. Design is almost always multiobjective [30]. Some objectives are well known, explicitly stated by the designer from an early stage, and are measurable/quantifiable. Others, meanwhile, can be tacit, immeasurable, or even subjective in nature. This is compounded by the fact that the relative importance of objectives is not necessarily easy to ascertain. Correspondingly, designers may not be aware of certain biases or forms of fixation that affect their decision-making.

Yet, none of this detracts from the fact that trade-offs exist and that some end products perform better than others, whether we are able to model all the objective functions or not. Ultimately, it is well established that designers design with the optimum in mind. Opportunism [35, 97], trade-off knowledge [33, 124], and of an apriori understanding of which constraints are active [59, 132], have all been found to be key indications of a designers experience. In turn, the involvement of *expert* designers increases the likelihood of success in the early stages of development [35].

Based on this, we could argue that successful designers strive to synthesise the solution that yields the end-product with the *best* performance - i.e. the best optimum. Yet, the definition of optimality is the *best* solution within available means [12] - i.e. what is feasible. How do we then design a system in a way that affects the available means, yielding the *best proportional optimum* - i.e. the "ideal" solution?

Generally speaking, the design of most products involves multiple objectives, meaning the optimum will be a set rather than a point unless we know the relative weighting of design objectives or there are no trade-offs at play. Furthermore, the end product comes with an associated cost (be it economical or environmental) which is to an extent driven by design complexity. Thus, for the purposes of the ensuing developments, we posit that designers will often strive to synthesise products that:

1. Involve no trade-offs
2. Have the *best* possible performance
3. Are low cost

The design of any product can be described as an optimization problem, consisting of any and all design objectives, constraints, variables and parameters of relevance to the development of the end product. These objectives and constraints range from those that can be modelled (e.g. mechanical efficiency) to those that cannot (e.g. user-friendliness). Viewed from this perspective, the above characteristics can be translated into optimization terms. Designers *ideally* want to synthesise designs with an optimum is a point rather than a set, positioned

as close to *origin* as possible, defined by a few design variables as possible (as this at least correlates to structural complexity [13]).

To put this into more formal terms and in negative-null form, we introduce three conjectures describing what designers strive for in synthesis. These describe the hypothetical *ideal design* and lead to a set of conditions that must be fulfilled in order for a design to be ideal. The conjectures and conditions are put forward in a hierarchy, in that the second conjecture is only put forward under the presumption that the first is true and its condition upheld, while the third is put forward presuming that the two preceding conjectures are true:

Conjecture 1 The First Conjecture of Ideal Design Synthesis
In the ideal design, $\text{argmin} f_i(\mathbf{x}) = \text{argmin} f_j(\mathbf{x})$ *for any pair of design objectives,* i *and* j*,* $i \neq j$*, meaning no trade-offs exist.*

Recall from theorem 5 that the Pareto set cannot exist if there are no trade-off variables in the optimization problem (globally or regionally) after the back-substitution of all active constraints. Thus, we can state the following condition for the First Conjecture:

Condition 1 Avoidance of trade-off variables
For a design to be ideal, it cannot contain trade-off variables, meaning $x_i \notin \underline{\overline{\mathbf{x}}}$*, for any variable i*

Recall that by definition, trade-off variables have oppositely monotonic relationships with two or more objectives, either globally (meaning the variable is monotonic) or regionally (meaning the variable is non-monotonic). From condition 1 it thus follows that the ideal design only involves objectives that are either only dependent on monotonic variables, on non-monotonic variables that are not shared with other objectives, or on non-monotonic variables that by chance have the same value at the minimum of each objective. If no trade-offs exist, we can move on to the question of the optimum of each objective:

Conjecture 2 The Second Conjecture of Ideal Design Synthesis
In the ideal design $f_i^* \to 0 \wedge -\infty$ *for any design objective* i*.*

Were it not for the First Conjecture and Condition 1, this conjecture would, on its own, simply imply that the ideal Pareto set is infinite. Recall from theorem 6 and its corresponding proof that harmonious variables and their bounds in part determine the position of the Pareto set. If condition 1 is fulfilled, the location of the optimum is only determined by constraints and the existence of interior optima (which implies non-monotonicity). Hence, we can state the following condition for the Second Conjecture:

Condition 2 Boundedness in the *Improving Direction*
For a design to be ideal, the bounds of its design variables must be infinite or asymptotic in the improving direction, meaning $\overline{x_i} \to \infty$ *for* $\mathbf{f}(x_i^-)$ *and* $\overline{x_j} \to 0 \wedge -\infty$ *for* $\mathbf{f}(x_j^+)$ *for any* i *and* j*.*

As a consequence of the activity theorem of constrained optimization [12], the optimum would never reach 0 or $-\infty$ if this condition is not fulfilled. Assuming both conditions are fulfilled, we can state the third and final conjecture:

Conjecture 3 The Third Conjecture of Ideal Design Synthesis
The ideal design has as few design variables as possible, meaning $\dim \mathbf{x} \to 1$*.*

Given the stated prerequisite that all design problems are multiobjective, it follows that $\dim \mathbf{f} \geq 2$. From this, a condition arises, without which the Third Conjecture would result in trade-offs:

Condition 3 Harmonious Variables
For a design to be ideal, all design variables must be harmonious, meaning $\operatorname{argmin}\mathbf{f}(\mathbf{x}) = \overline{\overline{\mathbf{x}}} \wedge \underline{\underline{\mathbf{x}}}$.

With these conjectures and conditions, we have a basic definition of what "good" design constitutes and what this implies about the dependencies and bounds in the system being designed. This definition is consistent with the developments made in the previous chapters. However, it goes beyond the Design Improvement Criterion from Chapt. 5, as it does not involve a comparison with preexisting design, meaning it applies to synthesis.

6.1.2 Converging Towards the Ideal

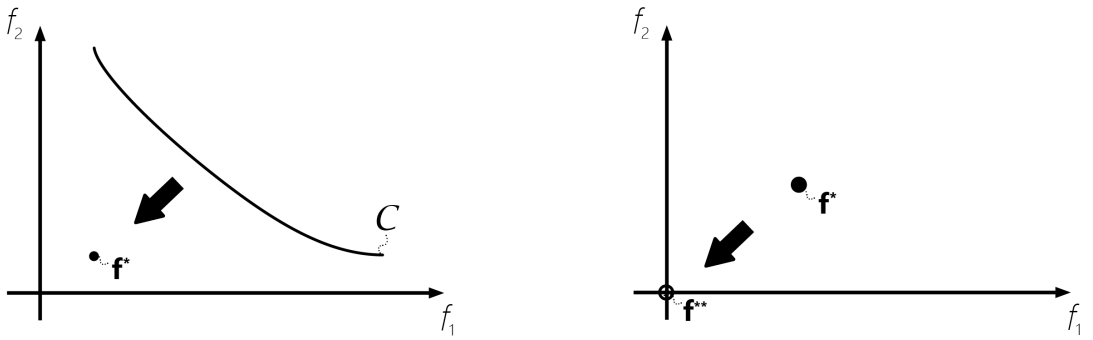


Figure 6.1: Converging toward the Ideal Design involves the targeted avoidance of trade-off variables, the widening of the feasible domain of harmonious and independent variables, and the reduction of variables/parts in the system.

The developments in the previous subsection might seem overly formal, hypothetical, and entirely academic. Obviously, no functional intent can be realised with a single design variable, just as trade-off variables can never be completely avoided. Correspondingly, a design problem with an infinite or asymptotically bounded feasible domain is poorly bounded, meaning any optimization model built to identify the optimal proportions of the system would fail to converge.

Yet, the Conjectures and Conditions of Ideal Design are put forward in support of design synthesis. We can, in practice, never actually fulfil any of the conditions described; we can only converge towards them. If we accept the conjectures as fact, then a design is *closer* to fulfilling the conditions than another, which will, by definition, be better. Viewed from this convergence perspective, the Conditions of Ideal Design have certain implications for synthesis:

***Condition 1*:** Implies that dependencies are not a problem, so long as they do not introduce trade-offs between any of the design objectives of relevance in the given design context. Hence, the more trade-off variables we are able to avoid through targeted decisions in synthesis and configuration, the better the design.

Condition 2: Implies that the bounds of monotonic harmonious and independent variables can have a substantial impact on the location of the optimum and thus how close the concept or configuration design is to being ideal. Hence, we can leverage monotonicity information to arrange the parts and features in a system in a way that widens the feasible domain of our design in the improving direction.

Condition 3: Implies that harmonious variables ultimately allow the realisation of more product functionality with less complexity. Hence, if we're able to avoid trade-offs and overly restrictive constraints without introducing new design variables/parts or remove

variables/parts without introducing trade-offs or introducing new constraints, the design is closer to the ideal.

Herein lies one of the core contributions of this thesis. Based on the theorems and proofs in the previous chapters and the conditions developed here, it would seem that good mechanical design synthesis is a matter of avoiding trade-off variables between design objectives and avoiding that the active constraints introducing additional trade-off variables while achieving as many harmonious variables as possible with as nonrestrictive constraints as possible. Hence, to converge towards the ideal design, we need to take measures in synthesis to ensure as wide a feasible domain as possible, few trade-off variables, and a low complexity. When would we apply such measures?

In Pahl & Beitz' Product Development process [6], the conceptual design phase begins with the identification of key product requirements, the identification of the essential problems that the end product is aimed at mitigating, and the identification of the desired functions and sub-functions of the end product. From an optimization perspective, the high-level design objectives and constraints for the end product arise as a consequence of the decisions made in this process - be they related to the physics of the product, the needs of the organisation, the needs of the user, etc.

Yet, the objective functions themselves, constraint functions, and the dependencies between them are determined by what comes after [133]. In the identification, selection, and development of working principles, the overall system structure, and the embodied design, the physics and practical limitations of the end product are determined. Based on experience or an understanding of the underlying physics, the designer might very well be aware of some of the monotonic relationships between the design variables and the objectives and constraints, even without any formal analysis.

In Section 6.1.3, design rules and guidelines are presented and demonstrated using the Novo Nordisk FlexTouch Device and the SOMA device. These guidelines support the convergence towards the ideal design and have emerged out of the Conditions of Ideal Design and out of the developments made in the previous chapters through a hypothetical-deductive approach. In unison, they answer two basic questions which arose in this process. Firstly, what decisions can we make in synthesis to get closer to the ideal? Secondly and of equal importance; what decisions can we make in synthesis to avoid some of the issues we might identify through Pareto-set Dependency Analysis and the Configuration Redesign Principles?

These rules and guidelines mostly apply irrespective of context (e.g. specific design objectives) and are aimed at aiding the selection of working principles and the combination of these into the first synthesised embodiment (the preliminary embodiment [6]). Later in the design process, they might also support loop-backs into the embodiment or conceptual design stage. Nonetheless, as the guidelines rely on the designer having some prior knowledge, there are some steps one must undertake in the early design process before they can be applied:

1. Define the overall design objectives based on the desired functionality.
2. Map out the potential working principles and structures.
3. Consider which monotonic relationships exist between the design objectives, the design variables the working principles/structures give rise to, and the different constraints that they may give rise to.
4. Map out the potential trade-offs and active constraints these relationships might give rise to.

5. Apply the Design Guidelines to support the refinement and selection of the principal solution or the synthesis of the preliminary embodiment.

6.1.3 Guidelines for Ideal Design Synthesis

This section puts forward a set of design guidelines under the notion that most designers are actually aware of (or able to identify) some of the basic monotonic relationships involved in their design problems from very early on in the design process. Even at a point where the first system sketch has not been developed, the designer may still be aware of some inherent relationships between the design objectives and the design variables that will arise as a consequence of the selection of working principles and their combination into a principle solution and subsequent preliminary embodiment.

Hence, many of these guidelines are put forward at a very basic level of abstraction; *design variables* here do not necessarily reflect the dimensions on a drawing or the variables in an optimization model. Rather, they might also represent the designer's overall understanding of how the working principles and the configuration of parts influence the design objectives - e.g. "*the larger the output from working principle A, the less mechanical efficiency we can achieve using working principle B*".

The design guidelines are put forward in three categories, based on which of the Conditions of Ideal Design they relate to. Given the hierarchy between the conjectures and conditions, there can be certain interdependencies between the guidelines. To distinguish between them, the design guidelines are listed with a prefix, $G_{\overline{\chi}}$ for guidelines related to Condition 1, G_{χ} for Condition 2, and $G_{\dim(\mathbf{x})}$ for Condition 3. Under each category, the guidelines are grouped depending on what type of design activity or mode of design reasoning they relate to. Some of the guidelines are reductive in nature and are related to existing heuristics, while others are more inductive in nature and rely on the designers understanding of the dependencies and monotonic relationships that are involved in the overall design problem. This is not an exhaustive set of design guidelines but rather a collection of heuristics that can be applied in synthesis, early decision making, and refinement of the first embodiment to reach a system that is closer to the ideal.

Case Introduction: The FlexTouch Device

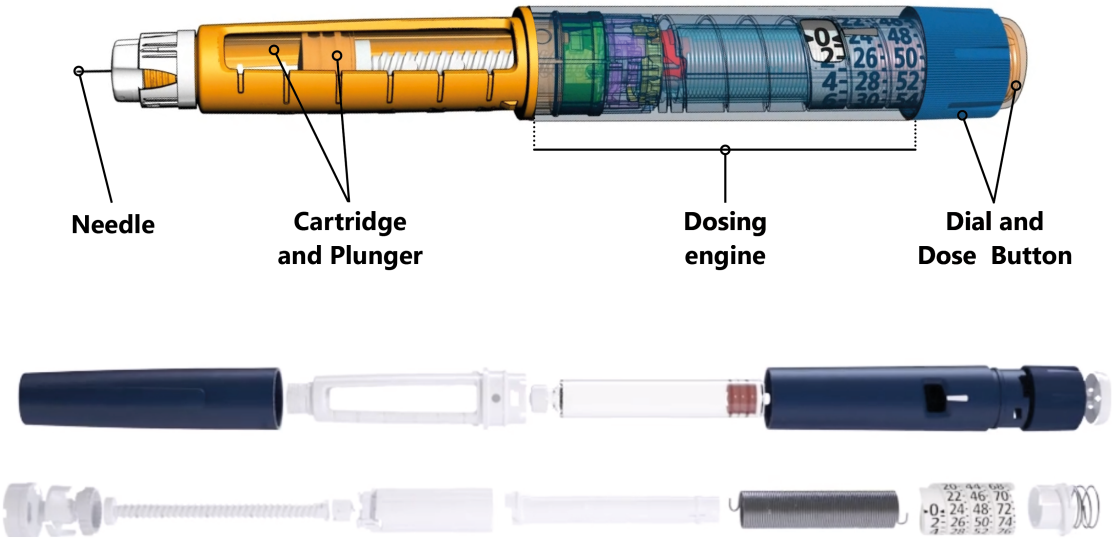


Figure 6.2: The FlexTouch pen is used to exemplify some of the guidelines for ideal design synthesis

In the following, a new case - the FlexTouch injection pen (see Fig. 6.2) - is used to exemplify some of the design guidelines, along with the SOMA case and more general examples from literature. The FlexTouch is a platform product designed and manufactured by the case company for the subcutaneous injection of active pharmaceutical (API) ingredients such as insulin, human growth hormone, and *glucagon like peptides* (GLP1). Such APIs are essentially large proteins, meaning they cannot be delivered orally, as the gastrointestinal system breaks down proteins by its very nature, resulting in little or no systemic uptake. As such, the FlexTouch is used by people living with diabetes, human growth hormone deficiency, and obesity, to perform daily or weekly injections. The device is pre-filled, meaning it contains a set amount of API and is disposed of once empty. It is used several times, delivering the dose the user selects via a sterile needle which is replaced by the user before each dose.

In essence, the device is designed to *autodose*, meaning that the user selects a dose prior to inserting the needle, after which the device delivers the dose at the push of a button. In selecting the desired dose using the dial (see Fig. 6.2), the user winds up a torque spring inside the device, with the set dose being shown on the scale. Said torque spring drives the dosing mechanism, rotating a lead screw through a stationery nut, thereby pushing a plunger through a cartridge filled with an API (e.g. insulin) in liquid form. By inserting the needle and activating the dosing mechanism using the button at the end of the device, the user injects the API into their *subcutis* (a tissue layer under the surface of the skin). Upon pressing the button, the torque spring mechanism is released from a rotational lock, thereby turning the lead screw. Several ratchet mechanisms create click sounds and haptic feedback to assist dose setting, indicate dose progress and indicate when dose delivery is complete. As such, numerous simultaneous functionalities occur in this process. The accuracy of dose delivery is essential, as even the slightest under- or overdoses can have significant long-term health effects. Given this, the substantial production volume, and the safety-critical nature of the product, the FlexTouch device is embodied with as few components as possible. This ensures a low cost, as few tolerance contributions as possible (which affect dose accuracy), and high reliability.

However, this also means that each individual component contributes to numerous subfunctions, resulting in a highly interdependent design, which involved numerous challenging trade-offs in development. In fact, the device took in excess of 6 years to develop, from initial sketch to running production, largely driven by this interdependent nature. That said, the design of the FlexTouch actually reflects that many targeted decisions have been made in the development process to avoid trade-offs, widen the feasible domain, and allow a low part count. These decisions were made in the selection of working principles and the configuration of the system and were the result of solution exploration, design iteration, and experience. The device was not used in the development of the guidelines but was rather found a posteriori to be consistent with them.

Condition 1 - Guidelines for Trade-off Avoidance

The avoidance of trade-offs through independence between design objectives is a common recommendation in engineering design literature (e.g. Suh [13], Pahl & Beitz [6], and Skagoon [96]). Yet, as shown and discussed in chapter 5, there are other routes towards avoiding trade-offs or reducing their influence, which might be equal or even preferable to independence in many contexts. If we expand the redesign principles derived in Chapter 5 - *Separate*, *Flip monotonicity*, and *Scale* into the decisions such as the selection of working principles, the synthesis of the working structure (aka. the *layout* of parts and subsystems) and the resulting preliminary embodiment, a set of inductive and reductive guidelines emerge. These apply to any trade-off variable and any set of objectives, whether they exhibit globally monotonic behaviour or not. Some might seem entirely obvious, but this merely reflects the importance

of considering potential contributors to trade-offs upfront.

$G_{\underline{\bar{x}}1}$: Select working principles with trade-off variables in mind

Some of the first decisions made in the design process can yield the most challenging trade-offs. With each of the different working principles that can be used to perform a desired function, different dependencies and thus corresponding trade-offs follow. It follows that certain working principles are more suited to allowing simultaneous optimization of certain sets of design objectives. Considering this while selecting working principles might lead to avoidance of detrimental dependencies or aid in the invention or identification of new working principles:

$G_{\underline{\bar{x}}1.1}$ Select working principles based on the objectives at hand.

$G_{\underline{\bar{x}}1.2}$ When possible, avoid working principles that do not allow the simultaneous improvement of all the objectives they contribute to or affect.

$G_{\underline{\bar{x}}1.3}$ Introduce compensating functionality that allow the reduction of trade-off between design objectives, when independence or like-monotonicity is not possible. This is especially applicable for intrinsically conflicting objectives.

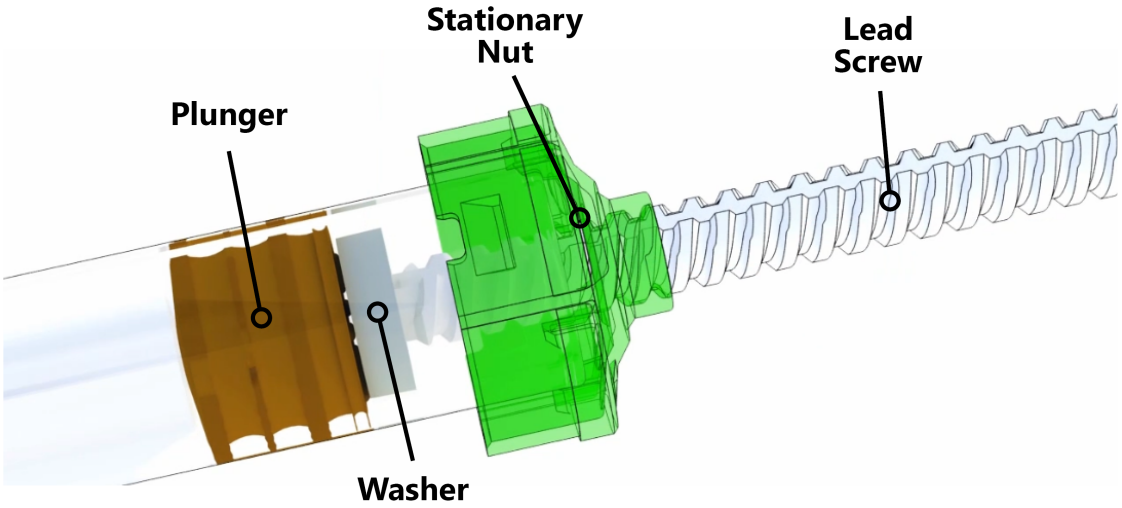


Figure 6.3: Example 1: The selection of a lead Screw rather than a rack and pinion in the FlexTouch device, allows long friction without elongating or widening the device.

Example 1: *Selecting Working Principles based on the Objectives - the FlexTouch Lead Screw*
Relevant Design Objectives: Minimise device diameter and length, maximise dosing speed (reduced by friction), maximise dose accuracy.
Design Guidelines Involved: $G_{\underline{\bar{x}}1.1}$, $G_{\underline{\bar{x}}1.2}$, and $G_{\underline{\chi}2.1}$

The decisions made in the development of the FlexTouch w.r.t. the selection of working principles and configuration of parts clearly show that the designers had these objectives in mind in synthesis. Primary amongst these decisions is that the entire dosing engine works in rotation, which is converted into a linear movement.

One aspect in that regard is the specific use of a lead screw driven by a torque spring, which is particularly beneficial in regards to avoiding several trade-off variables between dosing speed and dosing accuracy on one side and device size (diameter and length) on the other. The benefit of the lead screw lies in that the frictional loss is reduced with its diameter.

Meanwhile, the net force exerted by the lead screw is determined by its pitch, diameter, and torque delivered by the engine. Alternatives to a lead screw could be a helical rack and pinion or a linear ratchet rod (which we will get back to in Example 7). As discussed in Chapter 5, the accuracy of a rack and pinion increases monotonically while the friction decreases with an increase in the diameter of the pinion and the length of the rack's sliding joint. In other words, its efficiency and accuracy grow with its size, which the lead screw avoids.

Having selected a lead screw as a basic principle and placed it in the centre of the device (allowing the smallest screw diameter), the designers of the FlexTouch have avoided contributions to several size-related trade-offs. However, as we reduce the diameter of the screw, its accuracy becomes more sensitive to geometric variation in the thread and to errors in its rotational position (e.g. due to play between components), affecting dosing accuracy. This aspect will be revisited in Example 3.

$G_{\underline{\overline{x}}2}$: Synthesise preliminary embodiments with trade-off variables mind

In the combination of working principles into an overall principle solution, dependencies arise due to parts/variables contributing to several functionalities (and therefore objectives) simultaneously. Correspondingly, the realisation of the preliminary embodiment creates to further trade-off variables, as the embodiment gives rise to design constraints that may introduce trade-off variables when active. Even the slightest monotonicity information and identification of potential dependencies may guide this process towards avoiding trade-off variables:

$G_{\underline{\overline{x}}2.1}$ Assess monotonicity during morphology exploration: Systematically explore alternative combinations of the required working principles into different system layouts. Compare and select based on avoiding as many potential trade-off variables as possible.

$G_{\underline{\overline{x}}2.2}$ If a potential trade-off variable becomes evident, redistribute functionality among the parts/ subsystems/ functional elements in the system, or rearrange the parts themselves, to achieve independence or a scaling of the trade-off variable.

$G_{\underline{\overline{x}}2.3}$ Avoid geometric dependencies between oppositely monotonic variables. E.g. positioning a geometric feature with a monotonically decreasing influence on one objective inside a feature with a monotonically increasing influence on another objective.

$G_{\underline{\overline{x}}3}$: Avoid common drivers of trade-offs

Certain design "mistakes" can be made in the combination of working principles into a preliminary embodiment that drive potentially avoidable trade-offs. Many context-specific examples of such can be found in existing heuristics in engineering design literature; see, e.g. French [14] and Skagoon [96]. If not directly intended or necessary to realise the desired functionality, these drivers should be avoided.

$G_{\underline{\overline{x}}3.1}$ Avoid designing towards objectives being interdependent through equilibria, c.f. the steam engine example from Chapt. 1.

$G_{\underline{\overline{x}}3.2}$ Avoid temporal conflicts - e.g. a part ideally being infinitely stiff in for optimal performance in one system state, and infinitely soft in another [16]. Redistribute functionality, or introduce new parts to mitigate such scenarios.

$G_{\underline{\chi}3.3}$ Avoid force loops that overlap unnecessarily, especially if these work in the opposite direction [14].

$G_{\underline{\chi}3.4}$ Avoid unbalanced and asymmetric loads, unless they are required to fulfill a given functional intent [6, 14, 29]

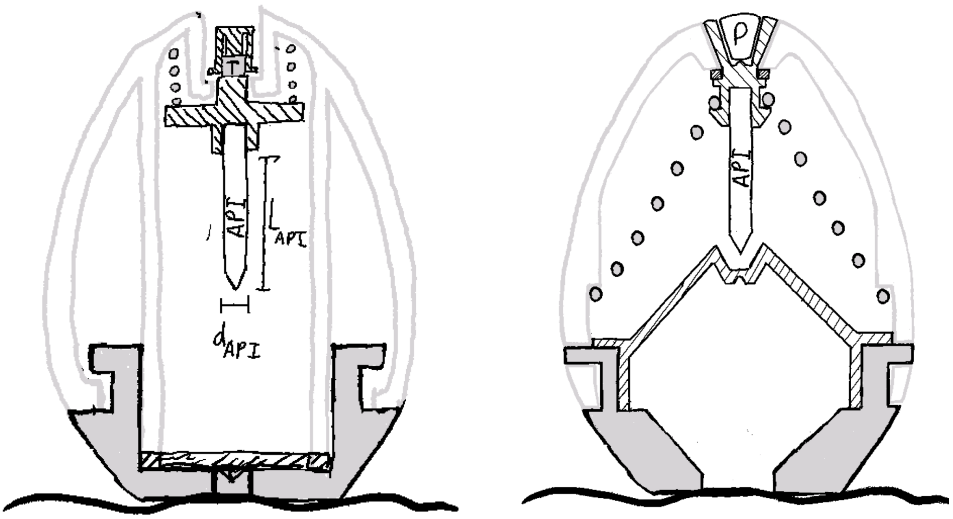


Figure 6.4: In the SOMA device, moving the spring closer to the base to reduce the self-orientation vs impact velocity trade-off, and avoiding the constraint-driven trade-off related to the serial arrangement of the spring and needle, can only be achieved through a change in working principle.

Example 2: Redefining the working principle and structure in the SOMA
Relevant Design Objectives: Minimise device diameter, while maximising impact velocity, self orientation and API capacity
Relevant Inequality Constraints: Radial fits between parts, vertical tolerance chains, needle tip clearance relative to the valve.
Design Guidelines Involved: Primarily $G_{\underline{\chi}1.2}$, $G_{\underline{\chi}2.2}$, $G_{\underline{\chi}2.3}$, but also $G_{\chi 2.3}$, and $G_{\chi 2.4}$ to an extent.

Revisiting the 11^{th} SOMA redesign iteration while considering the design guidelines, we see that we might actually have identified the redesign without the extensive analysis and redesign effort. Viewed from a working principle perspective, an actuator driven by a compression spring will always have to be mounted as far from the bottom of the device as possible, as the needle exits through the base, and the spring needs to interface with the needle hub. This inherently worsens the trade-off between self-orientation and impact velocity.

Conceptually speaking, a tension spring does not have this problem. We might have realised this without analysis had we considered the oppositely monotonic influence of the spring wire diameter upon the two objectives and how it is worsened by the serial arrangement of components necessitated by the use of a compression spring. This serial arrangement also creates a minor trade-off between impact velocity and API capacity, as any elongation of the spring (which follows with an increase in wire thickness) results in a shorter needle or less acceleration stroke. Interestingly, the change to a tension spring also allows functionality to be redistributed amongst the components, eliminating several design variables along the way, as can be seen by the elimination of the guiding cylinder and most of the trigger geometry. In unison, the valve and tension spring now fulfil the linear guidance functionality, which ensures that the needle actually goes through the hole in the base. In turn, this eliminates

material that worsens self-orientation and increases device size. So, in this case, the change in working principle allowed increased integration and the scaling of several trade-offs.

$G_{\underline{\underline{x}}4}$: Be Pragmatic

Trade-offs are not always worthwhile avoiding - some will occur due to inherently conflicting objectives (e.g. low mass vs high stiffness), while others exist between objectives on vastly different orders of importance. This should be considered in synthesis and redesign, as accepting these situations might open new opportunities:

$G_{\underline{\underline{x}}4.1}$ Accept the existence of a trade-off variable if the relative importance of the two objectives is vastly different or if the loss in utility caused by the existence of the dependency is negligible.

$G_{\underline{\underline{x}}4.2}$ "*The needs of the many...*": The existence of a trade-off variable might be acceptable if most of the objectives involved are of like monotonicity w.r.t the variable.

$G_{\underline{\underline{x}}4.3}$ If a constraint seems to be difficult to fulfil, treat it as an objective, and explore how the trade-off variables between it and the existing objectives can be avoided.

Example 3: *Scale and Needs of the Many in the FlexTouch Dosing Ratchet*
Relevant Design Objectives: Maximise dosing accuracy, dosing click volume and dosing speed, minimize device cost, diameter, and length
Relevant Inequality Constraints: Avoid lead screw buckling, feature sizes above a certain limit, stress in ratchet interface, radial play between components, injection moulding constraints.
Design Guidelines Involved: $G_{\chi 4.2}$ and $G_{\chi 1.3}$

The diameter of the lead screw in the FlexTouch is as small as it can feasibly be, given the constraint that it cannot buckle while dosing and size limitations related to straightness tolerances and injection pressure during moulding. Yet, the potential loss of accuracy that would have arisen from using as slender a lead screw as possible has been mitigated through conceptual/embodiment design - illustrated in Fig. 6.5.

Upon the activation of the dosing mechanism, leading to the torque being transferred from the spring mechanism to the lead screw, the screw itself does not interface directly with the spring mechanism. Rather, an intermediary - the purple part in Fig. 6.5 - controls its rotational position through a key interface and interfaces with the spring mechanism when dose delivery is activated. The part has a pair of elastic arms that are engaged with a ratchet surface in the outer housing, creating a one way-lock, meaning the lead screw can only rotate in the "dosing" direction. When the part rotates, the elastic arms create a click-sound thanks to the ratchet interface, indicating to the user that dose delivery is in progress, producing 24 clicks pr. revolution (equal to the resolution of the ratchet interface).

The inclusion of this intermediate part has a substantial positive influence on the dosing accuracy of the device. Thanks to the elastic arms, the interface between the lead screw and the key feature is virtually free of tangential play. Furthermore, the resolution of the ratchet substantially reduces the influence of errors in the rotational position of the dosing mechanism. It also scales down the influence of geometric variation in the dosing ratchet due to a gearing effect stemming from the difference in diameter between the key and ratchet.

So, beyond contributing to a new sub-function (the "dose progress" clicks), this intermediate component all but eliminates the trade-off involved in reducing the diameter of the lead screw to the bare minimum and several contributors to inaccuracy stemming from manufacturing

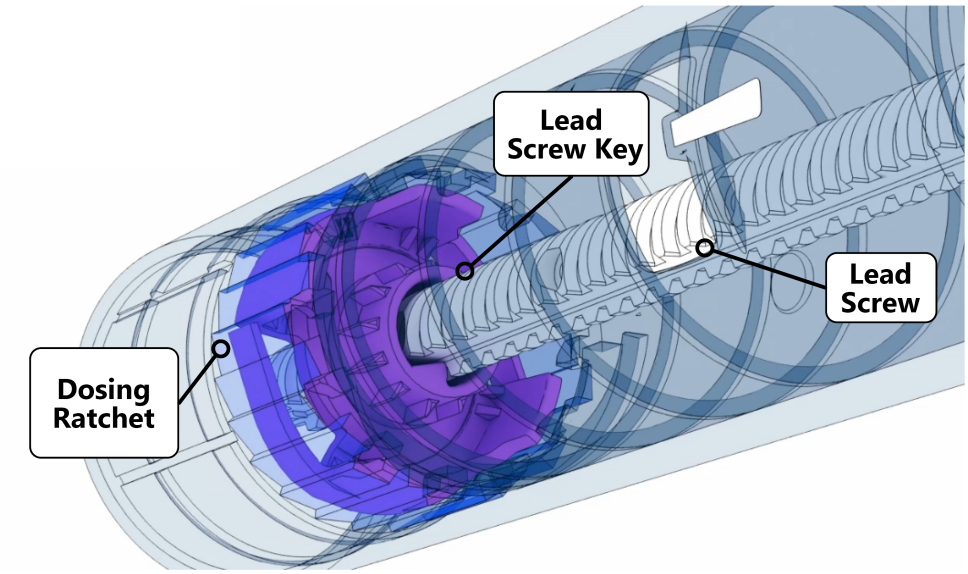


Figure 6.5: The purple component - specifically its dosing ratchet and lead screw key - compensates for most of the issues that arise from having as slender a lead screw as possible, thereby scaling the trade-off between efficiency and accuracy.

constraints. This only comes at the slight cost of efficiency, stemming from the deflection of the click arms (strain energy and frictional dissipation). The loss involved is much smaller than if the lead screw had a large diameter instead.

Condition 2 - Guidelines to Improve the Feasible Domain

As the active constraints in a design problem can influence the location of the Pareto set, they can by extension also render trade-offs unimportant if the location of the Pareto set is such that the trade-offs lead to little loss of utility. Hence, synthesising mechanical systems with bounds in mind can have a substantial effect on the performance of the end product. Considering inherent constraints while selecting working principles, and systematically arranging parts and geometric features based on the objectives at hand in order to avoid creating overly restrictive constraints, may allow the widening of the feasible domains of the variables in the design in an improving direction.

Knowing a priori which constraints are active can be challenging - especially when it comes to variables that are involved in several nonlinear phenomena. Yet, this should not prevent the designer from attempting designing around constraints that are *likely* going to be active. If we, for instance, wish to minimise the mass of a system, we can be fairly confident that manufacturing constraints relating to wall thickness, and constraints related to the avoidance of failure phenomena, will likely be active. Correspondingly, if we are interested in minimising size, the geometric fits between components and the capabilities of the manufacturing processes will definitely come into play. Yet, the designer does not necessarily need to know *which* constraints are active if the design can be manipulated in a way that affects several potentially active constraints at once. The following guidelines apply to all harmonious and independent monotonic variables and to non-monotonic variables that are bound at the optimum:

$G_{\chi 1}$: Consider inherent constraints when selecting working principles

Some constraints are inherent to the working principles in the system. Rather than being caused by decisions made in regards to the configuration and shape of parts,

they stem from the underlying physics involved or unavoidable practical limitations such as manufacture or assembly. For instance, in designing a suspension system, the constraints that arise from the selection of a pneumatic solution (e.g. seal integrity and radial piston fit) solution are wildly different from those involved in mechanical springs (e.g. shear stress, fatigue, spring index limits). Hence, the selection of working principles can drastically influence feasible domains; both of the design variables that arise with the specific principle (e.g. a piston diameter in the suspension example) and those that exist in the system irrespective of what principle is selected (e.g. suspension mounting points).

$G_{\chi 1.1}$ When possible, select working principles that avoid constraints that are not inherent to the design problem itself.

$G_{\chi 1.2}$ Introduce new functionality to eliminate active constraints - e.g. overload protection, active damping, designing to allow in-use adjustment or maintenance, etc.

$G_{\chi 1.3}$ When possible, rely on the principles of self-help [6, 16] to eliminate constraints related to mechanical failure.

$G_{\chi 1.4}$ When possible, change the design to make active constraints dependent on additional *decreasing* variables and fewer increasing variables - e.g introducing additional load bearing surfaces for a specific structural load case, eliminating load paths, etc.

$G_{\chi 2}$: Use monotonicity knowledge to widen χ through configuration

The relative arrangement of parts and geometric features in an assembly has a substantial effect on what constraints are imposed on the proportional optimization problem. How the whole system fits together creates geometric fit constraints (e.g. one part fitting inside another), tolerance chains, and force paths/loops. Hence, knowledge of monotonic relationships between the design objectives and the key dimension(s) of a functional element/part should be used to support the identification of the ideal system layout.

$G_{\chi 2.1}$ Base the layering and spatial configuration of the parts in a system on the monotonicity of its harmonious and independent variables, i.e. moving decreasing variables outward, increasing variables inward in the assembly.

$G_{\chi 2.2}$ Layer components from inside to out based on the influence of their variables on the objectives; positioning the most influential decreasing variable furthest out, and the most influential increasing increasing influence furthest in.

$G_{\chi 2.3}$ If a part contains increasing and decreasing variables that are geometrically interdependent, split the part in two or re-allocate functionality to other parts.

$G_{\chi 2.4}$ Arrange components and interfaces to take advantage of scaling/gearing effects. For instance, a rule of thumb is to locate surfaces that control the position of parts or are heavily loaded in the location in the assembly that allows the widest/largest possible dimension while locating rotating components as far inward as possible.

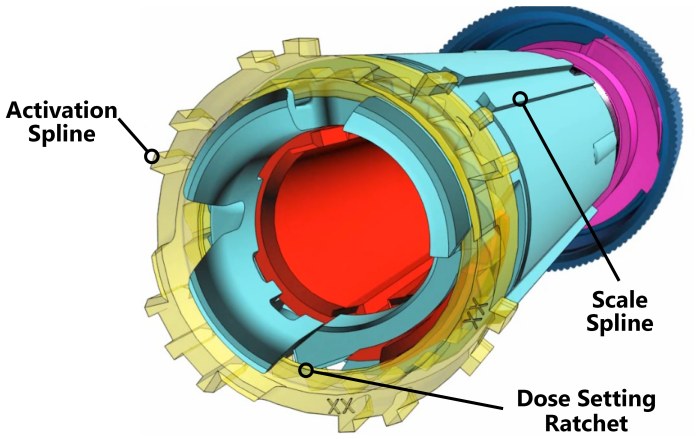


Figure 6.6: The location of the activation splines in the FlexTouch is beneficial for several reasons. The torque spring is mounted between the teal component and the red component

Example 4: Beneficial layering in the FlexTouch activation mechanism
Relevant Design Objectives: Minimise device diameter, minimize activation force, maximise dose accuracy.
Relevant Inequality Constraints: Interface stress in the activation spline, tangential assembly clearance in the spline interface, feature size (i.e. molding injection pressure).
Design Guidelines Involved: $G_{\chi 2.1}$ and $G_{\chi 2.2}$

The user activates autodosing in the FlexTouch by pressing the button at the end of the device. As mentioned in Example 1, the dosing engine works in rotation. By pushing the button, the user pushes a set of splines on a clutch component (the *Activation Splines* on the yellow part in Fig. 6.6) out of their engagement with splines in pen housing and into engagement with the purple key component from Example 3 which is free to rotate in the dosing direction. Prior to activation, the activation splines lock the spring mechanism against the housing, creating a closed force loop. This functionality could have been achieved in numerous ways but has specifically been located on the widest possible internal diameter of the device. Again, it would seem that the designers have striven for the ideal in locating this interface. From the user's perspective, a small device diameter is preferable, as is a low activation force. Pushing a button with a high force can cause considerable pain, given that this is done after the needle has been inserted.

By placing the clutch splines in the outer-most layer of the device, the designers have achieved the largest possible contact diameter. The activation force is primarily driven by the friction in the spline interface, stemming from withholding the torque spring. A large contact diameter results in a small tangential force, and therefore low friction. The low tangential force means that the mechanical stress in the device is low, meaning less material use and ultimately a smaller device. A large contact diameter also allows a high resolution of activation splines as there is a lower limit to how small features can be manufactured. Combined with lead-in surfaces for re-engagement that lower the angular error, this high resolution improves the dosing accuracy. Finally, the large diameter also *scales* down the influence of the geometric variation (which occurs in manufacture) of the spline surfaces, as a predefined absolute spline width or position tolerance has less influence on the angular error, the larger the diameter the spline exists on. In conclusion, the designers have been aware of a potentially harmonious relationship and used this knowledge to configure the parts in the system.

$G_{\chi 3}$: Design toward hitting hard constraints

While many constraints can be manipulated through design - e.g. eliminating contributions to a tolerance chain or increasing the achievable load bearing area of a snap feature by moving it to another location in the assembly - other constraints are hard and unaffected by a change in configuration. Designing towards these constraints becoming active, rather than the feasible domain being defined by (ultimately) avoidable constraints, widens the feasible domain as much as possible in the improving direction.

$G_{\chi 3.1}$ In the ideal design, all harmonious variables are determined by their general limits (irrespective of context), rather than a specific limit determined by the manner in which the functional intent has been realised.

$G_{\chi 3.2}$ If a variable is bound by a hard constraint that cannot be manipulated through configuration design change, explore changes to the overall concept or potential for parametric change (e.g. a change in the production process, material selection, etc.).

$G_{\chi 4}$: Manage the parametric contributions caused by active constraints

Oftentimes, constraints will include parameters (e.g. properties related to the material and production process), which cannot directly be manipulated by the designer. Yet, their influence and importance can still be considered in the process of synthesis and redesign:

$G_{\chi 4.1}$ When possible, relax active constraints through design rather than parametric change. Parameters can almost never be adjusted freely (toward zero or infinity) and often indirectly represent objectives beyond the designer's direct control (e.g. allowable cycle time in an assembly step, sourceable material grades, etc.). As a general rule, it is hence preferable to widen the feasible domain through design change.

$G_{\chi 4.2}$ Avoid letting features that do not directly relate to objectives but are rather necessitated by constraints affect the feasible domains of important harmonious variables (c.f. the design of the housing snap in the SOMA case discussed in Section 5.5.1).

Example 5: Layering and Lowest Theoretical Bound in the SOMA spring-trigger fit
Relevant Design Objectives: Minimise device diameter and maximise impact velocity and self-orientation
Relevant Inequality Constraints: Radial fits between parts, wall thickness limit, spring index limit
Design Guidelines Involved: $G_{\chi 2.2}$ and $G_{\chi 3.1}$

May of the initial redesign iterations in Chapter 5 might have been avoided by applying the simple guideline to layer components based on monotonicity and their influence on the objectives. Even without a fully constructed optimization model, we might realise that the configuration of components in the original design may be inopportune. With knowledge on the mechanics of springs, as found in classical books on spring theory such as Wahl [112], we might realise a spring's stiffness decreases monotonically with its coiling diameter and that its volumetric efficiency decreases monotonically with the spring index. Hence, we might realise without analysis that re-configuring the system to move the spring inward would be of benefit to the velocity objective while reducing spring-mass. That said, we might not have realised that that impact velocity is ultimately determined by the load-bearing trigger needing

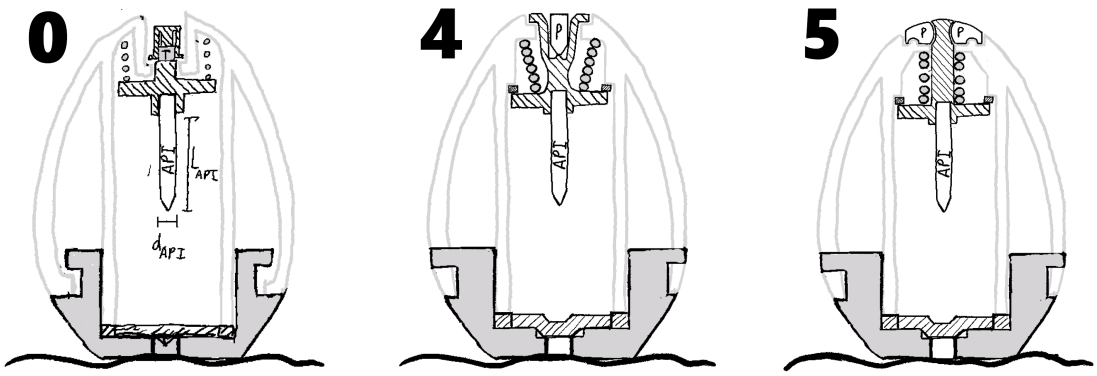


Figure 6.7: The 4th and 5th SOMA redesign iterations exhibit the notion of layering and designing towards the lowest theoretical bound

to fit inside the spring, as this issue arises due to multiple active constraints.

Yet, redesign iterations 4 and 5 (see Fig. 6.7) illustrate the guidelines related to widening the feasible domain quite well. Conceptually, the seal has no impact on impact velocity and little impact on the other objectives, so there is simply no trade-off involved in moving it outside the spring, which has a substantial impact on all of the objectives. Yet there is little utility in reducing the coiling diameter of the spring beyond a certain limit, as springs can generally not be manufactured below a spring index (the ratio between coiling and wire diameter) of 4 [112]. As discussed in Chapter 5, redesign iteration 5 most likely reaches beyond this limit, meaning that redesign iteration 4 also exemplifies the *design towards the hard minimum bound*.

Condition 3 - Design Integration

As mentioned previously, the more harmonious variables we can achieve, the less complex and conflicting the system will be. Assuming the first priority is to avoid introducing new trade-off variables, attempts to achieve low complexity in synthesis or redesign hence inevitably involves avoiding/eliminating redundant variables and increasing the number of objectives the remaining variables contribute to. Such an increase in *design integration* implies that each component in the assembly contributes to or effects more functionality [29]. One could, in other words, argue that design integration is the opposite of modularisation.

Design integration has multiple potential benefits beyond the oft claimed correlation [7, 13] between "structural" complexity (i.e. the number of design variables) and the end manufacturing cost. Examples include an increased potential for robust and reliable performance [29], given that fewer parts or fewer measures on a drawing ultimately mean that there are fewer potential sources of failure or variation. Studying efforts to reduce part counts in jet engines, Frey et al. [26] also found that complexity reduction can, in some cases, result in improved system performance, despite an increase in dependency.

In increasing the degree of integration, however, the inevitable challenge is to avoid creating new problems, either through the introduction of trade-off variables or new or more restrictive constraints. Hence, the following design guidelines may be applied to support the convergence towards fulfilling Condition 3 without moving further away from fulfilling the other conditions.

$\mathbf{G}_{\dim(\mathbf{x})\,1}$: Integrate functionality with trade-offs and constraints in mind

From a synthesis perspective, integration may involve designing components that are involved in the embodiment of multiple working principles. In redesign meaning, increasing integration might involve change such as combining parts, introducing new geometric features to existing parts and adding a state-change to the system. If care is not taken, one can easily end up making decisions that introduce new contributors to trade-offs or worsen the proportional optimum. Hence, the following design guidelines may apply:

$\mathbf{G}_{\dim(\mathbf{x})\,1.1}$ Whenever possible, integrate functionality that results in harmonious variables or the elimination of a constraint without the introduction of a trade-off variable.

$\mathbf{G}_{\dim(\mathbf{x})\,1.2}$ Integrate additional functionality as long as it does not shift bounds substantially in the non-improving direction.

$\mathbf{G}_{\dim(\mathbf{x})\,1.3}$ If variables/parts can be eliminated through the redistribution of functionality in a manner that does not introduce trade-off variables or new constraints, these variables/parts are redundant.

$\mathbf{G}_{\dim(\mathbf{x})\,1.4}$ Integrate whenever multiple functions can be performed over the same axis of operation (e.g. rotation around a given axis), so long as this does not introduce non-scalable trade-off variables, overly restrictive bounds on important harmonious variables, or result in an overconstrained mechanism.

$\mathbf{G}_{\dim(\mathbf{x})\,1.5}$ Design towards achieving *state changes*. As a rule of thumb, the more kinematic state changes (e.g. parts changing interfaces or kinematic degrees of freedom, and load paths being redirected) a designer is able to build into a mechanical system, the more functions and objectives each part can contribute to. This does not necessarily create trade-off variables or necessitate additional design variables given that independence is achieved *in time* rather than geometry. Hence, this is somewhat analogous to the *Separate in Time* heuristic from TRIZ.

$\mathbf{G}_{\dim(\mathbf{x})\,2}$: Separate to avoid trade-off variables or inherent constraints

Oftentimes, separation becomes the only recourse, as some forms of functionality cannot be integrated without creating trade-off variables that cannot be scaled or inherent constraints that cannot be relaxed. TRIZ [16] contains a quite expansive treatment on different approaches to separation, so the following guidelines are only stated in the specific context of a designer trying to get as close as possible to fulfilling the Conditions of Ideal Design:

$\mathbf{G}_{\dim(\mathbf{x})\,2.1}$ Avoid integrating *physically contradicting*[16] functionality in the same parts/subsystem - e.g. requiring a part to be stiff and compliant, insulating yet conductive, etc.

$\mathbf{G}_{\dim(\mathbf{x})\,2.2}$ Split parts or introduce new ones and redistribute functionality, if the alternative is a an active constraint or a trade-off variable that cannot be scaled.

$\mathbf{G}_{\dim(\mathbf{x})\,2.3}$ Only modularize and parallelize the system when the alternative is a trade-off or a substantially narrowed feasible domain in the improving direction.

This will often be the case in products that are maintenance-heavy or in architectures with a high degree of part re-use, where increased integration might lead to increased cost.

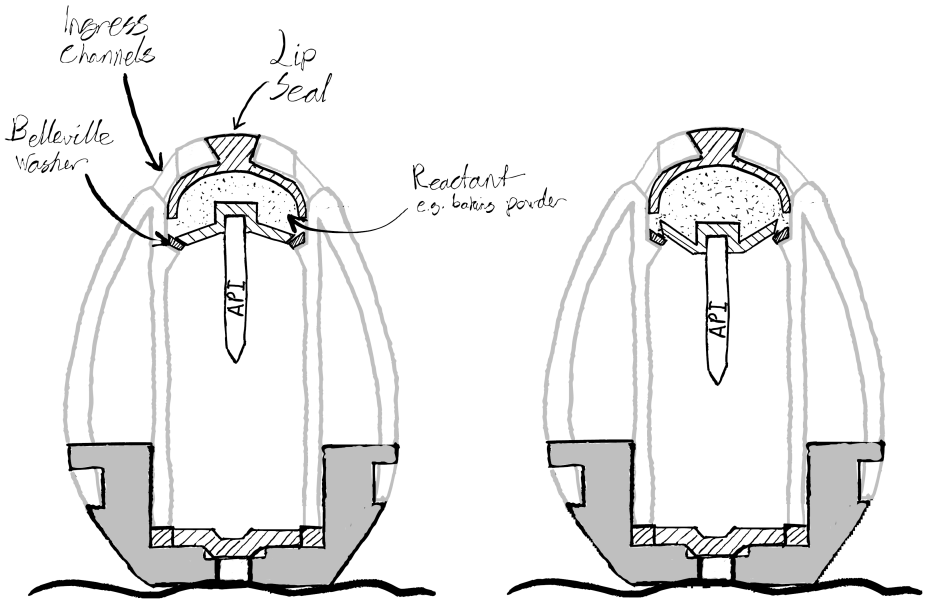


Figure 6.8: An alternative actuation principle that allows the actuator to trigger injection-using dynamic rather than static loading - permitting fewer parts and eliminating multiple constraints and potentially also trade-off variables.

Example 6: Selecting working principles in the SOMA
Relevant Design Objectives: Maximise impact velocity and self orientation
Relevant Inequality Constraints: Spring yield, trigger creep load, spring manufacture, and radial and axial fits
Design Guidelines Involved: $G_{\dim(\mathbf{x})\,1.1}$ and $G_{\chi 1.1}$

Looking at each redesign of the SOMA device in Chapter 5, one could argue that the concept remains unchanged. They are all self orienting devices with crystalline API. They all inject the API by releasing a pre-loaded spring after the dissolution of the component that held it in place. These principles inevitably lead to several constraints and trade-off variables that are inherent to a statically loaded system. The pre-loaded spring means that the device needs to be designed with a load-bearing structure that can withstand a long term static load yet remain lightweight. This would never be avoided completely using the Configuration Redesign Principles.

In applying the design guidelines for ideal design, an obvious question is: what if the device were not pre-loaded but triggered by the build-up of a force/pressure? In essence, **integrating** the actuator and trigger functionality, thereby eliminating some of these globally active constraints. Such a system implies some sort of state change, needing to occur after the user swallows the device. One answer to this is a chemical reaction with the contents of the stomach resulting in the build-up of gas inside the device (e.g. baking powder, which would produce CO_2), or some sort of swelling reaction (e.g. a hydrogel). Correspondingly, a build-up of pressure requires a sealing system, and some form of elastic or bi-stable release mechanism would also be required.

This line of thinking led to a conceptual sketch of such a system is shown in Figure 6.8. Essentially, gastric acid/fluid could enter the device via channels covered by a self-reinforcing lip seal or a membrane. The API is mounted in a bi-stable Belleville washer, which seals against a guiding cylinder in the housing. Behind the washer, a reactant of some sort, causing pressure build-up upon contact with gastric acid, to a point where the Belleville washer reverts to its opposite position and deflects past a ratchet surface in the housing, thrusting the needle and washer forward. This eliminates the constraints stemming from the static load from the spring, the spring itself, and the need for the dissolving plug. Furthermore, the reactant can be distributed much more freely in the device (volumetrically) than a spring, potentially allowing less mass in the top of the device. That said, we cannot claim that this is necessarily a solution that is closer to the ideal design, as there are multiple open questions. Examples include the integrity of the seal/membrane on the top of the device, sealing integrity between the Belleville washer and guiding cylinder, energy density of the reactant, etc. Yet, it serves to illustrate that the guidelines can lead to vastly different solutions, with other (perhaps less restrictive) constraints and trade-offs involved.

$G_{\dim(\mathbf{x})\,3}$: Consider the hierarchy of trade-off avoidance

If we accept the Third Conjecture of Ideal design, Condition 3 leads to the realisation that when possible, it is preferable to avoid trade-offs through design decisions that do not introduce new design variables/parts. If not, we would simply be mitigation trade-offs by increasing complexity. Hence, if we wish to fulfil Condition 1 and 3 simultaneously, there is an order of preference as to how to eliminate (or reduce) a contribution to a trade-off:

- $\mathbf{G}_{\dim(\mathbf{x})\,3.1}$ Flip monotonicity over all else - attempt to achieve like monotonicity in the selection of working principles and their combination into an overall system.
- $\mathbf{G}_{\dim(\mathbf{x})\,3.2}$ Eliminate the trade-off variable, by removing the its influence on one objective entirely. This especially applies to *unnecessary* influences (see $G_{\dim(\mathbf{x})\,4}$).
- $\mathbf{G}_{\dim(\mathbf{x})\,3.3}$ Separate the trade-off variable by redistributing functionality to existing geometry/design variables/parts.
- $\mathbf{G}_{\dim(\mathbf{x})\,3.4}$ Separate the trade-off variable by introducing new design variables/features onto existing geometry/design variables/parts.
- $\mathbf{G}_{\dim(\mathbf{x})\,3.5}$ Scale the trade-off variable using existing variables. This can for instance be achieved by relaxing the constraints on variables that act as a multiplier/divisor to the trade-off variable.
- $\mathbf{G}_{\dim(\mathbf{x})\,3.6}$ Separate the trade-off variable by introducing new parts/subsystems
- $\mathbf{G}_{\dim(\mathbf{x})\,3.7}$ Scale the trade-off variable by introducing new parts/subsystems

$G_{\dim(\mathbf{x})\,4}$: Avoid unnecessary influences

Trade-offs are caused by dependency. Yet, some forms of dependency are not inherent to the concept of configuration but rather arise unintentionally due to what can essentially be viewed as noise. These situations should be avoided whenever possible:

- $\mathbf{G}_{\dim(\mathbf{x})\,4.1}$ Aim for kinematically correct designs, as static indeterminacy can lead to non-linearities and unintended dependencies [14, 39].
- $\mathbf{G}_{\dim(\mathbf{x})\,4.2}$ Avoid the associated/parasitic loads that arise from asymmetric parts and

load paths, or from unbalanced moments[14]

$\mathbf{G}_{\text{dim}(\mathbf{x})\,4.3}$ Isolate negatively interacting subsystems from each other to avoid e.g. unintended friction, heating, vibration, competing working directions, etc. [94, 115].

$\mathbf{G}_{\text{dim}(\mathbf{x})\,4.4}$ Whenever possible, avoid designing geometric features required for manufacturing and assembly in a manner that influences functionality, to limit cost vs. performance trade-offs.

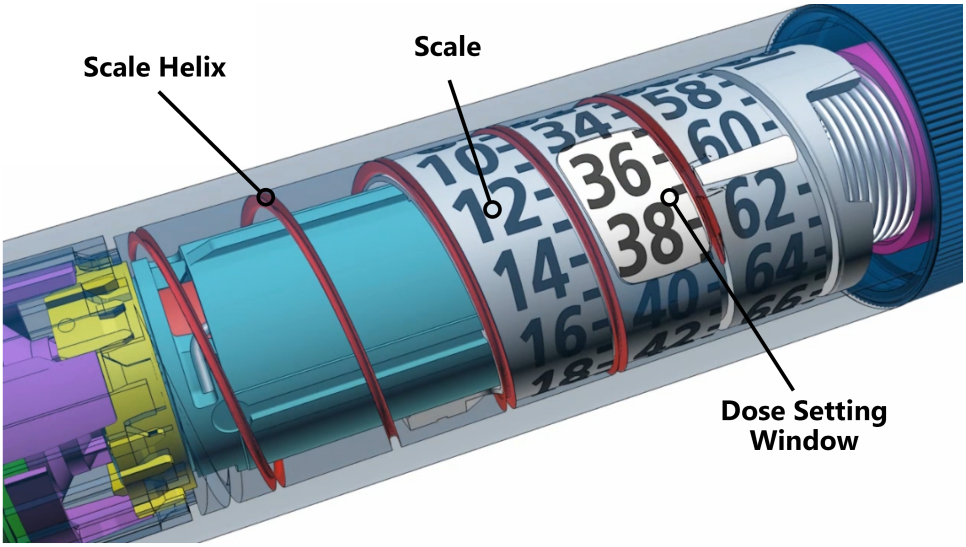


Figure 6.9: By driving most of the functionality in the dosing engine using a single torque spring, and using working principles for sub-functions which all rely on rotation, the FlexTouch device has avoided numerous contributors to the trade-offs between synchronisation and angular accuracy on one side, and mechanical efficiency on the other

Example 7: Beneficial Integration in the FlexTouch Scale and Click System
Relevant Design Objectives: Minimise device diameter, scale display error, maximise dosing speed, and scale number size and resolution,
Design Guidelines Involved: $\mathrm{G}_{\text{dim}(\mathbf{x})\,1.1}$, $\mathrm{G}_{\text{dim}(\mathbf{x})\,1.4}$, $\mathrm{G}_{\chi 2.1}$, $\mathrm{G}_{\chi 2.2}$

The entire FlexTouch dosing engine is designed to act in rotation, with the exception of the activation mechanism (which has benefits, as Example 4 illustrates). Dose setting, autodosing, dose clicks, and the dose scale are all driven by the torque spring mechanism and by the user turning the dial. The scale is rotationally locked to the spring mechanism via a set of splines (see Figs. 6.6-6.9) and is mounted on a helix inside the housing, meaning that it is *screwed* back and forward as the user sets a dose and the device auto-doses. Beyond acting as a means of communicating the dose setting, the scale also acts as a rotational lock, preventing the device from being dialled above the maximum dose setting and preventing it from dosing below its "zero" setting.

When viewed from a single-objective perspective, there might very well have been benefits in designing single modules to perform some of these functions independently - e.g. getting as loud clicks as possible or achieving larger on the scale or a finer resolution. This might also have had modularity benefits - allowing more variants and more treatment regimes to be covered by the same device platform.

Yet, as examples 1, 3 and 4 might already indicate, substantial benefits arise thanks to the integration of functionality. Chief among these is that a less integrated design would likely have resulted in a larger device and less accurate device. Especially given that additional parts/variables add to the number of sources of geometric variation. At its lowest dose setting (e.g. one standard unit of insulin), the device converts the rotational movement of the dial into app. 0.18mm of axial movement inside the drug cartridge. The use of a single actuator (the torque spring), and the use of the scale the rotational end-stop, minimises the deviation between the dose setting shown on the scale and the size of the delivered dose. Having located the max/min rotational stops on the scale also maximises the accuracy and repeatability of the device, as it is the outer-most component inside the device, yielding the largest possible contact diameter. The single actuator has also allowed the parts to be layered in a manner that minimizes friction without a loss of accuracy.

The use of rotation as a working direction also any geometric variation to be geared down, as any error in angular position is transformed into a comparatively small error in axial displacement thanks to the lead screw. An axial spring mechanism would likely require a serial arrangement of spring and the functional element that pushes the plunger in the cartridge (replacing the lead screw), which would lengthen the device or reduce the size of the deliverable dose. Furthermore, such a solution might involve a ratchet rod instead of a lead screw, which would worsen the accuracy of the device, given that linear tolerances would translate directly into a positional error of the ratchet rod. Not to mention that it would be challenging to get a fine enough dose setting resolution (as each increment equates to 0.18mm of movement).

6.2 Systematic Iterative Design

> "*It is characteristic of the search for alternatives that the solution, the complete action that constitutes the final design, is built from a sequence of component actions. The enormous size of the space of alternatives arises out of the innumerable ways in which the component actions, which need not be very numerous, can be combined into sequences.*"
>
> - Herbert Simon
> The Science of the Artificial [134]

As the above quote in part communicates, embodiment/configuration design is challenging, especially in synthesis and redesign. There are often so many forms of changes one might make that identifying how and when to change a design is by no means trivial. Now that we have developed guidelines for synthesis that are consistent with the developments of the previous chapters, we can move on to the question of how to apply the methodological developments systematically in iterative design. This section hence involves a discussion of how the developments of the previous chapter actually allow a systematic approach to guiding configuration design towards improvement and presents perspectives on how this is done in practice.

6.2.1 Supporting Design with Optimization

The breadth of options available to the designer as to how to embody the desired concept is substantial. The same functionality can at times be achieved through countless system layouts, working principles, part combinations, and so on. As a result, the transition from concept design to embodiment design is commonly supported by some form of morphological analysis [6, 9] and is heavily reliant on iteration.

As argued by Pahl & Beitz [6] and by Andreasen [8] when a desirable or promising preliminary system layout (the initial embodiment) has been identified, the process of refining said layout

begins. This might involve adding additional sub-functionality, redesign aimed at "*Design for X*" (e.g. manufacture, assembly, safety, robustness, aesthetics, etc.), and changes aimed at general improvement of product performance. In this process, design changes and improvements are made at different degrees of change, some of which can be supported through the use of optimization methods. Design optimization typically iterates on the design of the system to achieve optimality with respect to given objectives while satisfying a set of given constraints. Such optimality is limited to the particular analytical model, and so it is often more accurate in practice to think of optimization results as design improvements. Such improvements are associated with (i) proportional design changes, where design variables are resized; (ii) parametric changes, where parameters such as material properties or production processes are modified allowing for relaxation of design constraints; and (iii) configuration changes where the embodiment of functions or distribution of sub-functions is modified.

As covered in Paper B, proportional design has historically been facilitated by advancements in computational modelling, computing power, and optimization techniques. More robust and higher fidelity numerical solutions in size optimization have reduced risk in constraint relaxation for increasingly larger and more complex problems. Parametric design with gradual constraint relaxation has been a primary approach for product performance improvement in industry [131]. This has allowed designers to work within enlarged feasible domains thanks to new or improved materials, production processes, and a deeper understanding of failure phenomena, all leading to reduced design margins, e.g., as seen in the design of combustion engines and turbines [131]. With this in mind, the comparatively low focus on configuration/embodiment change in academia and industry is somewhat surprising. Entire fields of engineering are devoted to material, and process improvement and increasingly advanced mechanical analysis and modelling, but the question of embodiment/configuration design is largely reliant on a more practical approach, relying on a mix of creative synthesis, analysis, and engineering judgement [6, 12], being largely dependent on the proficiency and experience of the designer [10, 33, 35].

As discussed in chapter 3, configuration/embodiment design remains elusive in design optimization, with the notable exception of topology optimization [67], and to an extent combinatorial methods such as grammars [106] and graph-based methods allowing simultaneous synthesis and optimization [107, 108].

The fundamental challenge of optimization-based configuration design is the lack of appropriate mathematical modelling capabilities: different configurations require different mathematical problem formulations. The success of topology optimization (TO) methods is due to the introduction of a unified mathematical model across configurations. TO allows optimization of a functional representation of the design without the actual embodiment of the functions, and the results inform configuration design. However, both TO and combinatorial approaches have limitations in the context of early iterations in configuration design. Topology optimization is reliant on a predefined set of boundary conditions and loads, meaning it either involves the design of single components or requires a fixed configuration of components. Meanwhile, combinatorial optimization techniques can only capture the configurations that the designer has accounted for in the model, meaning that these have to a certain extent, been synthesised a priori. Both combinatorial and tensor-based techniques also fail if the physics or the boundary conditions change with the configuration.

The ultimate goals of these methods - identifying an *optimal configuration* - perhaps disregard some of the fundamental challenges engineering designers face in practice. Their work is highly iterative [6, 24], in part due to changing product requirements, and in part due to the gradual aggregation of design knowledge [10], permitting informed design changes.

Product functionality is also not static - it changes and grows as development progresses [8]. As a result, what constitutes an *optimal* design changes over time. Furthermore, designers are inherently opportunistic, "perceiving those requirements which are likely to constrain the design most critically, and using a knowledge of the design repertoire and experience" to inform decision-making [97]. They also often "try to identify, remove/reduce, then optimize the trade-off" [124]. Thus, as the designer progresses through successive iterations, the configuration design changes along with the objective and constraint functions. In effect, the designer manipulates the Pareto set by introducing changes that improve performance. They also gradually change its dimensionality by adding sub-functions or considering more requirements (e.g. manufacturability), thereby adding objectives and constraints.

With this in mind, it is perhaps more pertinent to consider how design optimization might be used to inform decision making in the iterations between the selection and combination of working principles, to the synthesis of the preliminary layout and the definitive one (i.e. from the start to the finish of embodiment design). Disregarding the potential *loop-backs* [24] into conceptual design, this configuration/embodiment design involves designers implementing a mix of proportional, parametric, and configuration related changes, along with the introduction of additional features (new functionality or new requirements). In practice, all three alternatives might often be explored simultaneously, for instance leading to the parallel exploration of a material change, resizing, and reconfiguration in attempting to solve an issue or improve the system's overall performance.

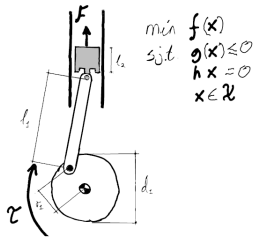
Rather than attempt to merely identify the optimum or automatically synthesise optimal configurations, the analysis and redesign methods presented in the previous chapters are suggested as an approach to using design optimization techniques interactively in the embodiment design process. Specifically, they can be used to formally explore what is possible within the proportional and parametric domains of design change while also identifying the limitations of the configuration design and how these might be mitigated. Furthermore, the guidelines for synthesis developed in the previous section allow the upfront avoidance of many of these limitations, potentially providing a *better* starting point for the subsequent configuration design iterations.

This constitutes an entirely different category of quantitatively driven configuration design, separate from the aforementioned topology and combinatorial optimization methods. The work done by Jain and Agogino [102], Cagan and Agogino [32], and Ishii and Barkan [103], all touches upon this aspect, but not to the extent of presenting an overall methodological framework for interactive optimization and configuration design.

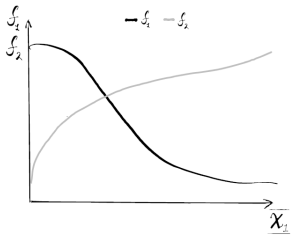
6.2.2 Interactive Trade-off Analysis and Mitigation

So, because configuration design involves gradually changing product functionality and requirements and an aggregation of knowledge, a more flexible approach is necessary if we wish to reach the "best" configuration/embodiment design before transitioning into detail design. Looking at the steps involved in embodiment design described by Pahl & Beitz [6], we see that embodiment design is a mix of synthesis, evaluation and optimization, and decision making (e.g. selecting the preliminary layout). Thus embodiment, as discussed, requires a mixture of creativity, an aptitude for analysis, and engineering judgement.

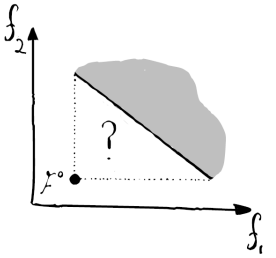
With this in mind, the interactive use of proportional multiobjective design optimization, trade-off analysis, and trade-of mitigation is proposed as a rigorous framework under which to structure the embodiment design process, as illustrated in shown in figure 6.10. This process may be applied after the synthesis of the first embodiment. Applying the Guidelines for Ideal Design Synthesis on beforehand might help avoid many of the issues one might identify in analysis, potentially reducing the number of iteration cycles.



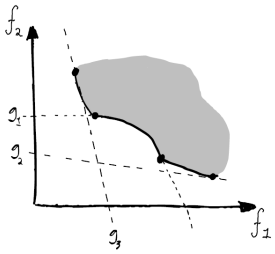
1 CONSTRUCT/UPDATE/ REFINE THE MODEL
Build a quantitative description of the design problem



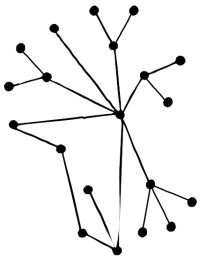
2 MULTI-OBJECTIVE MONOTONICITY ANALYSIS
Find trade-off variables and constraints causing them



3 NUMERICAL SOLUTION
Identify the Pareto set, and additional active constraints not found in step 2.

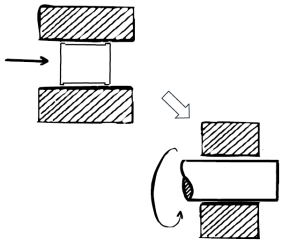


4 ε-MONOTONICITY ANALYSIS
Find Pareto-constraints and regional dependencies

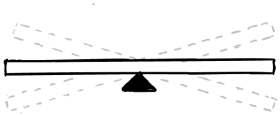


5 MAP OUT THE REDESIGN OPTIONS
Explore which of the design principles can realistically be applied.

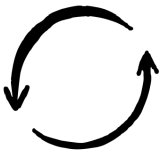
6 IDENTIFY REDESIGN SEQUENCE
Apply the prioritization scheme to ensure influential changes.



7 ITERATIVE SOLUTION EXPLORATION
Identify and implement design changes that fulfil the principles.



8 REDESIGN EVALUATION AND SELECTION
Evaluate the redesigns informally, opportunistically, or exhaustively, and select one.



••• INTERACTIVE ANALYSIS AND REDESIGN
Repeat the steps upon the introduce of new functions, or for further improvements.

Figure 6.10: Combining design optimization with Pareto set Dependency Analysis and Systematic Configuration Design Improvement and applying them interactively, one can rigorously guide the embodiment design toward improvement. Many of the issues identified and mitigated through this process could potentially be avoided through the application of the guidelines from Section 6.1.

Using the first outline of an embodiment as a starting point - preferably one that has been synthesised with the conditions of ideal design in mind - the application of Pareto set dependency analysis (steps 1-4) essentially involve the identification of the Pareto set and the dependencies and constraints that create and shape it. A critical part of this process is the

derivation of the objective functions and constraints of interest as a part of the model construction. Subsequently, Systematic Configuration Design Improvement (steps 5-7) uses the knowledge gained from this analysis to identify mitigations to the trade-offs in the design.

Given the theorems and proofs in chapters 4 and 5, the resulting redesigns should fulfil the definition of *design improvement* from chapter 5 w.r.t. the modelled objectives. As shown in the SOMA case in chapter 5, the result of this analysis and redesign process can be a set of redesigns (as a consequence of step 7), the most promising of which needs to be selected (step 8) for further work. Different approaches to evaluating redesigns are described in Section 6.3. After the initial round of optimization, analysis, and redesign, the procedure can be applied continuously throughout the embodiment design process in an interactive manner. Here, *interactive* is used in two senses; firstly is meant in the sense that while methods do involve optimization, they do not constitute an *automation* of design in the same sense as topology optimization. Rather they allow the designer to make targeted design changes based on the results of optimization and systematic analysis and repeat the process with a new embodiment design(s). Secondly, it is also interactive in the sense that one might apply different parts of the procedure outlined in figure 6.10 at different stages or only repeat parts of it as design objectives, constraints, and new sub-functions are added. Examples of partial applications of this process might include:

1. *Acceptable Optimum* (steps 1-3): If the Pareto set contains solutions that are sufficient for the application - meaning there is little utility in improving the optimum through configuration redesign - one could in principle decide that the subsequent steps are unnecessary.

2. *Acceptable redesign* (steps 1-8, followed by steps 1-3): Upon the identification of new embodiment designs, based on the analysis and mitigation of the trade-offs in the initial one, one might conceivably reach an acceptable performance. Hence, it would not be necessary to repeat the process beyond step 3, which would reveal the optimal proportions of the new embodiment.

3. *Parametric Change* (steps 2-4): The influence of parametric changes - e.g. new materials or production processes - might be studied by simply re-running the numerical model and updating the analyses in steps 2 and 4. This might reveal new or different trade-offs that occur due to the changes in constraint activity that arise as a result of the parametric change.

4. *Qualitative trade-off knowledge* (steps 5-7): The designer might combine the results of steps 1-4, with tacit knowledge of unmodelled trade-offs, to repeat or redo the redesign steps in order to reach further design improvement, without redoing the analysis steps.

5. *Added objective(s) or constraint(s)* (steps 1-4): How *well* an embodiment design is able to fulfill new requirements, can be studied by simply updating the optimization model and trade-off analysis. Whether new or worsened trade-offs arise would then reveal whether it would be an advantage to change embodiment. This might, for instance, occur when new sub-functionality is added to the system, resulting in new or changed design objectives and constraints.

6. *Loop-back* (step 1-5 followed by step "zero"): As also argued in the quote from Pahl & Beitz [6] in Section 6.1., the limitations of a concept will often become evident through the analysis of a resulting embodiment. Hence, the application of the analysis steps of the iterative procedure may reveal limitations of a conceptual nature, which might only be mitigated by moving back to the conceptual stage, and making bigger changes

than what the configuration redesign principles may help reveal. Such changes might include a change in working principle, design objectives, or changing the functionality of the product itself. Here, the outputs of analysis, combined with the Guidelines for Ideal Design Synthesis, may help select more suitable working principles and system layout, resulting in a wholly different concept.

Hence, this process can be applied iteratively throughout the embodiment design process. With each successful repetition of this process, additional novel embodiments/configurations are identified, which might then again be optimized, analysed, and improved further. When applied systematically, this would lead to the gradual convergence toward the "best" possible configuration design, illustrated in figure 6.11.

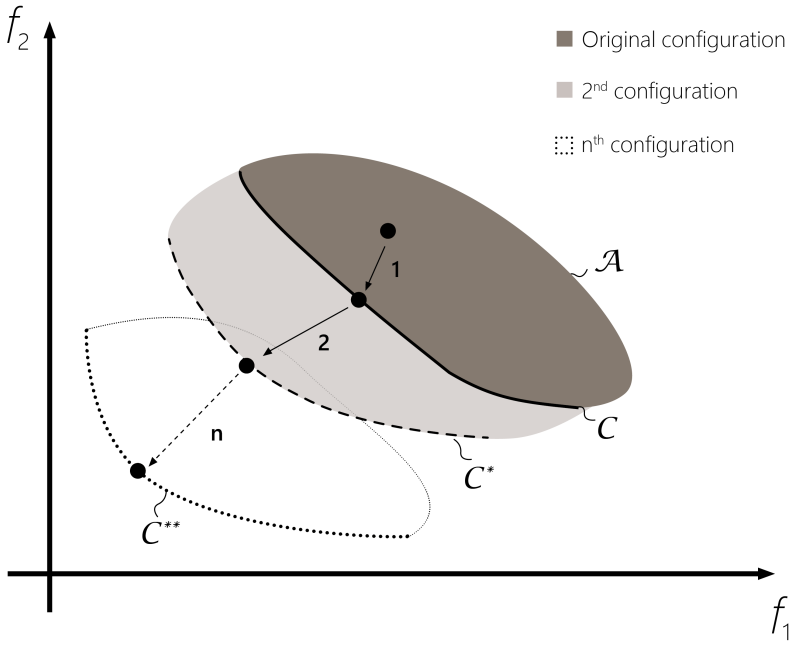


Figure 6.11: By interactively identifying the optimal set (1), mitigating the dependencies and constraints that cause it (2), and repeating the process, the configuration design can be systematically improved.

The interactive analysis and redesign process also has applications in practice beyond attempting to gradually and systematically identify improvements from an initial starting point (the first embodiment). It can also be used to explore whether a loop-back into design tasks of a more conceptual nature is necessary. Furthermore, one could repeat the process in part to account for the changes that occur to the system in the embodiment design process, which are not captured by the original model. One might, for instance, need to fulfil new requirements; the effect of these can be studied through the introduction of new constraints or objectives in an updated optimization model. Another common scenario is the need to introduce a new sub-system or add additional features to certain components to fulfil new functionality. Here, the interactive process can be applied to evaluate the effect of this and identify new drivers of which might be mitigated, thereby identifying more suitable overall embodiments when these new functionalities, which were disregarded in previous synthesis efforts, are taken into account. Finally, as one gains new knowledge, e.g. of the in-use conditions or of the physical phenomena that determine the constraints and objectives, the interactive process might be re-applied. This would allow more accurate identification of the Pareto set and reveal additional trade-off information, which could be used to derive more

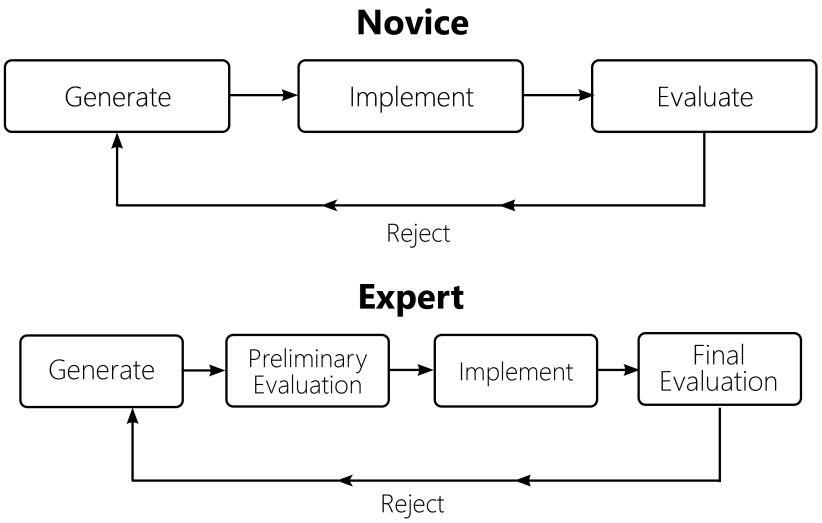


Figure 6.12: Figure adapted from [33]: The behavioural patterns of novices vs. experienced designers in synthesis and problem solving, as found by Ahmed et al [33].

embodiments that are *better* when taking this new knowledge into account.

6.2.3 A Formalisation of Experience

To understand why the proposed procedure is meaningful in the embodiment design process, we need to look into the question of how expert designers behave in practice. As argued by Finger and Dixon [99, 135], a key challenge in mechanical design research at practice is the development of simple analytical models that allow the designer to assess the *whole* rather than single disciplinary, single objective aspects of the behaviour of the overall system. In effect, Pareto-set Dependency Analysis allows a form of multi-disciplinary reasoning about the system. No rigorous analysis method can reveal whether a design is "good" or not, but in applying MOMA and ϵMA, we can at least find what dependencies and constraints limit its achievable performance. If these limitations can be mitigated, we will reach a "better" design. Here, we use the proportional domain to build an understanding of the limitations of the design on a configuration- or even conceptual level.

Importantly, this mimics the behaviour of experienced designers. As mentioned, experienced designers are inherently opportunistic [136], and are aware of the trade-offs [33] and active constraints involved in their design tasks [59, 132]. Ahmed et al. [33] found that a key aspect of this is that experts evaluate and reason about their designs in a different manner than novices, as illustrated in Fig. 6.12. Experienced designers are inevitably leveraging their knowledge of trade-offs, and active constraints in this *preliminary evaluation* step to identify improvements that can be implemented in the next step. As argued by Howard and Andreasen [20, 124], trade-off identification and avoidance are key steps in embodiment design, yet as found in Chapter 3, existing methods do not seem to support such efforts. This might explain why the designer's experience becomes the distinguishing factor.

One could hence argue that the iterative application of Pareto set Dependency Analysis and Systematic Configuration Design Improvement essentially replicates this mode of *functional reasoning*. By providing an approach to identifying the drivers of trade-offs and an understanding of how the active constraints locate and shape the Pareto set, the interactive procedure essentially provides the designer with the same knowledge that the "experts" are leveraging to increase their likelihood of success. Thus, one could view the Interactive Trade-

off Analysis and Mitigation approach as a systematisation of the craft of *good* embodiment design, allowing the designer to rely on analysis rather than tacit knowledge and intuition.

6.3 Redesign Evaluation: Comparing SOMA Redesigns

Recall that the previous chapter introduced a formal definition of design improvement; the Pareto-set of a redesign dominates the Pareto-set of the original design. With this basic definition, a set of theorems followed the proofs for which show that specific forms of design change will always result in design improvement. As seen with the SOMA case, one can apply the redesign methodology to synthesise a sequence of redesigns, based on the gradual application of the redesign principles, using the insights gained in analysis. Yet, one could conceivably also encounter cases with analysis results that open up alternative routes to improvement, leading to several alternative redesign sequences involving mutually exclusive design changes. Drastically different configuration redesigns would arise as a result.

Hence, design evaluation is essential in the decision making involved in systematic iterative design. One might wish to evaluate whether the design changes introduced actually resulted in design improvement. In the case where there are multiple alternative redesigns, a question of equal importance is; which of the redesigns is most *desirable*? As seen in the SOMA case, some design changes introduce new trade-off variables, the influence of which may be small enough that the design improvement criterion is still fulfilled.

Exhaustively evaluating a redesign relative to the initial design would require a comparison of the Pareto-sets of the new designs with the initial Pareto set. This means that a new optimization model (or an updated version) would need to be constructed for each redesign being evaluated. However, the analysis and redesign methods from the previous chapters have been developed to be applied during the early phases of design. At this stage, it is by no means a given that the designer has the time or the resources to build multiple optimization models in sequence, especially if the analysis and redesign methods are being applied iteratively.

Hence, this thesis introduces three different levels of evaluation; (1) *informal evaluation*, (2) *opportunistic evaluation*, and (3) *exhaustive evaluation*, which can be used during or after an iterative design improvement process, depending on what knowledge the designer wants to gain. The higher the level, the larger the analysis effort, and the lesser the re-use of the information gained from the application of MOMA and ϵMA to the initial design. The underlying rationale here is that the knowledge gained during the analysis of the initial design is not necessarily lost just because the design is changed. On the contrary - so long as we have applied the redesign principles systematically, we know exactly which relationships have changed and which ones have not. Thus we can, to an extent, leverage this knowledge to evaluate whether certain trade-offs have been reduced/mitigated or whether the optimum of certain objectives has been improved, thereby shifting the Pareto frontier. In the following, each of these levels of evaluation is demonstrated using the SOMA case.

6.3.1 Level I: Informal Evaluation

If we were to attempt to evaluate a redesign w.r.t. Definition 5 using as little effort as possible, we could employ an informal approach. Constructing a new or updated optimization model to reflect the design changes made relative to the original design might be time-consuming, especially when it comes to setting up all the constraints, ensuring well-boundedness, and getting the model to converge. Hence, instead of comparing entire Pareto-sets, we might gain insights evaluating a redesign at a single point. By re-using information from the original analysis to informally model and -optimize the redesign toward a single *Pareto-adjacent* design point, we can get a useful indication of the influence of the design changes:

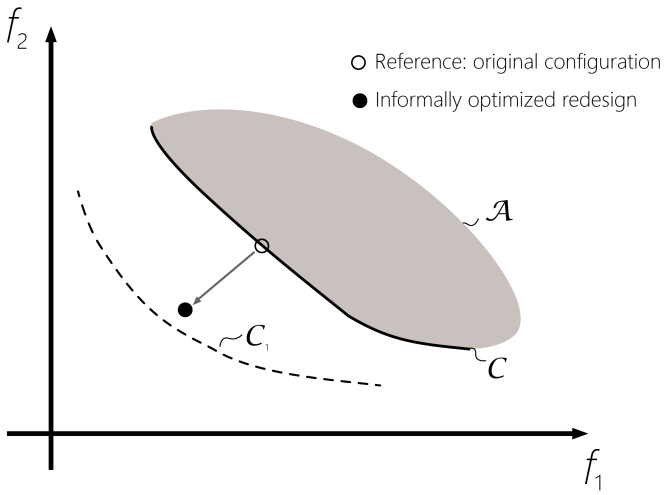


Figure 6.13: Informal evaluation involves the re-use of monotonicity and constraint activity knowledge from the preceding analysis to identify a single design point in the objective space for the redesign, allowing comparison.

Level of Abstraction: Evaluation of updated objective functions for a redesign at a single design point in the objective space, and then comparing it with the Pareto-set of the initial design.
Re-used Knowledge: Monotonicity and constraint activity information from the original analysis, combined with elements of the optimization model that are unaffected by design changes.
Steps: An informal evaluation process might involve:

1. If necessary, updating the objective functions to reflect the design changes.
2. Modeling the design in CAD, using monotonicity information to informally optimize it.
3. Using the updated objective functions or CAD/CAE to evaluate how well the informally optimized design meets the objectives.
4. Comparing with the initial or preceding design

Uses: Assessments of whether the configuration redesign principles have been successfully applied. Simple comparisons with alternative redesigns, assuming approximate knowledge of the relative importance of the objectives exists.
Potential limitations: If there are new or changed constraint functions, resulting in unforeseen constraint activity, then this design point may be infeasible unless this is accounted for during evaluation. Furthermore, the informal nature of the proportional design of the new configuration(s) means that we might be comparing the optima of the original design with a design point that is far from the new optimal set.

Essentially, we cannot identify a Pareto optimal design point for the new design without an optimization model. However, a lot of the knowledge gained in the original analysis can still be utilized. The monotonicities of the design variables which were not targeted by the systematic application of the configuration redesign principles will remain unchanged. Correspondingly, we should know exactly how constraint activities and the relationships between the objectives and their trade-off variables have changed as a consequence of the design changes. Thus, we might be able to informally optimise the system while dimensioning it in

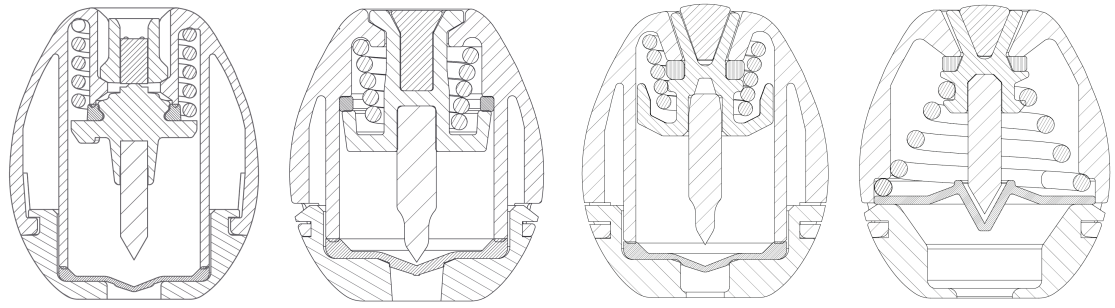
a CAD environment. This may take far less time than building an optimization model that is valid and convergent across the feasible design space. One might also utilise CAE tools to evaluate the objectives and constraints affected by the design changes rather than update the functions from the original optimization model.

The new point could then either be compared with the original entire Pareto set or with a reference point in it. Assuming no trade-off variables or new constraints have been introduced by the design changes, this might indicate how far the new Pareto-set has been moved or reshaped. In ongoing development projects, one might often have an already dimensioned version of the original design in CAD, at the point in time in which the methods discussed in Chapters 4 and 5 are applied. This can be used as a reference for comparison, as it at least to an extent reflects the relative weighting of the different objectives, given that it will have a certain relative performance. As seen in the SOMA case, the reference design used to build the optimization model was actually found to be quite close to being Pareto optimal already.

Should one find that this informally optimized design lies beyond the original Pareto-set, we might already be sufficiently confident that the design improvement criterion is fulfilled, given theorems 5 and 6, along with their associated corollaries. In extreme situations, the single new design point may even dominate the entire original Pareto set.

This knowledge might allow comparisons with other configuration redesigns for selection purposes, help decide whether the construction and solution of a whole optimization model is worthwhile, or substantiate that investment in prototyping and testing the redesigned configuration is worthwhile.

Case: CAD-based Evaluation of the SOMA Redesigns



	SOMA: *Reference design*	**Iteration 4:** *"Flipped seal"*	**Iteration 6** *"Wedge Trigger"*	**Iteration 11** *"Flipped Actuator"*
Z_{CM}	-0.65 h	-0.67 h	-0.68 h	-0.69 h
Ø	11.12 mm	8.37 mm	9.05 mm	9.05 mm
M_{API}	3.15 mg	3.4 mg	3.4 mg	3.4 mg
V_{IMP}	26.3 m/s	33.7 m/s	37.5 m/s	39.4 m/s

Figure 6.14: A comparison of the original SOMA device with proportioned realizations of three of the redesign iterations.

As three out of the four design objectives in the SOMA device can be evaluated through relatively simple means in a CAD environment, three of the redesign iterations were modelled and dimensioned in full in CAD (PTC Creo Parametric 4.0). Namely, the 4^{th} , 6^{th}, and 11^{th} design iterations were drawn and dimensioned, using monotonicity and constraint activity information from the original model. Specifically, they were drawn in a manner aimed at as

small an outer diameter, as large a spring force, as a large and acceleration stroke (z_{acc}), and as low a steel base (l_{t2}) as possible. Given that we know which constraints affect these and how they are affected by the design changes, it turned out to be relatively straightforward to model seemingly feasible proportional designs of new configurations. The same parameter values were used in CAD as in the original optimization model. Meanwhile, the API mass was kept comparable to that of the original SOMA device.

Different approaches were taken to evaluate how well these dimensioned realisations of the redesigned configurations meet the design objectives. The centre of mass was evaluated using a mass distribution evaluation routine in Creo Parametric 4.0, which took a matter of seconds compared to the quite time-intensive effort involved in building volumetric expressions for all of the changed components and subsequently applying these to evaluate the system centre of mass for each of the redesigns. The device diameter and needle mass were also simply measured using the CAD system. Finally, the impact velocity of each design was evaluated by updating the objective function (Eq. 3.49 in Chapter 3) to reflect the changed spring designs and measuring part masses and acceleration stroke (z_{acc}) in the CAD system. To ensure feasibility, the von Mises stress in each spring and the interface stress in the trigger system was evaluated for each design. These designs and the results of these evaluations are shown in Fig. 6.14 along with the original SOMA design as a reference.

As can be seen, each of the informally optimized redesigns exhibits improved performance w.r.t. every single objective. This indicates that the trade-offs between the design objectives have indeed been reduced. Recall that the reference design is was found to be very close to being Pareto optimal. In fact, all three redesigns dominate a substantial portion of the original Pareto set, given the substantial simultaneous reduction of diameter and increase in impact velocity. All of them have an impact velocity that lies beyond the maximum impact velocity seen in the original Pareto set (28.3 m/s). Further, all are below a diameter of 9.9 mm, which is the largest standard oral capsule size in the market today, making for a more swallowable device. Hence, based on this simple evaluation, we know that all three redesigned configurations at least dominate a region of the original Pareto set.

6.3.2 Level II: Opportunistic Evaluation

If we assume that the Pareto constraints studied in the original ϵMA are still active, we can use them to evaluate how the shape of the Pareto set, or certain vertices and regions of the set, have changed due to the configuration redesign.

Level of Abstraction: Evaluation of how important Pareto optimal activity cases are affected by the configuration design changes. This may reveal regional or global changes to the Pareto-set.
Re-used Knowledge: Monotonicity and constraint activity information from the original analysis, results of ϵMA, and what configuration redesign principles were applied where.
Steps: An opportunistic evaluation process might involve:

1. Update the MOMA to reflect any changes in constraint activity that result from the design changes.
2. Redo the model reductions in order to update the Pareto Optimal Activity cases identified in the previous application of ϵMA, to reflect the design changes.
3. Comparing the resulting trade-off expressions, or Pareto optimal vertices with those studied in the preceding design

Uses: An assessment of whether the Design Improvement Criterion is fulfilled locally or regionally in the new Pareto set. This allows the evaluation of whether the trade-offs between

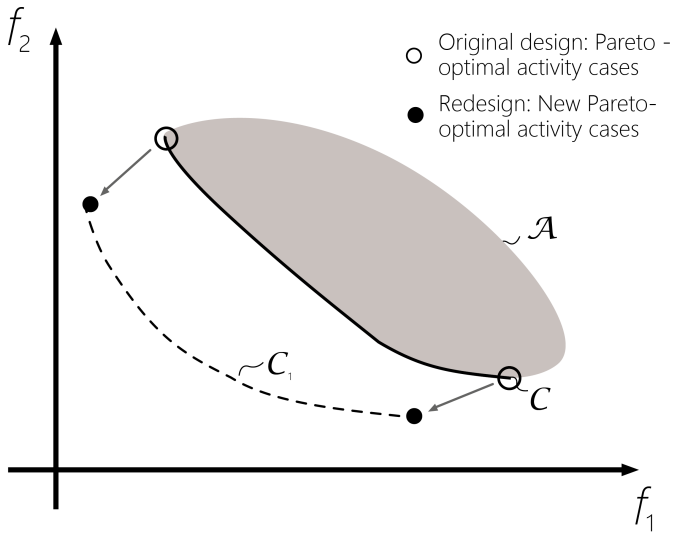


Figure 6.15: Opportunistic Evaluation allows the identification of potential vertices or bi-objective frontiers in the new Pareto set, by updating the Pareto Optimal Activity Cases from the analysis of the preceding design.

certain objective pairs have been reduced, or single objective optima improved..
Potential limitations: If there are new or changed constraint functions, resulting in unforeseen constraint activity, then the design points involved in the updated Pareto-optimal activity cases may be inactive or exist outside the attainable set.

Following the same basic logic as in Informal Evaluation, we can re-use much of the information gained in MOMA and ϵMA to explore how the extrema of the new Pareto set has been affected by the design changes. In cases where it was possible to reduce the original optimization down to a point where explicit expressions describing the relationships between the objectives at the optimum are revealed (e.g. of the form $\tilde{\epsilon_i^*}(\tilde{\mathbf{x}}, \tilde{\epsilon})$), opportunistic evaluation allows the comparison of Pareto frontiers between specific objective pairs. If not, then one might still be able to evaluate how some of the single objective optima have changed as a result of configuration redesign.

Case: Re-use of trade-off and activity knowledge to evaluate SOMA Redesigns

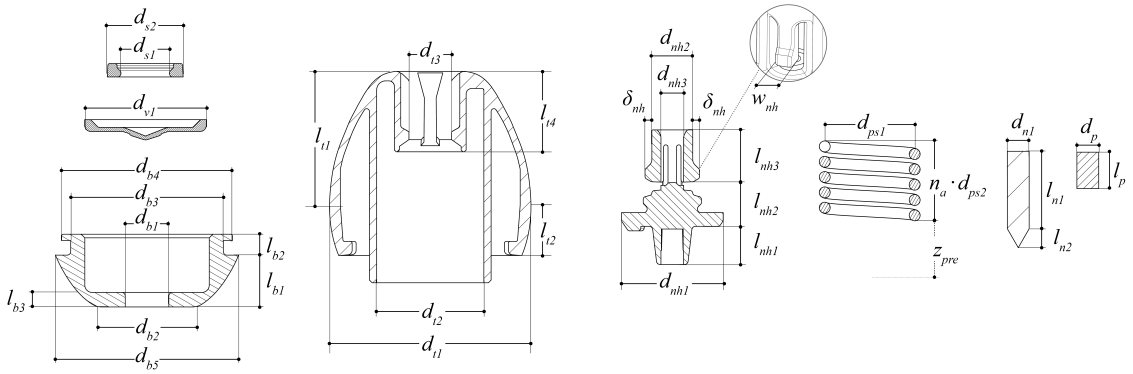


Figure 6.16: *From Chapter 3 and Paper A*: An overview of the key design variables in the SOMA device, included again in this chapter for the sake of readability

In the analysis of the original SOMA device, it was found that a set of radial fit constraints

limited the achievable combination of outer diameter and impact velocity. Looking at the Pareto Optimal activity cases, these constraints result in reveals some of the effects of the changes made to the configuration design.

Before the model was reduced, the radial fit constraint describing the fit between the top and base housings as a function of the location of the split line, had the form:

$$g_1(d_{t1}^-, l_{t1}^-, l_{t2}^+, d_{b5}^+) = d_{b5} - \sqrt{\frac{2(l_{t1}-l_{t2})d_{t1}^2}{l_{t1}} - \frac{(l_{t1}-l_{t2})^2 d_{t1}^2}{l_{t1}^2}} \leq 0 \tag{6.1}$$

The application of MOMA revealed that d_{b5} (shown in Fig. 6.16) is defined at the optimum by a set of active constraints allowing the partial minimization of the model using $d_{b5}^* = d_{t2} + 2R_{ov} + 6R_{wt} + 4R_{cl}$. Essentially, the diameter of the base at the housing split is defined by the inner diameter of the guiding cylinder (d_{t2}), the amount and thickness of the walls that exist between in and the outside of the device, the corresponding assembly clearances, and the overlap in the housing assembly snap.

$$g_1(d_{t1}^-, l_{t2}^+, d_{t2}^+) = d_{t2} + 6R_{wt} + 4R_{cl} + 2R_{ov} - \sqrt{\frac{2(l_{t1}-l_{t2})d_{t1}^2}{l_{t1}} - \frac{(l_{t1}-l_{t2})^2 d_{t1}^2}{l_{t1}^2}} \leq 0 \tag{6.2}$$

Correspondingly, the guiding cylinder is defined at the optimum by the spring needing to fit inside it ($d_{t2}^* = d_{ps1} + d_{ps2} + 2R_{cl}$), while the coiling diameter of the spring is defined by the spring needing to fit around the trigger ($d_{ps1}^* = d_{ps2} + 2\delta_{nh} + d_p + 4R_{wt} + 6R_{cl}$). When combined with the back-substitution bound device diameter objective, this yielded the reduced expression:

$$\begin{aligned} g_1(\tilde{\epsilon_1}^-, l_{t2}^+, d_{ps2}^+, d_p^+, \delta_{nh}^+) = 2d_{ps2} + 2\delta_{nh} + d_p + 10R_{wt} + 12R_{cl} + 2R_{ov} \\ - \sqrt{\frac{2(C_T\tilde{\epsilon_1} - l_{t2})\tilde{\epsilon_1}}{C_T} - \frac{(C_T\tilde{\epsilon_1} - l_{t2})^2}{C_T^2}} \leq 0 \end{aligned} \tag{6.3}$$

The further reductions depend on which constraint is active, with three possible Pareto Optimal activity cases, one of which revealed the single objective optimum of the device size ($\underline{\tilde{\epsilon_1}}$):

$$\overline{d_{ps2}} = 0.5(\tilde{\epsilon_1} - 4\delta_{nh}(d_{ps2}^+; \sigma_c, \sigma_y) - 6.9\text{mm}) \tag{6.4}$$

$$\wedge \underline{\tilde{\epsilon_1}}^* = 2d_{ps2} + d_p + 2\delta_{nh} + 7\text{mm} \tag{6.5}$$

$$\wedge \overline{l_{t2}} = l_{t2}(\tilde{\epsilon_1}^+, d_p^-, d_{ps2}^-, \delta_{nh}^-) \tag{6.6}$$

Comparing the 4th iteration with the original design, we have introduced several changes which affect these Pareto optimal activity cases. By changing the design of the housing snap to a hole in the top housing, we have eliminated $2R_{ov}$ from Eq. 6.2. Correspondingly, the inner diameter of the guiding cylinder has changed to $d_{t2}^* = 2R_{ov} + 2R_{cl} + d_{ps1,1} + d_{ps2}$, where $2R_{ov}$ is contributed by the radial thickness of the sealing o-ring, and $d_{ps1,1}$ is the major coiling diameter of the conical spring. As the trigger arms need to fit through the top of the spring without collision, the major coiling diameter is determined by $d_{ps1,1}^* = 2\delta_{nh} + d_{ps2} + d_p + 2R_{wt} + 4R_{cl}$. Due to the redesign of the trigger, two wall thickness contributions (R_{wt}) and two clearance contributions R_{cl} have been eliminated from the glb of the coiling diameter, compared to the glb of the original spring coil. Inserting these changes into Eq. 6.2 yields:

$$\begin{aligned} g_1(d_{t1}^-, l_{t2}^+, d_{t2}^+) = 2d_{ps2} + 2\delta_{nh} + d_p + 8R_{wt} + 10R_{cl} + 2R_{ov} \\ - \sqrt{\frac{2(l_{t1} - l_{t2})d_{t1}^2}{l_{t1}} - \frac{(l_{t1} - l_{t2})^2 d_{t1}^2}{l_{t1}^2}} \leq 0 \end{aligned} \tag{6.7}$$

Using this to update the Pareto optimal activity cases above yields:

$$\overline{d_{ps2}} = 0.5(\tilde{\epsilon_1} - 4\delta_{nh}(d_{ps2}^{+}; \sigma_c, \sigma_y) - 5.7\text{mm}) \tag{6.8}$$

$$\wedge \underline{\tilde{\epsilon_1}}^{*} = 2d_{ps2} + d_p + 2\delta_{nh} + 5.8\text{mm} \tag{6.9}$$

$$\wedge \overline{l_{t2}} = l_{t2}(\tilde{\epsilon_1}^{+}, d_p^{-}, d_{ps2}^{-}, \delta_{nh}^{-}) \tag{6.10}$$

As can be seen from these expressions, the trade-off between velocity and size has been reduced drastically via. the relationship between $\overline{d_{ps2}}$ and $\tilde{\epsilon_1}$, as has the single objective minimum $\underline{\tilde{\epsilon_1}}$. Inserting the global minimum feasible values of the remaining variables into the above expression, we find that the single objective minimum is $\underline{\tilde{\epsilon_1}}^{*} = 7.7\text{mm}$. Note that while $\overline{l_{t2}}$ remains unchanged w.r.t. which variables it depends on, the reduction in parametric contributions still applies.

We cannot directly calculate the single objective optimum of the impact velocity without calculating the system masses. That said, the minor coiling diameter $d_{ps1,2}$ is defined by the trigger arms needing to fit through the whole spring during assembly, meaning $d_{ps1,2}^{*} = d_{ps2} + 2\delta_{nh} + 2R_{wt} + 3R_{cl}$. Combined with the change to $\overline{d_{ps2}}$ and that $\underline{d_{ps1,1}}$ is smaller than the minimum coiling diameter in the original design, it clear that this redesign achieves a much stiffer and more volumetrically efficient spring coil than the original. Given this information, and given that we have identified the optimal size, we can conclude with a relatively high degree of confidence that the Pareto frontier between size and velocity has been improved substantially. If nothing else, we can hence conclude that the *Partial Design Improvement Criterion* defined in Chapter 5 is fulfilled.

Applying the same approach to redesign iteration 11 reveals even more drastic changes. As the spring now needs fit around the needle (d_n), the activity cases for wire diameter and device size have changed to:

$$\overline{d_{ps2}} = \frac{\tilde{\epsilon_1} - d_n - 4R_{wt} - 4R_{cl}}{2n_a} - R_{cl} \tag{6.11}$$

$$\wedge \underline{\tilde{\epsilon_1}}^{*} = d_n + 2n_a(d_{ps2} + R_{cl}) + 2.2\text{mm} \tag{6.12}$$

$$\tag{6.13}$$

In order to calculate a minimum size, we would hence have to update the yield stress constraint for the spring and the axial fit constraints. This would also give a basic idea of how much spring force and acceleration stroke can be achieved for a given device size. For the sake of brevity, we will not include this evaluation here. This does, however, serve to show that we can actually learn a lot about the relationships that exist at the optimum of the new design without building an entire new optimization model, as long as we are able to account for any new or completely changed constraints. We could, in principle, leverage this knowledge to attempt to identify further design improvements.

6.3.3 Level III: Exhaustive Evaluation

If we want certainty that a redesign has fulfilled the Design Improvement Criterion, the only approach is to build a multiobjective optimization model describing the redesign and identifying its Pareto-set. Luckily, this does not necessarily mean that one needs to build an entirely new model:

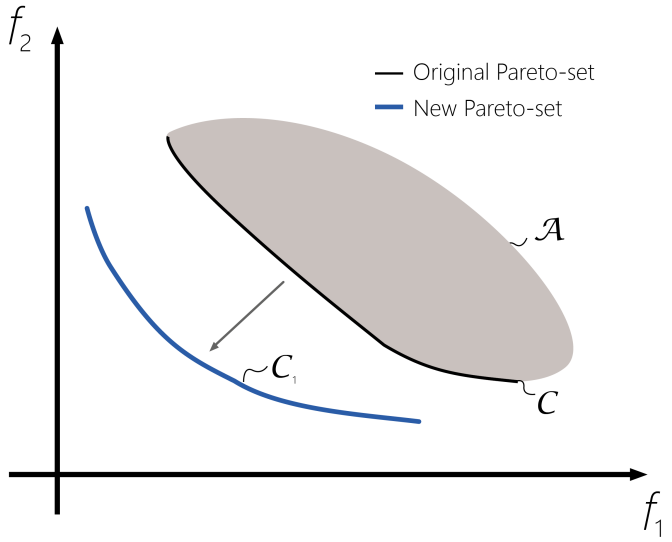


Figure 6.17: Exhaustive evaluation involves building an optimization model to identify the whole Pareto set, allowing an assessment of whether a redesign fulfils the Design Improvement Criterion.

Level of Abstraction: Comparing whole Pareto-sets using a rebuilt optimization model describing the redesign(s) in question.
Re-used Knowledge: Constraint and objective functions that are unaffected by the design changes.
Steps: The exhaustive evaluation process involves:

1. Constructing a new optimization model to account for the configuration design changes or updating the existing one if possible necessary.
2. Verifying the well-boundedness of said model and checking the validity of the expressions used.
3. Numerical solution across a broader range of ϵ values than the original model to explore the boundaries of the new Pareto set.
4. Identifying the meta Pareto-set and consequently evaluating whether the redesign fulfils the design improvement criterion.

Uses: Evaluating whether a configuration redesign is *better* than the original, irrespective of the relative importance of the different design objectives. Hence, this can be used for redesign selection and/or for an additional round of MOMA+ϵMA and subsequent configuration design improvement process.
Potential limitations: If the fidelity of the new optimization model(s) is not the same as the fidelity of the original optimization model - e.g. due to incomplete data or a lack of design maturity - the Pareto-sets are not necessarily comparable. Furthermore, constraints that do not exist in the original design might get overlooked due to a lack of knowledge surrounding the new, less matured configuration design.

The application of some of the redesign strategies will inevitably result in the need for an optimization model that is more distant from the original. If we, for instance, have only relaxed the constraints on harmonious variables through design change, it would be sufficient to rebuild the affected constraint functions. Manipulating trade-off variables meanwhile will always result in changed objective functions. Ultimately, the amount of re-use depends entirely on the types of objective and constraint functions involved. In the case of characteristics such as mass distribution (c.f. the SOMA device), even the smallest configuration design changes can have a drastic impact from a mathematical perspective. Nevertheless, the benefit in spending this effort on evaluating a redesign is that it allows us to repeat the application of MOMA, ϵMA, and the configuration redesign principles, to potentially reach new insights even more (and even better) configuration designs.

Case: Exhaustively Evaluating a SOMA Redesign

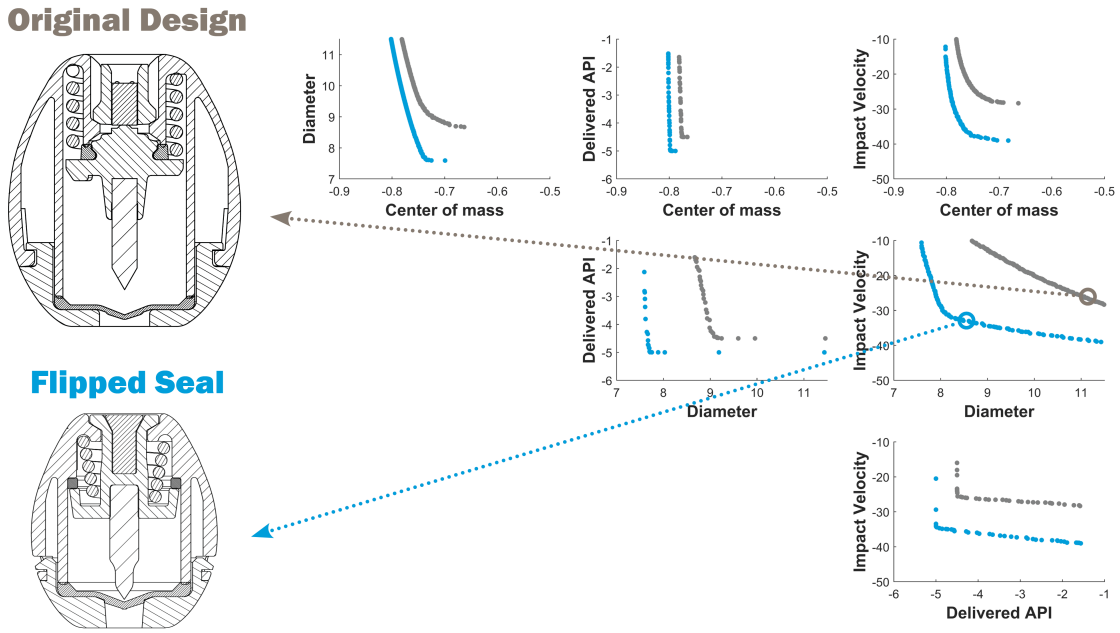


Figure 6.18: *From Paper B*: A head-to-head comparison of the original configuration (grey) against the 2nd redesign (blue), the *Flipped Seal*. The 4D Pareto-set is shown with a 2D projection showing the bi-objective Pareto frontiers between each objective pair, which shows how the redesign is a clear improvement on all accounts. The relative size of the two designs is to scale.

Exhaustive evaluation of every single redesign iteration would have been more time-consuming than what is worthwhile in practice for the SOMA case. Especially given the amount of effort involved in deriving an accurate analytical expression for the mass distribution of the system.

The two other levels of evaluation have already revealed valuable information about some of the redesigns. Furthermore, the theorems and corollaries from Chapter 5 already support that most of the redesigns must fulfil the design improvement criterion, given that most of them do not introduce new trade-off variables or more limiting constraints. In a product development context, we might hence be satisfied that we have indeed identified configuration redesigns that fulfil the design improvement criterion. Yet, to demonstrate the validity of the systematic redesign procedure, we can exhaustively evaluate a single redesign iteration and compare it against the original. This comparison is also described in Paper B, albeit with a lower level of detail. The 4^{th} redesign iteration - the *Flipped Seal* configuration - was selected

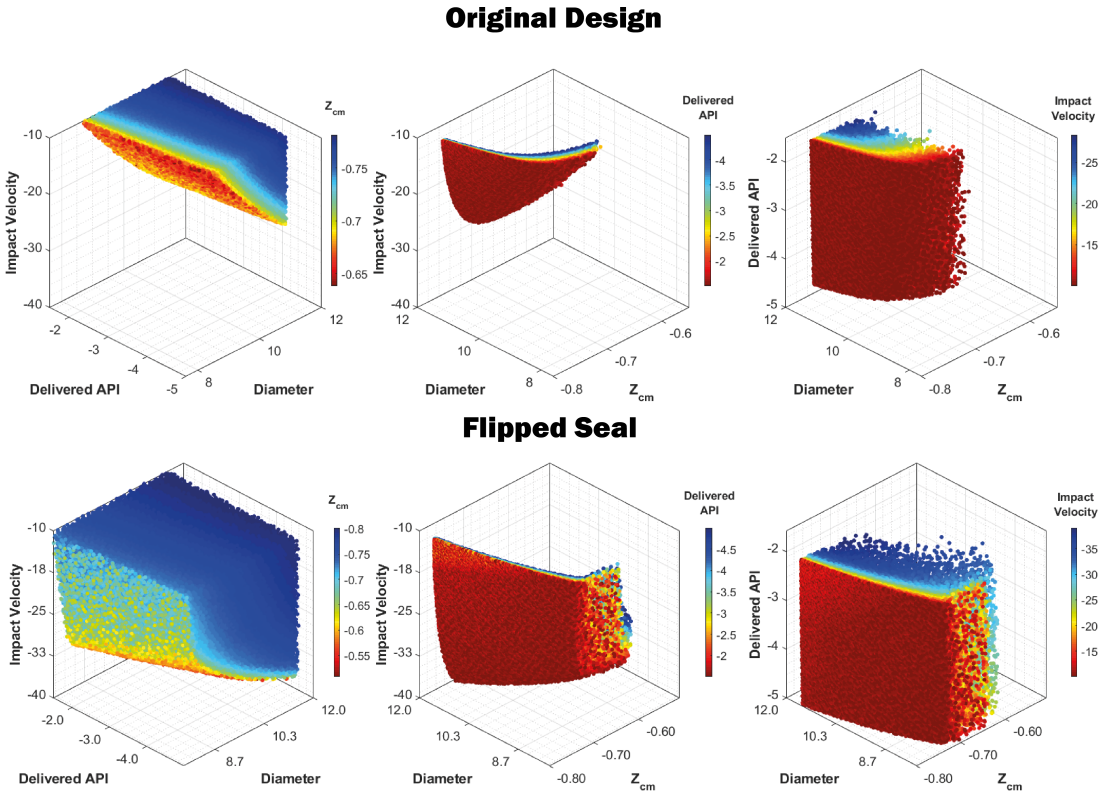


Figure 6.19: Different projections of the the 4D Pareto-set for the original SOMA design and the 4^{th} redesign iteration (the *Flipped Seal* configuration). To make the comparison clear, the plot axis limits for the original SOMA design have been set to match the axis limits of the redesigned SOMA.

for this comparison. It was selected based on three factors:

1. It allowed substantial re-use of expressions from the original optimization model.
2. Yet, it still involves influential changes of the configuration design, thereby demonstrating how large an influence trade-off variables, Pareto constraints, and restrictively bounded harmonious variables can have upon the achievable performance.
3. Out of all the redesigns, the SOMA project team was most interested in seeing how it compared, as making bigger changes to the configuration design might introduce more uncertainty compared to a more well-known system. The flipped seal exhibited no changes in working principles and quite resembled the original design.

Hence, a new optimization model was built. It contained new constraint functions to reflect the new part fits, new and updated expressions for mass distribution to reflect the changes in geometry, and changes in spring equations to reflect the conical shape. The model structure, governing equations, and level of detail remained unchanged. Of particular note are the radial fit constraints describing the fit of the trigger arms within the spring. Strictly speaking, the trigger arms only need to avoid collision with the spring during the injection. As the trigger arms can flex inward during assembly, the diameter of the lower portion of the spring coil is hence only constrained by the outer diameter of the trigger arms in their deflected state and not their free state. This allows for a more conical shape than otherwise.

This model was run with 200.000 iterations, with $\epsilon_{\mathbf{L}} = [7\text{mm}; 1.5\text{mg}; 10\text{m/s}]$ and $\epsilon_{\mathbf{U}} =$

[11.5mm; 5mg; 45m/s]. The results in Figs. 6.18-6.19 show the new Pareto set lying beyond the original one. For the union of the Pareto sets, $\mathcal{C}_{\mathcal{U}} = \mathcal{C}_0 \cup \mathcal{C}_2$ the meta Pareto-set was found to only consist of solutions from the 2^{nd} redesign, i.e., $\breve{\mathcal{C}} = \mathcal{C}_2$, and the single-objective optima of self-orientation has been improved by 2.63%, the size by 12.41 %, API payload by 11.11%, and velocity by 37.68%. We can thus conclude that the redesign is, in fact, a design improvement, as it meets the criteria in Definition 5. For the subsequent redesigns, it is clear that the achievable combination of impact velocity and self-orientation is improved even further, as the design changes are aimed at increasing the load-bearing area in the trigger system and shifting the centre of mass downward while increasing the acceleration stroke. The informal and opportunistic evaluations also support this.

6.3.4 Result Implications

In conclusion, the different approaches to redesign evaluation all show that the redesign procedure was successful for the SOMA device. At every level of abstraction, design points have been identified that improve upon the original Pareto set. In practice, most designers would probably rely on informal evaluation alone, given the minimal effort compared to analysis. This might then have been used to select a single redesign for exhaustive analysis.

That said, the opportunistic evaluation is, if anything, even less time-consuming. Yet, there is a risk that one might overlook new important constraints, which only become apparent when the design is drawn in a 3D environment. Furthermore, the self-orientation objective was not considered, but one could argue that there are clear indications that the two designs fulfil the Design Improvement Criterion, solely based on the theorems from Chapter 5 and the results of the opportunistic evaluation. If anything, the design changes made have likely improved the self-orientation, especially in the case of iteration 11, making it less critical to attempt to evaluate the effects of the design changes.

Still, had we not conducted the exhaustive evaluation, we could not have known for sure that the assessed redesigns are an improvement on the original, irrespective of the relative weighting of the objectives. As the multiobjective optimization of the 4th redesign iteration has revealed, the analysis and redesign methodology has allowed us to identify a design that can unequivocally be stated to be an improvement, w.r.t. to the objectives included in the model. Hence we can - at least for this case - conclude that the Configuration Redesign Principles are valid and that the overall Configuration Design Improvement procedure has value in systematically identifying and mitigating the limitations of a design. Furthermore, these results also confirm the practical relevance of the theorems and corollaries derived in Chapter 5.

Finally, as discussed in Example 2 in Section 6.1, the redesign process has actually led to a redesign that relies on different working principles and a different distribution of functionality amongst the parts in the system. Some might argue that this constitutes conceptual change. If that is the case, we have used a proportional optimization model describing a given configuration to reveal potential improvements to the overall concept. This has value in and of itself, as such situations may help identify potential opportunities in regressing back to concept design. This might otherwise be difficult to defend in practice, given the increased development cost involved in scrapping a relatively mature design.

7 Discussion

This chapter presents a discussion of the validity and limitations of the outcomes of this research. First, we revisit the Validation Square approach by Pedersen et al. [47] (introduced in chapter 2) to frame a discussion of the validity of this research. This is followed by a discussion of limitations such as the methodology's reliance on monotonicity and model reduction, the opportunistic nature of the redesign method, and the contextual nature of the research. Much of the content on the limitations of the developed methods stem from papers Papers A and B.

7.1 Result Verification and Validation

To systematically discuss the validity of the results, we return to the Validation Square method introduced in chapter 2, which was first suggested by Pedersen et al. [47]. In the following, the four aspects of validity outlined by the Validation square are discussed:

7.1.1 Theoretical Structural Validity

Is the underlying theory behind the design method well accepted, and is the method consistent? Are the constituent elements of a method, and the method as a whole internally consistent?

The developed methodological framework consists of the Pareto set Dependency Analysis method, the approach to Systematic Configuration Design Improvement, and their implications and integration in the design process described in Chapter 6. These have been developed with the aim of supporting design engineers in the understanding and systematic avoidance/reduction of trade-offs between design objectives through configuration design change and potentially also avoiding these through design synthesis.

'The focus in all of this work was to ensure a basic degree of mathematical rigour, given that the analysis and design methods are built around understanding and manipulating the relationships that create the Pareto set. This was enabled by the use and further development of monotonicity analysis (MA) methods. As an underlying theory upon which this work is based, MA is well accepted in the design optimization community, albeit perhaps being perceived by some as a somewhat niche and analysis intensive method. The original value proposition of MA was to allow systematic model verification and reduction, thereby reducing computational cost and the risk of non-real results and issues with convergence. In this research, MA has been adapted towards a quite different purpose - as a means of dependency analysis to support the identification of the root causes of trade-offs. The original theorems and principles of MA all have associated proofs, and this research has hence striven to underpin Pareto set Dependency Analysis with theorems and proofs that are consistent with MA and with optimization methods as a whole. As there are no apparent inconsistencies between the developments and existing design optimization theory, the methodological framework developed in this thesis can be argued to be externally consistent.

The systematic configuration design improvement methodology builds directly upon the outputs of Pareto set dependency analysis. The developments in chapter 6 are of a more practical and qualitative nature, but these simply arise as a consequence of the preceding developments. As such, they are inherently internally consistent as well.

7.1.2 Empirical Structural Validity

Are the example problems used to verify the method appropriate?

As the SOMA device case presents a configuration design task that involves numerous trade-offs at the early stage of mechanical design, it generally meets the scope and purpose of the methodological framework developed in this thesis. As a design problem, it involves non-standard components, a high-risk application, and a high volume manufacturing case. Thus, the design engineers involved in the development of the SOMA device have the design freedom and development resources to actually make and test the types of changes that the methodology results in.

Further, rather than being an artificial test case that might have been selected or designed to mask potential limitations in the methodology, the SOMA device is a real-world case from industry. The PhD fellow had no influence on or involvement in the development of the SOMA device prior to the action research, meaning that the methods were applied and the results were reached at a point where the knowledge gained was not otherwise available.

That said, the SOMA is free of many of the characteristics that would complicate the use of the methods. There are no discrete variables, nor are there any non-monotonic variables in the initial model. This is admittedly by design, as efforts were made in the model construction process to preserve monotonicity to the extent possible. Examples of such include the simplifications made to the design objectives (i.e. optimizing self-orientation by lowering the centre of mass) and the selection of the bound-objective formulation. Similarly, the SOMA optimization model is purely algebraic and is a relatively small design problem with a limited number of parts, objectives, and constraints.

As such, one might observe challenges or limitations to the methodology in situations that require more sophisticated models. However, the potential challenges ultimately come down to the question of analysis effort rather than general validity. Larger problems with, e.g. non-monotonic properties, will inevitably require more algebraic manipulations and numerical analysis in order to reach the same types of insights as seen in the SOMA case. That said, such systems will likely also require more effort and iteration from the designers perspective in order to achieve design improvement, meaning that the added analysis effort may still be worthwhile.

7.1.3 Empirical Performance Validity

Are the results of the application of the method useful in the studied case(s) - does the method meet its initial purpose?

In this regard, we return to the two questions posed in chapter 2:

1. Does the application of the developed methodologies result in designs that have measurably improved performance and reduced trade-off?
2. Do the optimization models used to identify redesigns, and subsequently, compare them, yield repeatable results, real design variable values (e.g. non-negative, zero, or infinite dimensions) and do they sufficiently approximate the behaviour of the real system?

As to the first aspect, the application of the analysis and subsequent redesign methodology to the SOMA case resulted in measurably improved configuration designs, although this can only be stated w.r.t the modelled objectives. Although each redesign was not evaluated exhaustively, the combination of the exhaustive evaluation of the "Flipped seal" redesign, and the explorative and opportunistic evaluation of select designs, demonstrates two things. First, that the Pareto set dependency analysis methodology does indeed reveal relationships

in the SOMA device that cause the existence, shape, and location of the Pareto set, as many of these relationships were removed or affected by the redesigns. Secondly, the redesign methodology provides an approach to mitigating some of these relationships systematically, with the "Flipped seal" design being demonstrated to fulfil the *Design Improvement* criterion defined in section 5.1. It is likely that the subsequent redesigns likely also fulfil the same criterion, given the results of the explorative and opportunistic redesign evaluations of other redesigns.

As to the second aspect, the optimization model used in the SOMA case demonstrated sufficient fidelity to drive decision making in redesign. Specifically, it was found to deviate less than 1% compared to existing numerical models. At the same time, the impact velocity objective yielded results that, to a large extent, matched observations made by the SOMA project in initial prototyping and testing. The optimization results also yielded designs that were very similar in proportion to the reference design, pointing to the fact that the SOMA project had already reached a Pareto adjacent design through iterative redesign rather than formal analysis. In part, this verifies that the optimization model developed in chapter 3 yields real results, given that we would expect that the model would yield vastly different and perhaps inconsistent results if it were far removed from reality.

As such, the methodological developments meet their initial purpose; the MOMA approach correctly predicted which objectives were in trade-off and revealed their root causes when combined with ϵMA, while the subsequent redesigns successfully reduced the trade-offs while improving optimality overall. The informal redesign evaluation also revealed that the fidelity of the optimization model was sufficient to support decision making in redesign. In other words, the SOMA case demonstrates that the methodological results of the research presented in this thesis at least have contextual validity. The redesign method also yielded real-world value, given that the redesign exploration resulted in several patent applications.

7.1.4 Theoretical Performance Validity

Is the usefulness of the support observed in the empirical case study generalisable beyond the case?

The generalisability of this research is to an extent guaranteed by the rigorous underpinnings of the analysis and redesign methods. Due to the formal developments (theorems and proofs) upon which the methods are based, they can be applied to any design problem that can be described through algebraic models and meta-models, which exhibit monotonic characteristics. Thus, the methodology is not limited to specific types of products or design objectives. Rather, it allows the analysis and potential mitigation of any trade-off that can be described through explicit models. Given that design engineers often rely on simple models to inform decision making at an early stage of development [99, 116], the methodological framework is potentially useful in the early embodiment/configuration design of most mechanical systems. Looking at the design of the SOMA, it is also worth noting that its constitute elements (springs, ratchets, snap fits, linear guides) and design objectives resemble those seen in a wide range of mechanical systems. Not to mention that mass distribution problems are seen in a multitude of design contexts.

Furthermore, a lot of design problems exhibit monotonic characteristics [58], especially when it comes to early decisions such as the configuration of parts in a system, which largely involves monotonic geometric relationships. Even in the presence of non-monotonicity, the foundational theorems and proofs of this research still apply, merely on a *regional* level in the design space, rather than globally. As discussed, this does detract from the practical applicability of the methodology, given that the effort in analysis grows in these situations. That said, this does not detract from its validity; regional monotonicity analysis is well accepted,

albeit perhaps seen by some as being a time-consuming approach compared to simply relying on computation. For this exact reason, the use of numerical data to support Pareto set dependency analysis has also been covered in chapter 4.

Finally, it is worth remembering that the purpose of this research was never to develop a computationally efficient approach to optimization or design analysis. Rather, the purpose was, and is, to aid designers in understanding the causes of the optimal set and how the set can be improved by changing the design itself (and, therefore, the optimization problem). One could say that design optimization is an approach to identifying the *best* solution to the design problem that is. In contrast, the methods developed in this thesis are an approach to identifying a *better optimal design problem* in and of itself. Given that this is largely tacit knowledge today [12, 31, 33], providing a systematic approach to gaining this knowledge and implementing it in configuration design thus brings *good* mechanical design closer to being a science rather than being a craft.

7.2 Limitations

Today, successful configuration/embodiment design requires a mix of creativity, systematic analysis, qualitative reasoning, and engineering judgement [6, 12]. This phase is characterised by both aleatory and epistemic uncertainty [10], meaning that the experience of the designer plays a vital role [33, 35]. As discussed in chapter 5, other works have sought to prescribe principles for achieving "good" design", but these are either context-specific or quite general in nature.

Hence, attempting to codify a systematic approach to identifying trade-offs and their root causes at this stage, and using the knowledge to identify design improvements, will always involve limitations. To misquote George Box, the author would argue that all design methods are wrong, but some are useful. In the following, the limitations of this research and the limitations of the usefulness of the developed methodological framework will be discussed.

7.2.1 Limitations of the Research Approach and Application

Most prescient among the limitations of the design of this research is the fact that the project was carried out in a single case company. This leaves the risk of contextual observations affecting the direction of the research and biasing results. Furthermore, one might also question whether the topic itself is of relevance to any product developing organisation or whether the discussed challenges with trade-off management are contextual to the case company.

These limitations are not unique to this research project; the study of design is fraught with challenges ultimately stemming from the lack of controllable laboratory conditions. Conducting in-depth research of embodiment design practices in parallel in multiple organisations operating in different contexts would simply not have been possible for the purposes of this research. If nothing else, such work would pose substantial IP-related concerns.

For the same reason, the focus in most of the research activities presented in this thesis has been to shy away from developing context-specific heuristics and methods, focusing on generality. A key component of this has been to ensure a degree of rigour in the analysis and redesign methods, allowing application independent of context-specific design objectives, trade-offs, or product embodiments. The methods developed in chapters 4-6 are not aimed toward "*Trade-off Management in Medical device design*"; they can be applied in the design of any mechanical system which exhibits (regionally) monotonic behaviour and is developed with multiple objectives in mind. Furthermore, findings in existing research also support that the topic itself is of general interest. Trade-off identification and avoidance are a key to how expert designers solve problems[33] and reason about design [124], and the embodiment

design itself ultimately defines the proportional optimum [12] and the trade-offs the designer will need to manage [12, 13, 20].

That said, the studies presented in this thesis still comprise a limited data set. This especially influences the Descriptive Study II, in that the evaluation of the developed analysis and design methods have been limited to a single phase in a single product development project - the SOMA device. While this evaluation revealed that the methods fulfil their purpose, it is not a given that they do so in every context. That said, the Guidelines of Ideal Design, which emerge out of the development from the prior chapters, at least indicate that many of the same modes of design change as suggested by the Configuration Redesign Principles are apparent in the decisions made on the level of conceptual and configuration design in the FlexTouch device.

Blessing and Chakrabarti [4] argued that the *usability* of new methods for engineering design is a key success criterion in design research. However, all of the analysis and redesign work presented in this thesis was conducted by the PhD fellow. Hence, the broader usability of the methodology in the hands of other design engineers remains to be seen. The methodological developments might have limitations that reveal themselves if/when they are applied by design engineers that work in different contexts and with different interests and backgrounds than the author.

Finally, the question of the industrial impact of this PhD remains open. Given that the development of the SOMA is very much ongoing, the ultimate influence of this research remains to be seen. Whether or not the work ultimately has any impact in practice, e.g. on the overall lead time of the project, the performance and robustness of the end product or its in-market success, is not simple or necessarily possible to evaluate. This would not only require a longer research horizon than that of a three year PhD project but also the evaluation of two product development projects running in parallel under the same working conditions. Hence, smaller experiments evaluating the success and usability of the methodology - e.g. with students or fictitious test cases - remains a part of further work.

7.2.2 Limitations of the Methodological Developments

On the Analysis Methods (*in part covered in Paper A*)

The developments to monotonicity analysis (MA), gathered in the Pareto set dependency analysis methodology (i.e. MOMA and ϵMA performed in sequence), are just as MA itself, opportunistic in nature. While they involve rigorous analysis and model reduction steps, they can only be applied when the monotonicity of the design problem can be ascertained. This opportunistic nature reveals some key limitations. Firstly, not all design problems are monotonic or even differentiable. This might be dealt with using techniques for local [69], and regional MA [68] if the expressions are regionally differentiable. This comes at the cost of increased analysis effort, which might be offset using sampling-based computational experiments (e.g., design space exploration or DoE) to reveal regional properties in non-monotonic or non-algebraic problems.

Secondly, MA mostly relies on algebraic manipulations, and some design problems are too complex to be expressed algebraically. Yet, that certain aspects of a design's behaviour can only be expressed numerically does not necessarily imply a lack of monotonicity, e.g. as is often the case with stiffness and deflection. In such situations, implicit MA [12] procedures and monotonicity analysis of meta-models might be used. This would reveal the variables and constraints that cause trade-offs, albeit without the derivation of explicit expressions of the relationships that exist in the Pareto set.

It is also well accepted that purely algebraic models can play a substantial role in practice

[116] in both conceptual and configuration design. These phases are often characterized by a lack of sophisticated quantitative models to support decision making due to requirement uncertainty and the modelling effort involved, compared to how quickly and often the design changes [137]. Configuration design also often involves the combination and arrangement of well-known types of parts and modules, which might be described algebraically, for example, as seen in the machine elements, engines, hydraulics, and thermal systems.

That said, the validity of the decisions made based on the analysis methods will inevitably be dependent on the fidelity of the model being used. This may lead to configuration redesigns that actually worsen the performance of the end product or solve issues that only exist in an overly pessimistic model. Luckily, many of the conclusions one may identify through the application of Pareto set Dependency Analysis, stem from the monotonicities of the design problem and are not necessarily dependent on the optimization model identifying the true optimum. Increasing the fidelity of a model does not necessarily affect monotonicity or constraint activity. This ultimately depends on whether the inclusion of factors such as nonlinear phenomena introduces new dependencies or results in a change in the dimensionality of the constraint functions to a point where previously active constraints become dominated or vice versa.

Hence, constraint activity and the resulting sequence of model reductions could remain unchanged as a development project moves from the initial simple analytical expression to the more sophisticated late-stage numerical models that take a slew of non-linearities into account. While this may result in more accurate identification of optima, many of the relationships that exist at the optimum (dependencies and active constraints) could remain the same. This means that despite the early stages of development being characterised by a large degree of uncertainty, and a lack of knowledge and resources to build high-fidelity analysis models, MOMA could still be applied very early to simple models and provide the analyst/designer with useful insights. Especially when there is confidence surrounding the general validity of the monotonicity information.

In regards to analysis effort, it is worth noting that the effort in Pareto set dependency analysis is proportional to the number of objectives, constraints, and variables in the problem. This effort is amplified by non-monotonicity and by regionally active constraints. Thus the bookkeeping and algebraic effort required to reduce a multiobjective model systematically may be prohibitive if the problem is large. This might limit the applicability of the analysis methodology, given that most mechanical design problems involve larger systems than the SOMA device.

Here, the use of symbolic solvers can help reduce the effort in back-substitution and model reduction (e.g. MATLAB symbolic, Maple, PTC MathCAD). In that regard, quite some work was done (with some success) on automating MA in the 1980s [69, 71]. In this context, there is potential in attempting to improve automation of MA (and thus MOMA and ϵMA) by leveraging the achievements made in computational techniques such as machine learning, AI, and data analysis and clustering, since MA methods were last in vogue. Given the advances in meta-modelling since then, it is also not unlikely that more complicated non-algebraic models might be analysed using the methods described in this thesis. Hence the (partial) automation of MOMA and ϵMA is possible future work.

Ultimately, the value of the analysis methodology comes down to the cost involved in analysis vs the expected benefit in discovering better configurations. As discussed in chapters 1 and 3, trade-off knowledge and decision-making are largely experience-driven in early-stage design. Finger and Dixon [135] highlighted the dearth of quantitative design analysis and evaluation methods for the early stages, especially those which allow multiobjective analysis and support

the identification of alternative configurations and concepts. The presented methodology addresses some of these unmet needs.

When the cost vs benefit in performing such analyses is positive, the methodology might be used to target iterative configuration redesign, guide morphological studies to identify alternative solutions, or simply to explain the results of an optimization model from a design perspective. For small, highly interdependent systems such as the SOMA device, the value in discovering the non-obvious influence of certain variables and constraints on trade-offs more than justifies the analysis effort. For a larger system, the methodology can be worthwhile if the system is obviously monotonic or if the optimization model is constructed at an architectural level of abstraction that limits the number of design variables and expressions to analyze, focusing on the relative dimensioning and arrangement of important modules and parts.

On the Redesign Methods (*in part covered in Paper B*)

With the use of the Systematic Configuration Design Improvement method and the developments in chapter 6, a degree of rigour is brought into the iterative design process, allowing the designer to utilize optimization to qualify the introduction of design changes. Thus, it is not singularly a configuration design or design optimization methodology - it is both. As seen with the SOMA device, the actual changes required to achieve an improved Pareto set can be relatively simple. Inversion, change of working direction, and changes to how the components fit together. Still, the impact on the Pareto set is substantial, as seen in the dominance of the *Flipped Seal* redesign over the original design.

Yet, as with the analysis methodology, this does have limitations. Perhaps the most obvious is that the analysis involved would seem onerous to most designers. Systematic analysis does not necessarily fit into organisational cultures where a *design thinking* mindset leads to a bigger emphasis on iteration and creativity than more classical engineering considerations. Again, a cost-benefit mindset comes in: if the benefit gained through redesign is accrued over a production volume counted in hundreds of thousands, millions or billions (as is the potential with SOMA), then the cost of analysis becomes almost meaningless.

The methodology's success also depends entirely on whether all objectives and constraints of importance have been taken into account in the model. Therefore, the importance of a restrained approach to applying the redesign principles cannot be understated. As exemplified in chapter 6, constraints and objectives that are either tacit or simply not included in analysis might lead to design changes that introduce new trade-offs, worsen the product offering, or compromises key functionality. An addition to this challenge is that embodiment design involves a concretisation process [6, 20], where necessary sub-functionality and features are gradually synthesised and integrated into the overall system. Just as with unmodelled constraints, these might need to be considered somehow in the analysis and redesign process to avoid introducing design changes that hinder the introduction of new functionality.

Finally, the methodological framework presented in this thesis can only lead the designer in a certain direction; it is not inherently generative in nature. Experienced designers might hence legitimately argue that they are capable of identifying trade-offs and mitigating them without the potentially arduous analysis efforts. As experience is contextual and time-consuming (i.e. costly for employers) to gather, the systematisation of trade-off analysis and mitigation permitted by the developments in this thesis is both of value to novice designers and to experts in situations where they face new design tasks and challenges.

On the Perspectives for the Design Process and for Synthesis

The methods developed in chapters 4 and 5 inherently rely on the existence of a system that has already been embodied. Hence, they support the transition between what Pahl & Beitz [6]

referred to as the *preliminary* and *definitive layouts*. To avoid the risk of spending substantial effort on analysis and mitigation effort might have been avoided in synthesis, the developments in Chapter 6 touch upon the question of how to apply some of the same underlying rationales behind the Configuration Redesign Principles in the context of synthesis.

As with the Configuration Redesign Principles, the Guidelines of Ideal Design Synthesis do not change the fact that the ideation involved in the synthesis of the preliminary layout of a mechanical system, and the subsequent identification of design improvements, still relies heavily on the creativity and ability of the designer. To a large extent, the developments in Chapter 5 onward only truly help the designer realise what to strive for in design and what types of changes can be introduced when the initial design involves a dissatisfactory optimal set or seems to be too far removed from the ideal design during the even earlier stages of product development.

The guidelines also rely on the Conjectures of Ideal Design, which may very well be disproven, or have important exceptions not identified in this research. Correspondingly, the guidelines are not exhaustive, nor are they on a level of specificity that ensures that it would always be obvious to the designer how to get close to the ideal design through synthesis. This will always be a challenge in the prescription of any heuristic. Yet, as they have been developed towards being consistent with the methods developed in the preceding chapters, they at least relate to a more rigorous foundation. That said, this aim of consistency may also have biased the development of the guidelines, resulting in an incomplete set of prescriptions or recommendations that directly conflict with other frameworks or perspectives of importance to design synthesis.

The Interactive Trade-off Analysis and Mitigation procedure essentially gathers all the developments of this thesis into a coherent process that supports the progression from the synthesis of the first embodiment design to the completion of the definitive one. A key limitation of this process is its over-reliance on analysis. In practice, design objectives and constraints are not static. They develop and change along with the design, as the product development project gradually eliminates uncertainties. Meanwhile, the outcomes of Pareto set Dependency Analysis and Systematic Configuration Design Improvement are entirely contextual to the initial optimization model and the underlying decisions and prioritisation it represents. Hence, there is a risk of spending a substantial amount of effort on finding the ideal configuration, only to realise the work is based on the wrong premise. Ultimately, this comes down to a key challenge in design science; the epistemic and aleatory uncertainty involved in product development. In any attempt to systematize the early design process, one can only use the knowledge at hand. Hence, the interactive nature of the design process prescribed by Section 6.2 is both a strength and a weakness, as its success, value, and required effort is dependent on the extent to which all the objectives and constraints of relevance are considered systematically. If not, one would end up performing more iterations to reach the same result, increasing the *cost* in applying the methodology.

Finally, the novel evaluation methods developed in Section 6.3. - i.e. the informal and opportunistic redesign evaluation procedures- have only been possible due to the development of MOMA and ϵMA. Given the reliance on the designer re-using information from the initial analysis and updating the results to reflect the changes made in redesign, it can be difficult to assess the validity of the results of evaluation. The changes made may - beyond the designer's knowledge - have created new and problematic dependencies and constraints, thereby invalidating the results of evaluation. Yet, this is also exactly why the developments are *informal* and *opportunistic* in nature - they will not always provide valid insights.

8 Conclusion

This chapter concludes this thesis by revisiting the research questions and hypothesis of the PhD project, describes the core contributions of the research, and describes potential avenues of further research.

8.1 Findings

The research conducted during this PhD was guided by three research questions. Besides the slight change to RQ2 mentioned in Chapter 4, these have largely remained constant throughout the project period. As covered in Chapter 2, these questions have been addressed through different work packages involving several studies. Theeir results have been demonstrated in Chapters 3-6 and submitted for publication in Papers A and B. To conclude this research, we revisit these research questions to summarise the findings of this PhD and subsequently test the original hypothesis. Given that the answers to these research questions involve new methodological developments, the research merely provides *an* answer to the questions, and not necessarily *the* answer. This is inevitable in that design is inherently a human activity, meaning there will always be alternative approaches/methods that one might apply in attempting to reach the same result (i.e. the systematic identification, understanding, and mitigation of trade-offs).

8.1.1 Answers to the Research Questions

Research Question 1

How can the trade-offs between design objectives be identified in the concept and embodiment design phase?

Trade-offs are omnipresent in design and are largely caused by decisions made during the early stages of design, as is well established in existing literature. Given their potential influence on the performance of the end-product, identifying trade-offs early in the design process is valuable, as some of them might be reduced or eliminated through redesign and targeted decision making. The issue today is that the designer does not necessarily have access to this knowledge early on in the design process, which is commonly referred to as the *design process paradox* [10].

Trade-offs are only possible in the presence of dependencies between the design objectives and numerous methodological frameworks exist, such as Axiomatic Design [13], TRIZ [16], the Design Structure Matrix [7], and the Quality-Function-Deployment [93]. These all allow some form of assessment of the dependencies between different design goals, and have applications in the early stages of design.

Yet, as found in this thesis, qualitative methods have their limitations in the context of trade-off identification. In an initial study, this project explored how a qualitative method (the Contradiction Index [17]) could be used to analyse trade-offs. Here, it was found that the influence of constraints was difficult to account for using qualitative analysis as one is either reliant on prior knowledge/experience or guesswork in figuring out which constraints are active. In the context of design, active constraints are basically requirements that relate to the feasibility of the end product and determine the achievable performance of a product.

The underlying challenge is that a trade-off situation (strictly speaking) implies Pareto optimality - i.e. that no further improvement to a single objective is achievable without detriment to another objective. If a design is not Pareto optimal, then further improvement to one objective is possible without detriment to another, meaning a trade-off situation has not yet been *reached*. The Pareto set itself exists at the boundary of what is feasible, meaning it is in part defined by active design constraints. As such, the trade-offs that affect the design of a mechanical system are defined by relationships that might be unique to the optimum. Thus, we cannot truly identify all the trade-offs in a given design unless we have already identified the Pareto set. Alteratnively, we would need to have exact knowledge of constraint activity and all the variables that are shared between the objectives. Identifying *all* the trade-offs in a design, therefore, requires the application of more quantitatively founded methods.

The challenge in applying quantitative methods at a very early stage of design is that they often come with a substantial amount of effort compared to more qualitative approaches. This effort may very well be wasted if spent too early in the design process, as the design itself and the objectives it is developed towards can change drastically as the product development project progresses. This likely changes the design problem to a point where any preceding analysis model is rendered obsolete or irrelevant. As a result, this research progressed towards exploring whether quantitative analysis could be performed in a less time-consuming manner or in a manner where the learnings from the analysis are still of value, despite a changing design. A second initial study involving a sampling-based approach was conducted, but again this revealed challenges w.r.t. design constraints. Hence, the approach yielded little knowledge beyond providing a snapshot of the trade-offs affecting the design problem.

As a result, the research progressed onward to exploring the combined application of monotonicity analysis (MA) and design optimization in the hope that this might reveal important information about the nature of the design problem, which might be leveraged throughout the design process. The basic logic behind MA is to utilise monotonicity information to systematically identify active constraints, allowing the *partial minimization* of the optimization problem. This has two benefits in the context of trade-off identification and analysis. Firstly, MA actually reveals which objectives are in trade-off through shared design variables. Secondly, in identifying which constraints are active, partial minimisation essentially reveals discontinuous dependencies between the objectives that are unique to the optimum. In other words, MA not only has the potential to help identify trade-offs; it may also reveal the design variables (global dependencies) and constraints (local/regional dependencies) that cause them. As a consequence of this, RQ2 was updated to the form shown below.

Research Question 2*

How can conceptual or configuration design limitations reflected in the Pareto set be identified rigorously? In particular, what specific design dependencies and constraints cause trade-offs?

While MA had potential for the purposes of this research, no existing theoretical developments allowed the systematic reduction of multiobjective problems or the rigorous identification of the dependencies that exist in the Pareto set. On the contrary, most existing applications of MA involve model verification and model reduction to decrease the computational cost in solving optimization models. While this is valuable in design optimization, the implications of MA in the context of early design remained somewhat unexplored. Furthermore, most existing design optimization methodology has been developed with a focus on the accurate and efficient identification of the optimal set for increasingly complicated

problems. Little prior work was found that focused on understanding what the optimal result implies about the design itself - i.e. the underlying issues in the (configuration) design that causes the existence, shape, and location of the optimal set.

As a result, the mathematics and procedures for multiobjective monotonicity analysis (MOMA) and the study of the vertices and boundaries of the Pareto set (ϵMA) were developed as a part of this research. In unison, these two allow Pareto Set Dependency Analysis - i.e. the exhaustive identification of the global and regional dependencies that exist in the Pareto set. This entailed the development of:

1. The foundational theorems, corollaries and proofs that allow the application of MA to multiobjective problems.

2. The foundational theorems that allow the study of regional or even local dependencies that exist with the Pareto set.

3. Practical approaches allowing the application Pareto set dependency analysis in as many design contexts as possible and as early on in the design process as possible. This includes the selection of a suitable multiobjective formulation and optimization algorithm. These support the preservation of monotonicity throughout model reduction and allow the post-optimality reduction of the optimization problem using numerical constraint activity data.

These methodological developments allow the rigorous identification of the design variables and constraints that cause trade-offs. This can be done for a design that can be described by an algebraic model or represented through a meta-model that exhibits at least regionally monotonic behaviour. As discussed in Chapters 4 and 7, the use of monotonicity information actually allows the application of the analysis methodology from a very early stage of product development. This was exemplified in the SOMA case study, where Pareto set Dependency Analysis revealed several shared variables and active constraints, which caused or worsened the trade-offs between four design objectives of critical importance to the overall performance of the SOMA device. Most of these were found to be inherent to the specific configuration design that was studied, while the remaining few related the concept itself.

In essence, Pareto set Dependency Analysis provides the first rigorous approach to dependency analysis, which allows the consideration of the effect of active constraints. It reveals the limitations of a configuration design, thorough analysis on a level of abstraction that falls within proportional design. As found in Chapter 6, this knowledge can be of substantial value to designers, in that trade-off and constraint activity knowledge has been found to be a key distinguishing factor between novice designers and experts [33, 59]. While the theoretical developments alone support that Pareto set dependency analysis allows the rigorous identification of the root causes of trade-offs and the location of the Pareto set, the work involved in answering RQ3 clearly demonstrates the validity of the analysis method.

Research Question 3

What approaches and solutions can be used in design to remove, mitigate, or reduce the influence of trade-offs?

In short, trade-offs can be removed, mitigated, or reduced through targeted changes to the configuration design or the overall concept. This involves designing the system in a manner that avoids or reduces the influence of as many of the variables that are shared between design objectives in a manner that causes trade-offs as possible. All the while ensuring that

the remaining independent and *harmonious* shared variables have as nonrestrictive bounds in the *improving direction* as possible.

By studying the mathematical implications of the outputs of Pareto set Dependency Analysis, the research found several generic forms of design change that would achieve this, resulting in an *improved* Pareto set. The theorems and proofs that emerged in this process led to the identification of a set of design principles - the *Configuration Redesign Principles*. These principles are founded on basic types of transformations that result in a changed optimization problem and are applied directly based on the results of Pareto set analysis to eliminate/reduce specific dependencies that cause trade-offs or identify design changes that improve all objectives at once. The principles were found to have commonality with many existing heuristics but are somewhat unique in that they are entirely context-independent. Their reliance on the outputs of the generic analysis of the root causes of trade-offs between the objectives allows their application to trade-offs between any pair of design objectives in any design, so long as its behaviour can be described through an explicit or numerical model that can be studied through MOMA and ϵMA.

The systematic application of these design principles involves a procedure called *Systematic Configuration Design Improvement*, which ensures the targeted application of these principles based on analysis. The outcome of this process is a set of configuration redesigns that should have better Pareto sets, so long as the redesign principles have been applied successfully. In principle, this procedure can be applied iteratively in sequence with the analysis methodology, thereby continuously identifying the limitations of a configuration design, deriving improved configurations, selecting the preferred one, and repeating the process over and over. When parts of this iterative procedure are performed *interactively* throughout the embodiment design process, the methodological developments can be used to support almost any of the design and decision making activities involved in embodiment design. When done successfully, this would, in principle, steer the development process in a direction that allows continuous improvement of the design.

These design methods were applied to the SOMA case based on the results of the aforementioned analysis. This yielded a sequence of 11 redesign iterations. Based on the developed theorems, most of these redesigns should exhibit an improved Pareto set. Applying several forms of redesign evaluation to a promising selection of redesigns revealed that the application of the redesign principles does, in fact, result in a substantially improved Pareto set. An exhaustive evaluation was performed on one of them, allowing its comparison with the original SOMA device. This evaluation revealed that the Pareto set of the redesign dominates the entire Pareto set of the original design, with an overall reduction in trade-offs and an improvement in optimality overall. Hence, the research has successfully used formal methods to support the identification of design improvements far beyond what is achievable through proportional or parametric methods alone. This not only confirmed the contextual validity of the redesign principles. It also proved that Pareto Set Dependency Analysis is indeed valid, given that the elimination of the trade-off variables, and relaxation of the problematic constraints that were identified through analysis, had a substantial influence on the location and shape of the Pareto set.

Finally, all of these developments were found to have implications beyond the redesign of an existing embodiment. By considering how one might synthesise a design that avoids trade-off variables, allows the least restrictive constraints possible, and has a low structural complexity, this research resulted in a set of conjectures and related conditions describing the mathematical characteristics of such an *ideal design*. These were used to prescribe a set of design guidelines for design synthesis and configuration design that allow the convergence

towards this ideal. These guidelines are not only consistent with the preceding developments, but they were also found to be observable both in an existing product that is out in the market and in the SOMA case.

8.1.2 Hypothesis Testing

Hypothesis

The end performance of mechanical systems is ultimately determined by the trade-offs that affect the design of the system. It is hypothesised, that the decisions made in conceptual and embodiment design in effect determine the trade-offs that exist between the different functionalities and objectives, that a product is designed towards. Some of these trade-offs can be limited or prevented through the deliberate attempts at avoiding certain detrimental dependencies during design synthesis and change.

When we relate the answers to the research questions, we see that there are strong indications that this hypothesis is valid. This is both supported by theoretical developments and by their application in industrial practice.

The rigorously founded analysis and redesign methodology developed in this thesis is based on theorems, corollaries and proofs. These clearly demonstrate the relationship between certain types of dependencies and the existence, shape, and location of the Pareto set. They also support that the Pareto set is transformed when such dependencies are manipulated through design change. Such changes go beyond the proportional and parametric domain, for instance, involving changes in the layout and shape of parts, changes in working directions and load paths, and even changes to the selection of working principles. Hence, the methodology can be applied in the conceptual and embodiment design phases in the development of mechanical products. The practical application of this work to the SOMA case reveals that the relationships that create and shape the Pareto set can indeed be identified relatively early on. Further, the SOMA case also demonstrates the extent to which the underlying root causes affect the Pareto set, in that their mitigation through configuration redesign resulted in a new Pareto set with a drastically changed shape and location.

Thus, the theoretical developments and practical applications shown in this thesis indeed support that trade-offs are highly influential upon the performance of mechanical systems. The results also support that trade-offs are caused by decisions made during the early stages of design, and that they can, in fact, be avoided or reduced through the structured redesign efforts that are informed by analysis results.

8.2 Core contributions

The core contribution of this research is a systematic and rigorously founded methodological framework for the identification, understanding, and mitigation of the trade-offs between design objectives, which may be applied during the early stages of product development. More specifically, the core contribution includes:

1. *Multiobjective Monotonicity Analysis:* A systematic and rigorous (albeit opportunistic) approach to the reduction of multiobjective problems for the purposes of trade-off identification, model verification, and reduction in computational cost. Provided the optimization problem can be reduced far enough and that the problem is at least regionally differentiable and monotonic, the developments can be used to reveal all of the objectives in trade-off in a given design.

2. *Pareto set Dependency Analysis:* The first analysis method that allows the study of the relationships that exist in the optimal set by allowing the consideration of the effects of active constraints. Existing methods for dependency analysis largely disregard the influence of the bounds of design variables. The insights gained through this analysis is of special value in early design, as this accumulation of this knowledge is largely related to the designer's experience today.

3. *Systematic Configuration Design Improvement*: An approach to the introduction of targeted design changes that will result in an improved Pareto set, which is steered by the results of quantitative analysis.

4. *Interactive trade-off analysis and mitigation*: The preceding developments are collated into a process involving the interactive analysis and mitigation of trade-offs. This allows the embodiment design process to be steered towards ending with the *best* possible design, as it accounts for the design problem (i.e. the objectives) changing as the solution is developed and refined.

5. *The Guidelines of Ideal Design Synthesis*: A set of guidelines aimed at avoiding many of the issues that might be identified and mitigated later in the design process using analysis and redesign methods. These guidelines are founded on a novel perspective as to what "good" design even constitutes, stated in the form of the Conjectures and Conditions of Ideal Design.

6. *SOMA Analysis, Redesign, and Evaluation:* The case studies performed on the SOMA device as a part of the theoretical developments, have yielded valuable insights and numerous redesigns that result in improved performance. This had led to several patent applications.

Based on the four different forms of validity discussed in the previous chapter, we can conclude that these contributions are founded on well-accepted and rigorous methods, empirically validated using a suitable case from industry, and largely context-independent. One cannot claim that the use of the methods developed in this thesis will always allow the design of superior products than if the methods are not employed. It will not always be possible or time-efficient to build optimization models at an early stage of development, and nor can they always be reduced through monotonicity analysis. Correspondingly, expert designers may, in many cases, be able to identify similar changes without any analysis. Yet, this gets to the heart of the basic challenge in design research; without a systematic methodology for design improvement, engineering design remains a craft rather than a science.

8.2.1 Academic Value

Pareto set dependency analysis presents an optimization-focused alternative to current techniques for dependency analysis (e.g. DSM and Axiomatic Design [13]), with MOMA allowing the analysis of dependencies unique to the optimum by accounting for the influence of constraints, and ϵMA revealing regional dependencies in the Pareto set. From a design optimization perspective, Pareto set dependency analysis is a rigorous approach to exploring the limitations of a given embodiment/configuration design, with the added benefit of a reduced computational cost. One could view Pareto-set Dependency Analysis as an approach to *checking the design* ahead of computation or *explaining the results* after computation [c.f. Paper A].

The principles of configuration redesign and the systematic procedure in applying them empower the designer to identify configuration design changes that actually improve the performance of the system. As this is steered by the outputs of analysis and is context independent,

the principles and procedure address a key limitation of many heuristics; that it can be difficult to know which heuristics to apply to improve a given design and when their application will actually result in improvement. These developments differ from previous prescriptions regarding dependency (e.g. Axiomatic Design [13]), given that they demonstrate that dependencies are not necessarily detrimental and that whether a constraint is active or not can have a substantial influence on the dependencies that exist between design objectives. One can therefore not disregard the influence of the bounds of design variables, nor can one claim that dependencies should be avoided at all costs.

Finally, the developments from Chapter 6 are distinct from existing prescriptions of" good" design, e.g. the Axioms of Axiomatic Design [13], the Ideal Final Result from TRiZ [16], and the Basic Rules of Embodiment Design by Pahl & Beitz' [6]. The Conjectures and Conditions involve basic goals that can be stated mathematically and actually account for the role of the bounds of design variables. Combined with the redesign methodology, these conjectures and conditions allowed the development of design guidelines, which only rely on the designer having a basic understanding of the monotonicity and bounds of the design variables at hand.

8.2.2 Industrial Value

The results of this research may have value to industry in numerous ways. First of all, given that it is well established that context-specific knowledge of trade-offs and constraint activity is a key differentiator between novices and experienced designers, one could argue that the systematic management of trade-offs is expensive today. Gaining experience is, by definition, time-consuming. Hence, industry might be paying the price for the lack of systematic methodologies for trade-off management, needing to accept that novice engineers will simply need to spend more time or fail more often in developing an acceptable configuration design. The alternative would be to simply hire a larger amount of experienced designers, which inevitably has an influence on wage costs.

Furthermore, the methods developed in this thesis allow the identification, understanding, and potential mitigation of trade-offs in early product development. While this does not guarantee an end-product with better performance, it has the potential to increase the likelihood of achieving good performance. Certainly, the SOMA case studies at least support that this is possible.

In industry, many product requirements and specifications are also not necessarily defined by necessity but rather relate to "*nice to haves*" or certain organisational desires. These may cause or worsen trade-off scenarios. The developments of this thesis might allow more upfront *negotiations* for such decisions, allowing the considerations of what trade-offs and resulting technical risk a given specification or desired new sub-functionality give rise to, avoiding situations where specification limits that yield no added utility, end up worsening the overall product. Correspondingly, an improved understanding of trade-offs allows them to be managed upfront and included in project management. This lessens the risk of project loop-backs resulting in increased lead time and the risk of unexpected late-stage compromises and worsened product specifications. Potential target groups that may gain value from the developments of this thesis include design engineers, systems engineers, analysis experts (optimization, structural mechanics, etc.), decision-makers in product development, and related stakeholders who influence the constraints and objectives products are designed towards (e.g. production, marketing, industrial design, etc.) .

8.3 Further Work

Throughout this PhD, numerous interesting perspectives were encountered, which could simply not be explored within the time frame of the project, or which were beyond the PhD fellow's competencies. This section presents a small portion of the broad palette of research topics that could build upon or complement the work presented in this thesis.

8.3.1 Automation and simplification of analysis

While limited successes have been achieved in the automation of monotonicity analysis[69, 70], its mostly manual nature is a key limitation in its application, especially to large and non-monotonic design problems. However, since the original development of Monotonicity analysis by Papalambros and Wilde [45], large strides have been made within the fields of data analysis, machine learning, and interactive computing. Applying such methods to identify monotonic properties through the analysis of large data sets or to allow automated/guided model reductions post-optimality may provide designers with some of the design insights achieved through Pareto set dependency analysis. This might reduce the analysis effort substantially, thereby heightening the value of design optimization methods to inform redesign processes in early design.

8.3.2 Mixed analyses - Quasi-Quantitative Trade-off Analysis

In practice, design involves objectives and constraints which are not easily modelled or measured. Characteristics such as aesthetics, circularity, usability, manufacturability can, in many cases, be translated into more quantifiable proxies but are difficult to describe in the form of a scalar objective function. Yet, such objectives can still be in trade-off with other objectives or amongst themselves. Even though they might not be directly quantifiable, they may in many cases still exhibit monotonic relationships with the design variables and parameters in a mechanical system. An example of such is that the negative environmental impact of a mechanical system increases monotonically with material consumption and decreases with mechanical efficiency and durability. Hence, we might apply the same approaches as developed for Pareto set Dependency Analysis in mixed analyses consisting of both quantitative objective functions and more qualitative/fuzzy objectives. While such analysis would have little meaning in the discussion of the Pareto set, its outputs could still be used in the application of the Systematic Configuration Design Improvement method, allowing the mitigation of trade-offs between quantifiable and non-quantifiable objectives, thereby further informing configuration design. Such aspects have gone untouched in this PhD, but might be a valuable topic of further research.

8.3.3 Designer Cognition in Trade-off Management

How designers think and reason when identifying, understanding, mitigating, or accepting trade-offs during the synthesis and improvement of mechanical systems, is an avenue of research that has been disregarded in this PhD. Nonetheless, an important question in this context is the extent to which a designer's cognitive biases, risk-willingness, and tendency towards design fixation, affects the way they design mechanical systems and (fail to) consider trade-offs throughout the design process. Some designers might be more prone to avoiding them altogether, while others may be overly reliant on analysis, become fixated on certain objectives over others, or falsely assume that an acceptable optimum exists. Given the findings of Ahmed et al. [33], Cross [35], and Howard & Andreasen [124], it seems clear that trade-off knowledge plays a key role in designers aggregation of experience and the decision making of some "expert" designers. Yet, for some reason, designers are still prone to relying on intuition over analysis [34], with cognitive biases [23] seemingly having an influence on the acceptance of trade-off studies. Research into the mechanisms involved may foster a better understanding of how to manage the development of integrated and multi-functional mechanical systems and have perspectives for the didactics of mechanical design as well.

8.3.4 Conceptual and Generative Design

As discussed in the previous chapters, the methodologies developed in this thesis are - besides the developments in Section 6.1. - largely dependent on the analysis of an already synthesised embodiment and on the creativity/ability of the designer to translate principles developed in this thesis into actualised design changes or new systems. There might very well be steps or rules that one needs to follow in conceptual design and in synthesis beyond the ones defined in the Guidelines for Ideal Design Synthesis. Research efforts that attempt to extend the developments in Chapters 5 and 6 into more synthesis and generative design-related areas might hence yield new perspectives of use in mechanical design synthesis. One approach to this might be to work on extending the methods from this thesis by applying and adapting them to optimization models built on another level of abstraction than the purely proportional. This may reveal rigorous redesign and synthesis methods with applications in conceptual design, just as the work in this PhD revealed configuration redesign methods using proportional optimization methods. In this context, it may be worthwhile building on the optimal configuration design approaches developed by Schmidt & Cagan [106] or by Bayrak et al. [107, 108], to see if approaches resembling MOMA and ϵMA might reveal important information about the limitations of an overall concept stemming from the selection and combination working principles into an overall working structure.

8.4 Concluding Remarks

This PhD project has been an enlightening and challenging journey, both on a professional and personal level. It has required learnings I had not imagined, involved challenges I had not foreseen, and experiences I would not want to be without today. Be it my immersion into the design optimization field, the sudden need to comprehend the intricacies of patent law, the re-planning necessitated by getting hit by a car on my way to work, or the unforeseen effects of a global pandemic, this PhD has, if nothing else, always been interesting.

I have always been of the opinion that challenging problems are the most interesting ones. In my time as a design engineer before this PhD, I encountered a no bigger challenge in mechanical design than the multi-disciplinary, multi-dimensional, and oft counter-intuitive nature of the trade-offs that one encounters or overlooks in the early stages in design. Upon the conclusion of this PhD, my opinion has not changed. As designers, we usually strive for the "*best*" design, but what best constitutes, and how it is achieved in the face of incomplete knowledge and seemingly conflicting requirements, is elusive to most except the *ingenious* few. I hope that the work presented in this thesis at least scratches the surface of the underlying science of "good" configuration (re)design. I look forward to continuing the application of this work, be it in my continuing employment in industry or through a continued affiliation with the academic community in some capacity. I sincerely hope that my career will continue to present opportunities for further learning and intellectual pursuits, albeit in a less intense and all-consuming form than what is at times required of a PhD student.

References

[1] Editors. *Industrial Revolution*. 2021. URL: https://www.britannica.com/event/Industrial-Revolution/.

[2] E. B. Woodruff et al. *Energy Conversion*. 2016. URL: https://www.britannica.com/technology/energy-conversion.

[3] N. P. Suh. *Axiomatic Design: Advances and Applications*. Oxford University Press, 2001.

[4] L. T. Blessing and A. Chakrabarti. *DRM, a design research methodology*. Springer Berlin Heidelberg, 2009. DOI: 10.1007/978-1-84882-587-1.

[5] P. Y. Papalambros. "Design Science: Why, What and How". In: *Design Science* 1 (2015). DOI: 10.1017/dsj.2015.1.

[6] G. Pahl and W. Beitz. *Engineering design — A systematic approach*. 1999. DOI: 10.1016/0261-3069(96)84970-3.

[7] K. Ulrich, S. D. Eppinger, and M. C. Yang. *Product Design and Development*. 7th ed. McGraw Hill, 2020.

[8] M. Andreasen and L. Hein. *Integrated Product Development*. IFS/Springer, 2000, p. 220.

[9] E. Tjalve. "Form Design - A Systematic Approach". In: *Schriftenreihe WDK (Workshop Design - Konstruktion)*. 1981.

[10] D. G. Ullman. *The mechanical design process*. 6th ed. McGraw-Hill New York, 2017.

[11] E. K. Antonsson and J. Cagan, eds. *Formal Engineering Design Synthesis*. Cambridge University Press, Nov. 2001. DOI: 10.1017/cbo9780511529627. URL: https://www.cambridge.org/core/product/identifier/9780511529627/type/book.

[12] P. Y. Papalambros and D. J. Wilde. *Principles of Optimal Design*. Cambridge University Press, Jan. 2017. DOI: 10.1017/9781316451038. URL: https://www.cambridge.org/core/product/identifier/9781316451038/type/book.

[13] N. P. Suh. "Axiomatic Design Theory for Systems". In: *Research in Engineering Design - Theory, Applications, and Concurrent Engineering* 10.4 (1998), pp. 189–209. DOI: 10.1007/s001639870001.

[14] M. J. French. *Conceptual Design for Engineers*. 1985. DOI: 10.1007/978-3-662-11364-6.

[15] H. G. Sillitto. "On systems architects and systems architecting: Some thoughts on explaining and improving the art and science of systems architecting". In: *19th Annual International Symposium of the International Council on Systems Engineering, INCOSE 2009*. 2009. DOI: 10.1002/j.2334-5837.2009.tb00995.x.

[16] G. Altshüller. *Creativity As an Exact Science*. Gordon and Breach, 1984. DOI: 10.1201/9781466593442.

[17] S. M. Göhler and T. J. Howard. "The contradiction index (CI): A new metric combining system complexity and robustness for early design stages". In: *Proceedings of the ASME Design Engineering Technical Conference*. Vol. 7. 2015. DOI: 10.1115/DETC201547255.

[18] P. J. Clarkson, C. Simons, and C. Eckert. "Predicting change propagation in complex design". In: *Journal of Mechanical Design, Transactions of the ASME* 126.5 (2004), pp. 788–797. DOI: 10.1115/1.1765117.

[19] G. Vianello and S. Ahmed-Kristensen. "A comparative study of changes across the lifecycle of complex products in a variant and a customised industry". In: *Journal of Engineering Design* 23.2 (2012), pp. 99–117. DOI: 10.1080/09544828.2010.542133.

[20] M. M. Andreasen and T. J. Howard. "Is Engineering Design Disappearing from Design Research ?" In: *The Future of Design Methodology*. Ed. by H. Birkhofer. London: Springer Verlag, 2011. Chap. 2, pp. 21–34. DOI: 10.1007/978-0-85729-615-3.
[21] S. M. Göhler, D. D. Frey, and T. J. Howard. "A model-based approach to associate complexity and robustness in engineering systems". In: *Research in Engineering Design* 28.2 (Apr. 2017), pp. 223–234. DOI: 10.1007/s00163-016-0236-1.
[22] M. Unal, G. P. Warn, and T. W. Simpson. "Quantifying the shape of pareto fronts during multi-objective trade space exploration". In: *Journal of Mechanical Design, Transactions of the ASME* 140.2 (2018), pp. 1–13. DOI: 10.1115/1.4038005.
[23] E. D. Smith. "Tradeoff studies and cognitive biases". PhD thesis. The University of Arisona, 2006, p. 225.
[24] D. C. Wynn and C. M. Eckert. *Perspectives on iteration in design and development*. Vol. 28. 2. Springer London, 2017, pp. 153–184. DOI: 10.1007/s00163-016-0226-3.
[25] S. M. Göhler. "Metric-driven Robust Design – Robustness Quantification of Complex Engineering Systems". PhD thesis. Technical University of Denmark, 2017, p. 173.
[26] D. Frey, J. Palladino, J. Sullivan, and M. Atherton. "Part count and design of robust systems". In: *Systems Engineering* 10.3 (Sept. 2007), pp. 203–221. DOI: 10.1002/sys.20071.
[27] N. S. Sigurdarson, T. Eifler, and M. Ebro. "Functional trade-offs in the mechanical design of integrated products - Impact on robustness and optimisability". In: *Proceedings of the International Conference on Engineering Design, ICED* 2019-Augus.August (2019), pp. 3491–3500. DOI: 10.1017/dsi.2019.356.
[28] W. B. Arthur. "Why Do Things Become More Complex?" In: *Scientific American* 268.5 (1993), pp. 144–144. DOI: 10.1038/scientificamerican0593-144.
[29] B. Matthiassen. "Design for Robustness and Reliability - Improving the Quality Conciousness in Engineering Design". PhD thesis. Technical University of Denmark, 1997.
[30] O. Isaksson and C. Eckert. *Product Development 2040: Technologies are just as good as the designer's ability to integrate them*. Tech. rep. September. Design Society Report DS107, 2020. DOI: https://doi.org/10.35199/report.pd2040.
[31] D. K. Sobek, A. C. Ward, and J. K. Liker. "Toyota ' s Principles of Set-Based Concurrent Engineering Toyota ' s Principles of Set-Based Concurrent Engineering". In: *Sloan Management Review* 40.2 (1999), pp. 67–83.
[32] J. Cagan and A. M. Agogino. "Innovative design of mechanical structures from first principles". In: *Artificial Intelligence for Engineering, Design, Analysis and Manufacturing* 1.3 (1987), pp. 169–189. DOI: 10.1017/S0890060400000275.
[33] S. Ahmed, K. M. Wallace, and L. T. Blessing. "Understanding the differences between how novice and experienced designers approach design tasks". In: *Research in Engineering Design* 14.1 (2003), pp. 1–11. DOI: 10.1007/s00163-002-0023-z.
[34] B. Kleinmuntz. "Why we still use our heads instead of formulas: Toward an integrative approach". In: *Psychological Bulletin* 107.3 (1990), pp. 296–310. DOI: 10.1037/0033-2909.107.3.296.
[35] N. Cross. "Expertise in design: An overview". In: *Design Studies* 25.5 (2004), pp. 427–441. DOI: 10.1016/j.destud.2004.06.002.
[36] *Novo Nordisk Annual Report 2019*. Tech. rep. Novo Nordisk A/S, 2019.
[37] W. C. Lee et al. "Medication adherence and the associated health-economic impact among patients with type 2 diabetes mellitus converting to insulin pen therapy: an analysis of third-party managed care claims data." eng. In: *Clinical therapeutics* 28.10 (Oct. 2006), pp. 1711–1712. DOI: 10.1016/j.clinthera.2006.10.004.
[38] Y. Wu and A. Wu. *Taguchi Methods for Robust Design*. ASME Press, Jan. 2000. DOI: 10.1115/1.801578. URL: https://doi.org/10.1115/1.801578.

[39] M. Ebro. "Applying Robust Design in an Industrial Context". PhD thesis. Technical University of Denmark, 2016, p. 150.
[40] D. Lowe. "The Latest on Drug Failure and Approval Rates". In: *Science and Translational Medicine* (2019). URL: https://blogs.sciencemag.org/pipeline/archives/2019/05/09/the-latest-on-drug-failure-and-approval-rates.
[41] A. Abramson et al. "An ingestible self-orienting system for oral delivery of macromolecules". In: *Science* 363.6427 (2019). DOI: 10.1126/science.aau2277. URL: http://science.sciencemag.org/.
[42] W. Sircus. *Human Digestive System - Structures of the human stomach*. 2021. URL: https://www.britannica.com/science/human-digestive-system/images-videos#/media/1/1081754/68634.
[43] J. E. Van Aken. "Management research as a design science: Articulating the research products of mode 2 knowledge production in management". In: *British Journal of Management* 16.1 (2005), pp. 19–36. DOI: 10.1111/j.1467-8551.2005.00437.x.
[44] K. A. Jørgensen. *Videnskabelige Arbejdsparadigmer (Scientific working paradigms)*. Tech. rep. Aalborg: Institut for Production, Aalborg University, Denmark, 1992, p. 5.
[45] P. Papalambros and D. J. Wilde. "Global Non-iterative Design Optimization Using Monotonicity Analysis". In: *Journal of Mechanical Design, Transactions of the ASME* 78 -WA/DE-17 (1978).
[46] N. K. Denzin and Y. S. Lincoln. *The SAGE Handbook of Qualitative Research*. Ed. by N. K. Denzin and Y. S. Lincoln. 3rd ed. Thousand Oaks: Sage Publications, 2005. URL: https://books.google.dk/books?id=X85J8ipMpZEC.
[47] K. Pedersen et al. "Validating Design Methods and Research: The Validation Square". In: *Proceedings of DETC '00 2000 ASME Design Engineering Technical Conferences*. September. 2000, pp. 379–390. DOI: 10.1115/detc2000/dtm-14579.
[48] D. D. Frey and C. L. Dym. "Validation of design methods: Lessons from medicine". In: *Research in Engineering Design* 17.1 (2006), pp. 45–57. DOI: 10.1007/s00163-006-0016-4.
[49] Y. Barlas and S. Carpenter. "Philosophical roots of model validation: Two paradigms". In: *System Dynamics Review* 6.2 (1990), pp. 148–166. DOI: https://doi.org/10.1002/sdr.4260060203. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sdr.4260060203.
[50] J. S. Arora. *Introduction to Optimum Design*. 2012. DOI: 10.1016/b978-0-12-381375-6.00004-8.
[51] D. G. Ullman, T. G. Dietterich, and L. A. Stauffer. "A model of the mechanical design process based on empirical data". In: *Artificial Intelligence for Engineering, Design, Analysis and Manufacturing* 2.1 (1988), pp. 33–52. DOI: 10.1017/S0890060400000536.
[52] W. Karush. "Minima of functions of several variables with inequalities as side conditions." PhD thesis. Thesis (S.M.)–University of Chicago, Department of Mathematics, December 1939., 1939.
[53] H. W. Kuhn and A. W. Tucker. "Nonlinear Programming". In: *Proceedings of 2nd Berkeley Symposium*. 1951. DOI: 10.1007/978-3-030-55404-0{_}3.
[54] R. T. Marler and J. S. Arora. "Survey of multi-objective optimization methods for engineering". In: *Structural and Multidisciplinary Optimization* 26.6 (2004), pp. 369–395. DOI: 10.1007/s00158-003-0368-6.
[55] T. W. Athan and P. Y. Papalambros. "A quasi-Monte Carlo method for multicriteria design optimization". In: *Engineering Optimization* 27.3 (1996), pp. 177–198. DOI: 10.1080/03052159608941405.

[56] Y. Y. Haimes and W. A. Hall. "Multiobjectives in water resource systems analysis: The Surrogate Worth Trade Off Method". In: *Water Resources Research* 10.4 (1974), pp. 615–624. DOI: 10.1029/WR010i004p00615.

[57] D. Carmichael. "Computation of Pareto Optima in Structural Design". In: *International Journal for Numerical Methods in Engineering* 15 (1980), pp. 925–952. DOI: 10.1017/S0022029900029393.

[58] P. Y. Papalambros. "Model Reduction and Verification Techniques". In: *Advances in Design Optimization*. Ed. by H. Adeli. New York: Chapman and Hall, 1994, pp. 109–138.

[59] C. M. Eckert and M. K. Stacey. "Constraints and Conditions: Drivers for Design Processes". In: *An Anthology of Theories and Models of Design: Philosophy, Approaches and Empirical Explorations*. Ed. by A. Chakrabarti and L. T. M. Blessing. London: Springer London, 2014, pp. 395–415. DOI: 10.1007/978-1-4471-6338-1{_}19. URL: https://doi.org/10.1007/978-1-4471-6338-1_19.

[60] P. Y. Papalambros. *Remarks on sufficiency of constraint-bound solutions in optimal design*. 1993. DOI: 10.1115/1.2919201.

[61] S. D. Eppinger and T. R. Browning. *Design Structure Matrix Methods and Applications*. 2018. DOI: 10.7551/mitpress/8896.001.0001.

[62] H. M. Kim, N. F. Michelena, P. Y. Papalambros, and T. Jiang. "Target cascading in optimal system design". In: *Journal of Mechanical Design, Transactions of the ASME* 125.3 (2003), pp. 474–480. DOI: 10.1115/1.1582501. URL: https://www.researchgate.net/publication/33984207.

[63] T. W. Simpson and J. R. Martins. "Multidisciplinary design optimization for complex engineered systems: Report from a national science foundation workshop". In: *Journal of Mechanical Design, Transactions of the ASME* 133.10 (2011), pp. 1–10. DOI: 10.1115/1.4004465.

[64] G. G. Wang and S. Shan. "Review of metamodeling techniques in support of engineering design optimization". In: *Journal of Mechanical Design, Transactions of the ASME* 129.4 (2007), pp. 370–380. DOI: 10.1115/1.2429697.

[65] S. Tosserams, M. Kokkolaras, L. F. Etman, and J. E. Rooda. "A nonhierarchical formulation of analytical target cascading". In: *Journal of Mechanical Design, Transactions of the ASME* 132.5 (2010), pp. 0510021–05100213. DOI: 10.1115/1.4001346.

[66] R. J. Balling and J. Sobieszczanski-Sobieski. "Optimization of coupled systems - A critical overview of approaches". In: *AIAA Journal* 34.1 (1996), pp. 6–17. DOI: 10.2514/3.13015. URL: https://doi.org/10.2514/3.13015.

[67] M. P. Bendsøe and N. Kikuchi. "Generating optimal topologies in structural design using a homogenization method". In: *Computer Methods in Applied Mechanics and Engineering* (1988). DOI: 10.1016/0045-7825(88)90086-2.

[68] P. Papalambros and D. J. Wilde. "Regional monotonicity in optimum design". In: *Journal of Mechanical Design, Transactions of the ASME* (1980). DOI: 10.1115/1.3254774.

[69] S. Azarm and P. Papalambros. "An automated procedure for local monotonicity analysis". In: *Journal of Mechanical Design, Transactions of the ASME* 106.1 (1984), pp. 82–89. DOI: 10.1115/1.3258566.

[70] J. Zhou and R. W. Mayne. "Monotonicity analysis and the reduced gradient method in constrained optimization". In: *Journal of Mechanical Design, Transactions of the ASME* 106.1 (1984), pp. 90–94. DOI: 10.1115/1.3258567.

[71] J. Zhou and R. W. Mayne. "Interactive Computing in the Application of Monotonicity Analysis to Design Optimization". In: *Journal of Mechanisms, Transmissions, and Automation in Design* 105.2 (June 1983), pp. 181–186. DOI: 10.1115/1.3258506. URL: https://doi.org/10.1115/1.3258506.

[72] N. F. Michelena and A. M. Agogino. "Multiobjective Hydraulic Cylinder Design". In: *Journal of Mechanisms, Transmissions, and Automation in Design* 110.1 (Mar. 1988), pp. 81–87. DOI: 10.1115/1.3258910. URL: https://doi.org/10.1115/1.3258910.
[73] M. Gobbi, F. Levi, G. Mastinu, and G. Previati. "On the analytical derivation of the Pareto-optimal set with applications to structural design". In: *Structural and Multidisciplinary Optimization* 51.3 (2015), pp. 645–657. DOI: 10.1007/s00158-014-1152-5.
[74] N. Riquelme, C. Von Lücken, and B. Barán. "Performance metrics in multi-objective optimization". In: *Proceedings - 2015 41st Latin American Computing Conference, CLEI 2015*. 2015. DOI: 10.1109/CLEI.2015.7360024.
[75] B. Frischknecht and P. Papalambros. "A PARETO APPROACH TO ALIGNING PUBLIC AND PRIVATE OBJECTIVES IN VEHICLE DESIGN". In: 2008.
[76] J. Hamel, M. Li, and S. Azarm. "Design Improvement by Sensitivity Analysis Under Interval Uncertainty Using Multi-Objective Optimization". In: *Journal of Mechanical Design* 132.8 (Aug. 2010). DOI: 10.1115/1.4002139. URL: https://doi.org/10.1115/1.4002139.
[77] S. Gunawan and S. Azarm. "Multi-objective robust optimization using a sensitivity region concept". In: *Structural and Multidisciplinary Optimization* 29.1 (2005), pp. 50–60. DOI: 10.1007/s00158-004-0450-8. URL: https://doi.org/10.1007/s00158-004-0450-8.
[78] C. A. Mattson and A. Messac. *Pareto Frontier Based Concept Selection Under Uncertainty, with Visualization*. Tech. rep. 2005, pp. 85–115.
[79] S. Bhattacharya et al. "Incorporating quantitative reliability engineering measures into tradespace exploration". In: *Research in Engineering Design* 29.4 (2018), pp. 589–603. DOI: 10.1007/s00163-018-0293-8. URL: http://dx.doi.org/10.1007/s00163-018-0293-8.
[80] K. N. Otto. "Imprecision in engineering design". In: *Journal of Mechanical Design, Transactions of the ASME* 117.B (1995), pp. 25–32. DOI: 10.1115/1.2836465.
[81] R. C. Purshouse and P. J. Fleming. "Conflict, Harmony, and Independence: Relationships in Evolutionary Multi-criterion Optimisation". In: *Evolutionary Multi-Criterion Optimization*. Ed. by C. M. Fonseca et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 16–30.
[82] I. Das. "A Preference Ordering Among Various Pareto Optimal Alternatives". In: *Structural Optimization* 18 (1999), pp. 30–35. DOI: 10.1007/BF01210689.
[83] J. C. Kelly, P. Maheut, J.-F. Petiot, and P. Y. Papalambros. "Incorporating user shape preference in engineering design optimisation". In: *Journal of Engineering Design* 22.9 (2011), pp. 627–650. DOI: 10.1080/09544821003662601. URL: https://doi.org/10.1080/09544821003662601.
[84] K. N. Otto and E. K. Antonsson. "Trade-off strategies in engineering design". In: *Research in Engineering Design* 3.2 (1991), pp. 87–103. DOI: 10.1007/BF01581342.
[85] C. A. Mattson and A. Messac. "Concept Selection Using s-Pareto Frontiers". In: *AIAA Journal* 41.6 (2003), pp. 1190–1198. DOI: 10.2514/2.2063. URL: https://doi.org/10.2514/2.2063.
[86] T. W. Athan and P. Y. Papalambros. "A note on weighted criteria methods for compromise solutions in multi-objective optimization". In: *Engineering Optimization* 27.2 (1996), pp. 155–176. DOI: 10.1080/03052159608941404.
[87] A. Abi Akle, S. Minel, and B. Yannou. "Information visualization for selection in Design by Shopping". In: *Research in Engineering Design* 28.1 (2017), pp. 99–117. DOI: 10.1007/s00163-016-0235-2.
[88] C. A. Mattson, A. A. Mullur, and A. Messac. "Smart Pareto filter: obtaining a minimal representation of multiobjective design space". In: *Engineering Optimization* 36.6

(2004), pp. 721–740. DOI: 10.1080/0305215042000274942. URL: https://doi.org/10.1080/0305215042000274942.

[89] T. W. Simpson, D. Carlsen, M. Malone, and J. Kollat. "Trade Space Exploration: Assessing the Benefits of Putting Designers "Back-in-the-Loop" during Engineering Optimization". In: *Human-in-the-Loop Simulations: Methods and Practice*. Ed. by L. Rothrock and S. Narayanan. London: Springer London, 2011, pp. 131–152. DOI: 10.1007/978-0-85729-883-6{_}7. URL: https://doi.org/10.1007/978-0-85729-883-6_7.

[90] A. M. Ross and D. E. Hastings. "The tradespace exploration paradigm". In: *15th Annual International Symposium of the International Council on Systems Engineering, INCOSE 2005*. Vol. 2. 2005, pp. 1706–1718. DOI: 10.1002/j.2334-5837.2005.tb00783.x.

[91] M. Unal, G. P. Warn, and T. W. Simpson. "Quantifying tradeoffs to reduce the dimensionality of complex design optimization problems and expedite trade space exploration". In: *Structural and Multidisciplinary Optimization* 54.2 (2016), pp. 233–248. DOI: 10.1007/s00158-015-1389-7. URL: http://dx.doi.org/10.1007/s00158-015-1389-7.

[92] C. M. Fonseca and P. J. Fleming. "Multiobjective optimization and multiple constraint handling with evolutionary algorithms - Part I: A unified formulation". In: *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans.* (1998). DOI: 10.1109/3468.650319.

[93] J. R. Hauser and D. Clausing. "The house of quality". In: *Harvard Business Review* 66.May-June (1988).

[94] J. M. Torry-Smith, N. H. Mortensen, and S. Achiche. "A proposal for a classification of product-related dependencies in development of mechatronic products". In: *Research in Engineering Design* 25.1 (2014), pp. 53–74. DOI: 10.1007/s00163-013-0161-5.

[95] T. Pimmler and S. D. Eppinger. "Integration Analysis of Product Descriptions." In: *Proceedings of IDETC 1994, ASME Design Engineering Technical Conferences*. September. 1994.

[96] J. G. Skakoon. *The Elements of Mechanical Design*. 2008. DOI: 10.1115/1.802670.

[97] M. J. French. "The opportunistic route and the role of design principles". In: *Research in Engineering Design* 4.3 (1992), pp. 185–190. DOI: 10.1007/BF01607946.

[98] M. J. French. "An annotated list of design principles". In: *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture* 208.4 (1994), pp. 229–234. DOI: 10.1243/PIME{_}PROC{_}1994{_}208{_}083{_}02.

[99] S. Finger and J. R. Dixon. "A review of research in mechanical engineering design. Part I: Descriptive, prescriptive, and computer-based models of design processes". In: *Research in Engineering Design* 1.1 (1989), pp. 51–67. DOI: 10.1007/BF01580003. URL: https://doi.org/10.1007/BF01580003.

[100] M. J. Chalupnik, D. C. Wynn, and P. J. Clarkson. "Comparison of ilities for protection against uncertainty in system design". In: *Journal of Engineering Design* 24.12 (2013), pp. 814–829. DOI: 10.1080/09544828.2013.851783. URL: https://doi.org/10.1080/09544828.2013.851783.

[101] I. M. Ilevbare, D. Probert, and R. Phaal. "A review of TRIZ, and its benefits and challenges in practice". In: *Technovation* 33.2-3 (2013), pp. 30–37. DOI: 10.1016/j.technovation.2012.11.003. URL: http://dx.doi.org/10.1016/j.technovation.2012.11.003.

[102] P. Jain and A. M. Agogino. "Theory of design: An optimization perspective". In: *Mechanism and Machine Theory* 25.3 (1990), pp. 287–303. DOI: 10.1016/0094-114X(90)90030-N.
[103] K. Ishii and P. Barkan. "Active Constraint Deduction - A Framework for Expert Systems in Mechanical Systems Design". In: *Advances in Design Automation - ASME Design Technology Conferences - The Design Automation Conference*. Vol. 10. 1987.
[104] K. Deb and A. Srinivasan. "Innovization: Innovating design principles through optimization". In: *GECCO 2006 - Genetic and Evolutionary Computation Conference* 2 (2006), pp. 1629–1636.
[105] A. Chakrabarti et al. "Computer-based design synthesis research: An overview". In: *Journal of Computing and Information Science in Engineering* 11.2 (2011). DOI: 10.1115/1.3593409.
[106] L. C. Schmidt and J. Cagan. "Optimal Configuration Design: An Integrated Approach Using Grammars". In: *Journal of Mechanical Design* 120.1 (Mar. 1998), pp. 2–9. DOI: 10.1115/1.2826672. URL: https://doi.org/10.1115/1.2826672.
[107] A. E. Bayrak, N. Kang, and P. Y. Papalambros. "Decomposition-Based Design Optimization of Hybrid Electric Powertrain Architectures: Simultaneous Configuration and Sizing Design". In: *Journal of Mechanical Design, Transactions of the ASME* 138.7 (2016), pp. 1–9. DOI: 10.1115/1.4033655.
[108] A. E. Bayrak, Y. Ren, and P. Y. Papalambros. "Topology Generation for Hybrid Electric Vehicle Architecture Design". In: *Journal of Mechanical Design, Transactions of the ASME* 138.8 (2016). DOI: 10.1115/1.4033656.
[109] N. Lyu and K. Saitou. "Topology optimization of multicomponent beam structure via decomposition-based assembly synthesis". In: *Journal of Mechanical Design, Transactions of the ASME* 127.2 (2005), pp. 170–183. DOI: 10.1115/1.1814671.
[110] K. S. Channer and J. P. Virjee. "The effect of size and shape of tablets on their esophageal transit". In: *Journal of Clinical Pharmacology* 26.2 (1986), pp. 141–146. DOI: 10.1002/j.1552-4604.1986.tb02922.x.
[111] U.S. Department of Health and Human Services Food and Drug Administration (CDER). "Guidance for Industry: Size, Shape and Other Physical Attributes of Generic Tablets and Capsules". In: *Pharmaceutical Quality/CMC* December (2013), pp. 1–11. URL: http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm.
[112] A. Wahl. *Mechanical Springs*. 1st ed. Penton Publishing Company, 1944.
[113] J. G. Lin. "Maximal Vectors and Multi-Objective Optimization". In: *Journal of Optimization Theory and Applications* 18.01 (1976).
[114] E. M. Kasprzak and K. E. Lewis. "Pareto analysis in multiobjective optimization using the collinearity theorem and scaling method". In: *Structural and Multidisciplinary Optimization* 22.3 (Oct. 2001), pp. 208–218. DOI: 10.1007/s001580100138. URL: http://link.springer.com/10.1007/s001580100138.
[115] U. Chouinard, S. Achiche, and L. Baron. "Integrating negative dependencies assessment during mechatronics conceptual design using fuzzy logic and quantitative graph theory". In: *Mechatronics* 59.April 2018 (2019), pp. 140–153. DOI: 10.1016/j.mechatronics.2019.03.009. URL: https://doi.org/10.1016/j.mechatronics.2019.03.009.
[116] G. A. Hazelrigg. "On the Role and Use of Mathematical Models in Engineering Design". In: *Journal of Mechanical Design* 121.3 (Sept. 1999), pp. 336–341. DOI: 10.1115/1.2829465. URL: https://doi.org/10.1115/1.2829465.
[117] G. Mavrotas. "Effective implementation of the ϵ-constraint method in Multi-Objective Mathematical Programming problems". In: *Applied Mathematics and Computation*

213.2 (2009), pp. 455–465. DOI: https://doi.org/10.1016/j.amc.2009.03.037. URL: https://www.sciencedirect.com/science/article/pii/S0096300309002574.
[118] I. Das. "On characterizing the 'knee' of the Pareto curve based on normal-boundary intersection". In: *Structural Optimization* 18.2-3 (1999), pp. 107–115. DOI: 10.1007/s001580050111.
[119] M. J. D. Powell. "A fast algorithm for nonlinearly constrained optimization calculations". In: *Numerical Analysis*. Ed. by G. A. Watson. Berlin, Heidelberg: Springer Berlin Heidelberg, 1978, pp. 144–157.
[120] P. Spellucci. "A New Technique for Inconsistent QP Problems in the SQP Method". In: *Mathematical Methods of Operations Research* 47.3 (1998), pp. 355–400. DOI: 10.1007/BF01198402.
[121] J. V. Burke. "A sequential quadratic programming method for potentially infeasible mathematical programs". In: *Journal of Mathematical Analysis and Applications* 139.2 (1989), pp. 319–351. DOI: 10.1016/0022-247X(89)90111-X.
[122] B. C. Williams and J. Cagan. "Activity analysis: Simplifying optimal design problems through qualitative partitioning". In: *Engineering Optimization* 27.2 (1996), pp. 109–137. DOI: 10.1080/03052159608941402.
[123] Mathworks. *Optimization Toolbox ™ Users Guide R2020b, retrieved November 27, 2020*. 2020. URL: https://www.mathworks.com/help/optim/index.html?s_tid=CRUX_lftnav.
[124] T. J. Howard and M. M. Andreasen. "Mind-sets of functional reasoning in engineering design". In: *Artificial Intelligence for Engineering Design, Analysis and Manufacturing: AIEDAM* 27.3 (2013), pp. 233–240. DOI: 10.1017/S0890060413000231.
[125] R. Jugulum and D. D. Frey. "Toward a taxonomy of concept designs for improved robustness". In: *Journal of Engineering Design* 18.2 (2007), pp. 139–156. DOI: 10.1080/09544820600731496.
[126] T. Goto, R. Isobe, M. Yamakawa, and M. Nishida. "The New Mazda Gasoline Engine Skyactiv-G". In: *MTZ worldwide eMagazine* 72.6 (2011), pp. 40–47. DOI: 10.1365/s38313-011-0063-8. URL: https://doi.org/10.1365/s38313-011-0063-8.
[127] M. Ebro and T. J. Howard. "Robust design principles for reducing variation in functional performance". In: *Journal of Engineering Design* 27.1-3 (2016), pp. 75–92. DOI: 10.1080/09544828.2015.1103844. URL: http://dx.doi.org/10.1080/09544828.2015.1103844.
[128] K. K. Fu, M. C. Yang, and K. L. Wood. "Design principles: Literature review, analysis, and future directions". In: *Journal of Mechanical Design, Transactions of the ASME* 138.10 (2016), pp. 1–13. DOI: 10.1115/1.4034105.
[129] A. Stock. *Comparing Performance and Efficiency of Linear Motors, Ball Screws, and Rack-and-Pinion Drives.*
[130] Park-Industries. *RACK AND PINION VS BALL SCREW*. URL: https://www.parkindustries.com/blog/rack-and-pinion-vs-ballscrew/.
[131] C. A. Mcmahon. "OBSERVATIONS ON MODES OF INCREMENTAL CHANGE IN DESIGN". In: *Journal of Engineering Design* 5.3 (1994). DOI: 10.1080/09544829408907883.
[132] B. Onarheim. "Creativity from constraints in engineering design: Lessons learned at Coloplast". In: *Journal of Engineering Design* 23.4 (2012), pp. 323–336. DOI: 10.1080/09544828.2011.631904.
[133] P. Y. Papalambros and K. Shea. "Creating Structural Configurations". In: *Formal Engineering Design Synthesis*. Cambridge University Press, Nov. 2001, pp. 93–125. DOI: 10.1017/CBO9780511529627.007. URL: https://www.cambridge.org/core/product/identifier/CBO9780511529627A013/type/book_part.

[134] H. Simon. "The Science of Design: Creating the Artificial". In: *Design: Critical and Primary Sources* (2017). DOI: 10.5040/9781474282932.0013.
[135] S. Finger and J. R. Dixon. "A review of research in mechanical engineering design. Part II: Representations, analysis, and design for the life cycle". In: *Research in Engineering Design* 1.2 (1989), pp. 121–137. DOI: 10.1007/BF01580205. URL: https://doi.org/10.1007/BF01580205.
[136] M. French. *Design Principles Applied to Structural Functions of Machine Components*. 1992. DOI: 10.1080/09544829208914759.
[137] R. Radhakrishnan and D. A. McAdams. "A Methodology for Model Selection in Engineering Design". In: *Journal of Mechanical Design* 127.3 (2005), pp. 378–387. DOI: 10.1115/1.1830048. URL: https://doi.org/10.1115/1.1830048.

Appendix 1: Terminology

Unfortunately, the design science field does not have a fixed, universally accepted terminology. Terminological differences exist amongst its different fields, just as there are geographical variations as well. This PhD exists in the intersection between embodiment design and design optimization fields, where this problem is especially apparent. For one thing, embodiment design is often referred to as configuration design in the design optimization field Papalambros2017,2001FormalSynthesis,Schmidt1998. For this reason, it is necessary to provide descriptions and/or definitions for some of the most foundational terms used throughout this thesis:

Conceptual design - An early phase of product development which "involves abstracting to find the essential problems, establishing function structures, searching for working principles, combining working principles into working structures, selecting a suitable working structure and firming it up into a principle solution (concept)" (Pahl & Beitz [6]).

Concept selection - Somewhat confusingly, it is a term that is often used in design research to mean describe any situation involving the selection of a solution among a set of alternatives, although the alternatives might not strictly be different *concepts*. Thus, concept selection is used just as broadly in this thesis, being taken to mean any decision between a set of alternative solutions, e.g. the selection of a concept, of an embodiment, of a subsystem or part design, etc..

Configuration design - A term used broadly in the design optimization field [11, 12, 106], which used synonymously with embodiment design in this thesis. The definition presented by Papalambros & Wilde [12], is most pertinent: "This involves decisions on the general arrangement of parts, how they may fit together, geometric forms, types of motion or force transmission, and so on... the designer creates a new configuration through a spontaneous synthesis of previous knowledge and intuition."

Constraint Activity - "An active constraint is one which if removed, would alter the location of the optimum" (Papalambros & Wilde [12]).

Designer - Just as in Pahl and Beitz [6], the term designer is used synonymously with *design engineer*, *development enginer* or *mechanical design engineer*. In this context, the distinction made by Ahmed, Wallace, and Blessing[33] between novices (less than 5 years of industrial experience) and experienced designers is also used in this thesis.

Dependency - An interrelationship between design objectives, constraints, or objectives and constraints, determined by design variables or parameters that are shared between them. Some dependencies contribute to trade-off. In this thesis, these are refereed to as trade-off variables.

Design analysis - The use of analysis models from engineering science to describe the behaviour of a design; thus design analysis is problem specific.

Design change - Any modification made to a system during the design process. Design change can be made on different levels describing the degree of change; i.e. conceptual change, embodiment/configuration change, proportional change, parametric change, etc.

Design constraints - "All the relations among the design variables that must be satisfied for proper function of the design" (Papalambros & Wilde [12])

Design iteration - A process of gradual change, concretization, and refinement of a design. It is a slightly ill-defined concept, given that any change to design can be perceived as an iteration. In this thesis, the taxonomy introduced by Wynn and Eckert [24] will be used when relevant.

Design methodology - An overall framework for doing design (Blessing and Chakrabarti [4]).

Design objectives - The criteria for defining, evaluating and choosing the" best" solution among a set of alternatives. Depending on the level of abstraction, design objectives can thus both be general to a in a given product development process, or specific to a given concept or embodiment.

Design variable - "variables that are regarded as free because we should be able to assign any value to them. Different values for the variables produce different designs.... The number of independent design variables gives the design degrees of freedom for the problem" (Arora, [50]).

Detail design - A phase in the design process, described by Pahl and Beitz [6] as "the phase of the design process in which the arrangement, forms, dimensions and surface properties of all of the individual parts are finally laid down, the materials specified, production possibilities assessed, costs estimated, and all the drawings and other production documents produced"

Feasible domain - The design variable values that satisfy the constraints.

Early stage design - This term is used synonymously with the conceptual and embodiment design phases.

Embodiment design - Using in the meaning defined by Pahl & Beitz [6]: "During this phase, designers, starting from a concept (working structure, principle solution), determine the construction structure (overall layout) of a technical system.... Embodiment design results in the specification of a layout." [6]. Used synonymously with configuration design in this dissertation.

Modelling - Meant in the sense of constructing a mathematical model (e.g. for design analysis), or in the sense of constructing CAD models.

Parameter - Fixed properties determined by decisions made by the designer between a set of alternative options (e.g. material selection, safety factors, part shape, and so on).

Proportional design - Design activities involving the resizing of the design variables in a system.

Pareto optimal - A Pareto optimal solution is an optimum in a multi-objective problem, where one where no single objective can be improved further without worsening another. A more formal definition is given in chapter 3.

Pareto set - The set of all Pareto optimal solutions.

Trade-offs - A balancing of factors all of which are not attainable at the same time. In the context of design, this is taken to mean situations where two design objectives cannot be optimized simultaneously, meaning compromise is necessary.

Trade-off management - The Identification, quantification, root cause analysis, and reduction/elimination of trade-offs in a design.

Working Principle - As defined by Pahl & Beitz [6]: "the combination of the physical effect with the geometric and material characteristics (working surfaces, working motions and materials) allows the principle of the solution to emerge... The combination of several working principles results in the working structure."

Appendix 2: Paper A

Title: Multiobjective Monotonicity Analysis: Pareto Set Dependency and Tradeoffs Causality in Configuration Design
Authors: Sigurdarson, N.S.; Eifler, T.; Ebro, M.; Papalambros, P.Y.
Publication: ASME Journal of Mechanical Design (2022), Vol. 144, Issue 3. The appended manuscript is the post-print of the final paper.

Journal of Mechanical Design. Received December 17, 2020;
Accepted manuscript posted August 31, 2021. doi:10.1115/1.4052444

Multiobjective Monotonicity Analysis: Pareto Set Dependency and Tradeoffs Causality in Configuration Design

Nökkvi S. Sigurdarson*
Mechanical Engineering,
Technical University of Denmark,
Kgs. Lyngby, Denmark,
noksig@mek.dtu.dk

Tobias Eifler
Mechanical Engineering,
Technical University of Denmark,
Kgs. Lyngby, Denmark,
tobeif@mek.dtu.dk

Martin Ebro
Device R&D,
Novo Nordisk A/S,
Hillerød, Denmark,
mixe@novonordisk.com

Panos Y. Papalambros
Mechanical Engineering,
University of Michigan,
Ann Arbor, MI 48109,
pyp@umich.edu

Multiobjective design optimization studies typically derive Pareto sets or use a scalar substitute function to capture design trade-offs, leaving it up to the designer's intuition to use this information for design refinements and decision making. Understanding the causality of trade-offs more deeply, beyond simple post-optimality parametric studies, would be particularly valuable in configuration design problems to guide configuration redesign. This paper presents the method of Multiobjective Monotonicity Analysis to identify root causes for the existence of trade-offs and the particular shape of Pareto sets. This analysis process involves reducing optimization models through constraint activity identification to a point where dependencies specific to the Pareto set and the constraints that cause them are revealed. The insights gained can then be used to target configuration design changes. We demonstrate the proposed approach in the preliminary design of a medical device for oral drug delivery.

Nomenclature

$\mathcal{A}$ attainable set
$\mathcal{C}$ Pareto Set
$\mathbf{c}$ vector of bound objectives in the upper bound problem
D_s indices of the constraint functions that depend on a shared variable x_i
f primary objective function in the upper bound problem
$f(x^+)$ a function increasing monotonically w.r.t. x
$f(x^-)$ a function decreasing monotonically w.r.t. x
$\mathbf{F^*}$ $[k,j]$-matrix of Pareto optima
$\mathbf{E}$ $[k\text{-}1,j]$ dimensional matrix of sampled values of ε
$\mathbf{g(x)}$ vector of inequality constraints for the design problem
$\mathbf{G^*}$ matrix of$\mathbf{g}(\mathbf{x}^*)$ values stored for every run
$\mathbf{h(x)}$ vector of equality constraints for the design problem
$\mathbf{H^*}$ matrix of $\mathbf{h}(\mathbf{x}^*)$ values stored for every run
j number of computational iterations ε is sampled over
k number of objectives
n number of design variables
$\mathbf{x}$ vector of design variables
$\underline{x}$ argument of the infimum of the design problem
$\overline{x}$ argument of the supremum of the design problem
$\overline{x_i}$ monotonic trade-off variable
$\mathcal{X}$ feasible domain
$\mathcal{X}\varepsilon$ feasible domain for a given upper bound value, ε
ε k-1 dimensional vector of upper-bound parameters
ε_i upper-bound parameter for the *ith* bound objective
$\varepsilon_\mathbf{L}$ lower limit of objective bounds
$\varepsilon_\mathbf{U}$ upper limit of objective bounds
$\tilde{\varepsilon}_i$ reduced-objective variable
$\tilde{\varepsilon}^*_{i,j}$ optimal value of $\tilde{\varepsilon}_i$ implied by the activity case where the Pareto-constraint $g_j(\mathbf{x},\tilde{\varepsilon})$ bounds $\tilde{\varepsilon}_i$
λ Lagrange multiplier vector of inequality constraints

*Corresponding Author

Journal of Mechanical Design. Received December 17, 2020;
Accepted manuscript posted August 31, 2021. doi:10.1115/1.4052444

1 Introduction

Designers naturally aim to embody solutions that trade off a range of functionality and production objectives. Over time, competitive pressures require designers to improve performance while integrating more features with each new product generation [1]. Additional trade-offs arise as a result. While optimization methods are commonly used at the embodiment stage, systematic, quantitative analysis of trade-offs is less common ahead of important decisions such as concept selection, iterative redesign, or requirement setting [2]. Whether due to time constraints, task complexity, or early-stage design uncertainties, knowledge about trade-offs is largely experience-driven in design practice [3].

Pahl & Beitz [4] note that optimizing the "carrier of several combined functions" can be difficult. Yet, they also argue that decisions on what parts and subsystems contribute to different aspects of product functionality are made early on, typically during embodiment design. This is a term used somewhat interchangeably with preliminary design, configuration or topology design [5], layout [6], system design [7], or system architecture [8]. Here, we use the term configuration design for consistency with the design optimization literature. Andreasen and Howard [9] similarly argue that identifying and managing trade-offs is a key challenge in embodiment design. Different configurations will ultimately be affected by different trade-offs, a notion well accepted also in design optimization.

Multiobjective Design Optimization (MODO) techniques study what is achievable in a design subject to trade-offs, typically identifying, comparing, simplifying, and visualizing Pareto sets. Procedures for selecting a point on a Pareto set are essentially post-optimality analyses. Occasionally, they lead to converting the multiobjective problem to a scalar one with a new objective at a higher level, such as going from engineering design to design for market systems, e.g., Shiau and Michalek [10].

There is paucity in discussing *why* the result is a set rather than a dominant optimum. Yet, existing methods seem to focus only on selecting points within the set or on measures to describe how the objectives compete. Selecting a point in a Pareto set includes work on modeling preferences [11, 12, 13], identification of compromise solutions by measuring the distance to a utopia point [14], scaling methods to account for objective weighting [15], and strategies for making trade-offs aggressively or conservatively [16]. Substantial work exists for sensitivity, robustness [17], uncertainty [18], visualisation [19], dimensional reduction [20], and identification of competing objectives in a n-dimensional objective space [11]. Furthermore, structural topology optimization (TO) [21] is a notable contribution in the context of multiobjective configuration design problems. Somewhat uniquely, TO optimizes a functional representation of the design without the actual embodiment of the functions, and the results inform configuration design.

Some post-optimality analyses aim at understanding how objectives compete. Multiple measures for Pareto frontier shape exist e.g., [11, 22, 23]. Frischknecht and Papalambros [24] developed metrics to measure the alignment of objective pairs and later suggested a Pareto set analysis using local measures of objective coupling [25] to compare system topologies. Metrics describing the *quality* of a Pareto set such as hypervolume, Pareto-spread, and generational distance [26] have also been suggested to compare Pareto sets for alternative configuration designs. To tackle the comparison of multi-dimensional objective spaces, Athan and Papalambros [27] introduced the notion of meta-Pareto sets, which consist of the union of Pareto sets of multiple alternative configurations. Mattson and Messac [18] similarly put forward an approach to concept selection using s-Pareto frontier to compare alternatives.

There are three challenges with the current MODO approaches to design. First, the main focus is on optimizing a fixed design rather than questioning why the objectives compete. Second, the analysis done at earlier time points in the product's evolution may become obsolete at a later design stage. Finally, if the Pareto set contains no points acceptable to the designer, e.g., due to non-modelled considerations, there is little guidance for what to do next. A rigorous approach to gain insights into the root cause of the trade-offs inherent to the design would substantially increase the value of optimization at an early stage of product development.

Originally developed by Papalambros and Wilde [28], Monotonicity Analysis (MA) is a rigorous, yet opportunistic, method used to identify active constraints. When applicable, it allows model reduction and assessment of model boundedness, and in some cases, it reveals global optima with little or no computation. Michelena and Agogino [29] expanded the method to multiobjective problems by applying MA to a weighted sum formulation. Gobbi et al. [30] and Mastinu et al. [31] later applied MA in a procedure to derive Pareto sets analytically. Unlike the other approaches discussed above, MA can be performed prior to numerical computation and even before a full optimization model has been built. To date, MA has largely been used as a model "debugging" tool.

Monotonicity analysis is of interest here for its implications in a design context. Jain and Agogino [32] demonstrated how MA could be used to support the conceptualization of a multi-speed gearbox and explore configuration changes that lead to superior designs compared to proportional changes alone. Ishii and Barkan [33] applied MA in an interactive expert system, intending to help designers identify bottlenecks in the design caused by active constraints. Cagan and Agogino [34] also used MA to reveal previously hidden relationships through back-substitution of active constraints into objective functions. They identified ways to expand the design space to widen the search for design improvements. Deb and Srinivasan [35] meanwhile, discussed the similarities between MA and their 'innovation through optimization,' or *innovization* procedure aimed at deriving design principles using commonalities among Pareto-optimal designs. They argued that both MA and their NSGA-II-based innovization approach help identify important relationships at the optimum.

Most of this prior work [32, 34, 35] has focused on understanding the common characteristics of Pareto-optimal designs to allow reuse in future designs but within a single

configuration. Yet, if a configuration has limitations or is just not very good, one would simply find the best compromise for a poor design. If MA can identify relationships for the design variables at optimality, then arguably, it might also be able to identify relationships that *limit* optimality. In a multiobjective formulation, such analysis could lead to the discovery of the root cause for trade-offs between objectives.

In the remainder, Section 2 articulates the aims of this work, Section 3 provides some theoretical foundation for the Pareto set Dependency Analysis method developed in Section 4. Section 5 presents the methodology applied to the SOMA (Self-Orienting Millimeter-scale Applicator) drug delivery device currently in development. We offer a discussion in Section 6 and conclude in Section 7.

2 Aim of this work

Multiobjective optimization quantifies trade-offs among competing objectives. While trade-offs can be studied computationally, understanding the underlying causes is typically left to designers' ability to interpret results and to identify redesigns aimed at improving performance. Optimization has been claimed to have an intrinsic value in the design process beyond just providing numerical optimal solutions [2, 34, 36]. How to extract such design knowledge systematically is left to the designer, particularly in the early design stages. This then begs the question:

How can conceptual or configuration design limitations reflected in the Pareto set be identified rigorously? In particular, what specific design dependencies and constraints cause trade-offs?

This work seeks to demonstrate how the limitations of a design configuration may be identified through rigorous analysis rather than through tacit knowledge and heuristics alone, using novel extensions to monotonicity analysis.

To this end, we apply MA to multiobjective problems posed in the upper-bound formulation, also known as the bound objective [14] or ε-constraint method [37]. While MA is often used to check the *validity* of a model, here we demonstrate extensions to MA that allow it to be used to check the design itself when it exhibits global or regional monotonic behaviour. We use constraint activity identification and systematic reduction of the model's degrees of freedom to reveal often hidden dependencies among variables and objectives at the optimum, which cause the trade-offs. Designers will still need experience and intuition to convert this knowledge into actionable redesign decisions, but these decisions are informed by deeper understanding from rigorous analysis.

3 Theoretical Foundation

The multiobjective design optimization problem is stated in negative-null form as:

$$\text{min.} \quad \mathbf{F}(\mathbf{x}) \tag{1}$$
$$\text{subject to} \quad \mathbf{g}(\mathbf{x}) \leq 0 \tag{2}$$
$$\mathbf{h}(\mathbf{x}) = 0 \tag{3}$$
$$\mathbf{x} \in \mathbb{P} \tag{4}$$

where **F(x)** is a vector of design objectives f_i, $i = [1,2,...,k]^T$, **x** is a vector of design variables, and **h(x)** and **g(x)** are the equality and inequality constraints respectively. If $\mathcal{X}$ denotes the feasible domain, then the attainable set $\mathcal{A}$ contains all values of $\mathbf{F}(\mathbf{x})|x \in \mathcal{X}$. A point $\mathbf{F_0}(\mathbf{x}^*)$ in the attainable set $\mathcal{A}$ is said to be Pareto-optimal if and only if there exists no point in the attainable set that fulfills:

$$\mathbf{F}(\mathbf{x}) \leq \mathbf{F_0}(\mathbf{x}^*) \qquad \wedge \qquad f_i(\mathbf{x}) < f_i(\mathbf{x}^*) \tag{5}$$

The set of all Pareto-optimal points is the Pareto set $\mathcal{C}$ sitting on the boundary of the attainable set $\mathcal{A}$ [38]. There are many ways to construct the Pareto set; in the trade-off analysis that follows, we use the *upper-bound formulation* also known as the ε-*constraint method* (see [14] for a methods overview).

Monotonicity Analysis [28] leverages any existing monotonic behavior of objective and constraint functions to check for boundedness and identify constraint activities thus reducing the problem's degrees of freedom. A function is said to be monotonically increasing with respect to a variable x if $f(x_2) > f(x_1)$ for any $x_2 > x_1$. This monotonic relationship between f and x is denoted $f(x^+)$. Correspondingly, a function is monotonically decreasing with respect to x if $f(x_2) < f(x_1)$ for any $x_2 > x_1$, and denoted $f(x^-)$. In the presence of monotonicity, the following principles [36] can be exploited in single-objective problems to identify activity of certain constraints, without computing the optimum first:

First monotonicity principle (MP1)
In a well-constrained minimization problem, every increasing variable is bounded below by at least one non-increasing active constraint.

Second monotonicity principle (MP2)
In a well-constrained minimization problem, every nonobjective variable is bounded both below by at least one non-increasing semi-active constraint and above by at least one non-decreasing semi-active constraint.

Constraint activity means that the location of the optimum is altered if the constraint is deleted. Active inequality constraints will be satisfied as strict equalities at the optimum, thus reducing the degrees of freedom accordingly. By identifying active constraints, one can solve the constraint functions with respect to (w.r.t.) one of their dependent variables and substitute the solution for that variable into the remaining constraint functions and objectives, thereby eliminating the active constraints and the substituted variables.

3.1 Modelling and Computation

As mentioned, multiobjective MA was originally demonstrated using a weighted-sum formulation [29]. For the trade-off analysis method development that follows in section 4, we use the upper-bound formulation [14]. This formulation involves converting the problem in Eq. 1-4 into

a single objective one:

$$\begin{aligned} \text{min.} \quad & f(\mathbf{x}) && (6)\\ \text{s.j.t} \quad & \mathbf{c}(\mathbf{x};\boldsymbol{\varepsilon}) \leq 0 && (7)\\ & \mathbf{g}(\mathbf{x}) \leq 0 && (8)\\ & \mathbf{h}(\mathbf{x}) = 0 && (9)\\ & \mathbf{x}, \boldsymbol{\varepsilon} \in \mathbb{P} && (10) \end{aligned}$$

In this formulation, originally put forward by Carmichael [37], $\mathbf{c}(\mathbf{x};\boldsymbol{\varepsilon})$) is a $k-1$ dimensional vector of *bound objectives* expressed in the form $c_i(\mathbf{x},\varepsilon_i) = f_{i+1}(\mathbf{x}) - \varepsilon_i \leq 0$ or $c_i(\mathbf{x},\varepsilon_i) = \varepsilon_i - f_{i+1}(\mathbf{x}) \leq 0$, $i = [1,2,..,(k-1)]$. The vector $\boldsymbol{\varepsilon}$ of parameters ε_i represents the upper bounds of the bound objectives. When $f(\mathbf{x})$ is minimised for given values of ε_i, then the solution $\mathbf{x}^*$ is Pareto optimal if all of the bound objectives are active with non-zero Lagrange multipliers. Pareto points are thus identified by varying $\boldsymbol{\varepsilon}$ systematically between lower $\boldsymbol{\varepsilon}_\mathbf{L}$ and upper limits $\boldsymbol{\varepsilon}_\mathbf{U}$. See [14, 37, 39] for an overview of works on the upper bound formulation, the underlying mathematics, and approaches to defining suitable limits for $\boldsymbol{\varepsilon}$. The Pareto set is constructed by sampling a set of $\boldsymbol{\varepsilon}$ parameter values

$$\mathbf{E} = (\boldsymbol{\varepsilon}_\mathbf{U} - \boldsymbol{\varepsilon}_\mathbf{L})\mathbf{R} + \boldsymbol{\varepsilon}_\mathbf{L} \quad (11)$$

where $\mathbf{R}$ is a matrix of uniformly distributed quasi-random numbers between 0 and 1 of the dimension [k-1; j], where j is the number of computational iterations, and k is the number of objectives. A low discrepancy quasi-random set (e.g., a Halton set) can be used to reduce bias in $\mathbf{R}$ to reduce the computational cost of achieving a Pareto set with low sparsity. After sampling, the optimization problem is solved iteratively, as in the following pseudo-code:

for i = 1..j **do**
  Set upper bound on constrained objectives, $\boldsymbol{\varepsilon} = \mathbf{E}(:,i)$
  Solve optimization problem w.r.t $\boldsymbol{\varepsilon}$
  Store optimum, $\mathbf{F}^*(:,i) = [f^*, \boldsymbol{\varepsilon}^T]^T$
  Store arguments, $\mathbf{X}^*(:,i) = \mathbf{x}^*$
  Store Lagrange multipliers, $\Lambda(:,i) = \lambda$
  Store constraint values ,$\mathbf{G}^*(:,i) = \mathbf{g}(\mathbf{x}^*)$ and $\mathbf{H}^*(:,i) = \mathbf{h}(\mathbf{x}^*)$
end for

The sparsity of the approximated Pareto set decreases as j increases, while the span increases with j and the difference between ε_U and ε_L. With an increased j, one identifies more Pareto points resulting in a more dense Pareto set. A high j can be necessary to approximate the shape of the Pareto set, should it have interactions between the objectives that exist locally in the attainable set, for instance creating knee like shapes [23]. Beyond a certain limit, the Pareto set will have been exhaustively constructed, meaning no additional feasible solutions can be found by further increasing the difference between ε_U and ε_L. Thus, one can also solve the MODO problem multiple times with a relatively low j, increasing the difference between ε_U and ε_L, until the boundaries of the Pareto-set seem to have been identified, and then subsequently increasing j to the desired level of density.

As discussed in [14], the $\boldsymbol{\varepsilon}$-constraint formulation does have certain limitations. It results in the identification of non-optimal solutions when the bound objectives are inactive, computational iterations are "wasted" on values of $\boldsymbol{\varepsilon}$ that lie between the Pareto set and the Utopia point, and it might only identify local optima in non-convex attainable sets. In situations where the problem $\mathcal{A}$ is non-convex or computationally expensive, one could use another formulation for numerical solution and only use the $\boldsymbol{\varepsilon}$-constraint formulation in pre- and post-optimality analysis. One could also rely on one of the many implementations of the $\boldsymbol{\varepsilon}$-constraint method that have addressed these limitations (e.g. AUGMECON by Mavrotas [39]). However, the aim of this contribution is to identity the properties that affect the optimum rather than to identify the optimum itself efficiently. As such, we will forego further treatment of specific implementations and computational efficiency and use the general problem form in eqs. 6-10 due to its benefits in MA:

1. *Maintaining monotonic properties*: Converting a set of objectives into a composite function, e.g., a weighted-sum, can result in loss of monotonic properties when the objectives share variables. Using an upper-bound formulation avoids this issue.
2. *Objective elimination*: Introducing objectives as constraints in an optimization model allows one to parametrically study the *activity* of the bound objective across the attainable set using monotonicity analysis. If a bound objective can be determined to be active through monotonicity analysis, the objective itself can be 'optimized out' of the model through back-substitution [36], revealing how the objectives affect each other at the Pareto frontier.
3. *Sensitivity data*: Solving a constrained optimization problem yields non-zero Lagrange multipliers for active constraints, revealing the local sensitivity of the optimum w.r.t. changes in each active constraint. In the upper-bound formulation, the Lagrange multipliers of the bound objectives describe whether and to which degree the bound objectives compete with the primary objective, which some term the *trade-off ratio* [40].

It is often suggested that the most important objective should be modelled as the function being minimised [14], while the remaining objectives should be bound. To simplify monotonicity analysis, however, the most suitable approach would be to select the objective with the largest number of design variables. Doing so allows the broadest application of *MP1* in problem reduction.

4 Pareto Set Dependency Analysis

This section develops novel theory for the systematic reduction of multiobjective problems (Subsection 4.1) and the analysis of the relationships that *bound* the Pareto set (Subsection 4.2). We then use these developments to define an overall analysis procedure that allows the identification of

the dependencies between the objectives and constraints that create the Pareto set (Subsection 4.3). These developments, collectively referred to as Pareto Set Dependency Analysis, are demonstrated on algebraic models. The methods could, in principle, also be applied to meta-models and numerical models through either computational experiments or implicit model reduction [36].

4.1 Multiobjective Monotonicity Analysis (MOMA)

The reasoning behind the desire to develop a systematic approach to multiobjective monotonicity analysis is as follows. Consider that the Pareto set $\mathcal{C}$ exists on the boundary of the attainable set $\mathcal{A}$ but is not necessarily defined by the constraints alone, as unconstrained multiobjective problems also yield Pareto sets [38]. It follows that the occurrence of Pareto sets must have two causes:

1. *Trade-off variables* In negative-null form, a variable x that influences two objectives, $f_1(x)$ and $f_2(x)$, causes a trade-off if $\arg\min f_1(x) \neq \arg\min f_2(x)$. In design, this mostly occurs when an objective pair is oppositely monotonic w.r.t. a variable, either globally or regionally.
2. *Active constraints* Active constraints reduce the degrees of freedom (DOF) in optimization problems, affect the feasible domains for the remaining DOF, and change the optimum. Eliminating and back-substituting active constraints into objective functions can introduce new variables to the expression, and can change its monotonicity w.r.t. the original variables, revealing additional trade-off variables *hidden* in constraints.

Hence, multiobjective monotonicity analysis (MOMA) may allow the systematic identification of trade-off variables and reveal relationships between the objectives at the optimum that are *hidden* by constraints. This can help designers understand the root causes of trade-offs in a configuration design. Demonstrating this requires certain extensions of MA to deal with multiobjective problems.

4.1.1 Definitions and Theorems

For upper-bound formulations, the extension of MA into multiple objectives is relatively straightforward as this merely involves handling more constraints. The principles and procedures originally developed by Papalambros and Wilde [36] mostly still apply. The exception is that the bound objectives, $\mathbf{c}(\mathbf{x};\boldsymbol{\varepsilon})$, cannot be treated as traditional inequality constraints. Firstly, as we wish to vary the upper-bound values, $\boldsymbol{\varepsilon}$, these cannot be regarded as fixed parameters when performing monotonicity analysis. Secondly, we seek to partially minimize the bound objectives, which has implications for the use of MP1 and MP2. Hence, it is necessary to introduce some theorems of relevance to how **c** is handled:

Definition 1 Trade-off Variables
If an objective pair f and c_i have a variable x_1 in common, but differ in monotonicity w.r.t. x_1, e.g., $f(x_1^+)$ and $c_i(x_1^-)$, then x_1 is said to be a trade-off variable, denoted $\underline{x}_1$. Correspondingly, an objective pair of like monotonicity w.r.t. a common variable, indicates that the variable is harmonious and can be used to partially minimise both simultaneously.

Theorem 1 Influence of Monotonic Trade-off Variables
In the presence of monotonic trade-off variables, no dominant minimum exists, resulting in a Pareto set. The proof for this is a corollary to MP1.

Proof. Let f_1 be monotonically increasing w.r.t. $x \in \mathbb{P}$ and f_2 monotonically decreasing, and let x be well bounded from above and below. Then by MP1, $\arg\min \mathrm{f_1(x)} = \underline{x}$, and $\arg\min \mathrm{f_2(x)} = \bar{x}$, meaning that the minimizers for the two objectives are defined by the *greatest lower bound* (glb)and the *lowest upper bound* (lub) respectively. Hence any feasible value of x will yield a unique Pareto point. ■

Corollary 1.1 Boundedness of trade-off variables
Following Theorem 1, multiobjective problems can only be said to be well-bounded if all trade-off variables are bounded from above and below.

For instance, if a bound objective, c_i, is critical w.r.t. a monotonic trade-off variable, $\underline{x}_1$, then the multiobjective problem is not well bounded, as $\underline{x}_1 \to \infty$ or $\underline{x}_1 \to 0$ when $\varepsilon_i \to \infty$ and f is minimised. This can either be handled by introducing additional constraints, or by selecting suitable limits for the upper-bound problem $\boldsymbol{\varepsilon}_\mathbf{L}$, $\boldsymbol{\varepsilon}_\mathbf{U}$.

In upper-bound formulations, we treat objectives as additional constraints and iteratively identify Pareto points, exploring $\bar{\mathbf{x}} \in \mathcal{X}$, for different values of $\boldsymbol{\varepsilon}$, as illustrated in Figure 1. If a bound objective is active, the model is essentially exploring a smaller region of the feasible domain $\mathcal{X}_\varepsilon \in \mathcal{X}$. From this, an additional theorem arises:

Theorem 2 Activity of Bound objectives
A bound objective $c_i(\underline{x};\varepsilon_i)$ can either be active, semi-active, dominated, or inconsistent with other constraints, depending on the value of ε_i. The change in activity of $c_i(\underline{x};\varepsilon_i)$ across $\mathcal{A}$ affects the shape of the Pareto set.

Consider an objective pair, $f_1(x_i^+)$ and $c_1(x_i^-,\varepsilon_1)$, with the design variable x being bounded from below by $g_1(x_i^-)$ and from above by $g_2(x_i^+)$, where ε is the upper bound parameter. Here, the value of ε determines constraint activity:

1. For the values of ε_1 where $g_1(x_i) < c_1(x_i)$, c_1 is active, and the result will be Pareto-optimal.
2. For the values of ε_1 where $c_1(x_i) < g_1(x_i)$, c_1 is inactive, and the result will not be Pareto-optimal
3. For the values of ε_1 where $g_2(x_i) < c_1(x_i)$, $\mathcal{X}_\varepsilon \in \emptyset$, and thus these constraints are inconsistent. In this case, g_2 shapes a boundary of the Pareto set.
4. For the value of ε_1 where $c_1(x_i) = g_1(x_i)$, the bound objective is semi-active, yielding the single-objective optimum for f_1. Correspondingly, $c_1(x_i) = g_2(x_i)$ yields the single-objective optimum for f_2.

Thus, exploring these changes in the activity of c_1 yields the Pareto set for the objective pair. We can hence utilise MOMA to identify the conditions under which a bound objective is active, dominated, or inconsistent. This can reveal

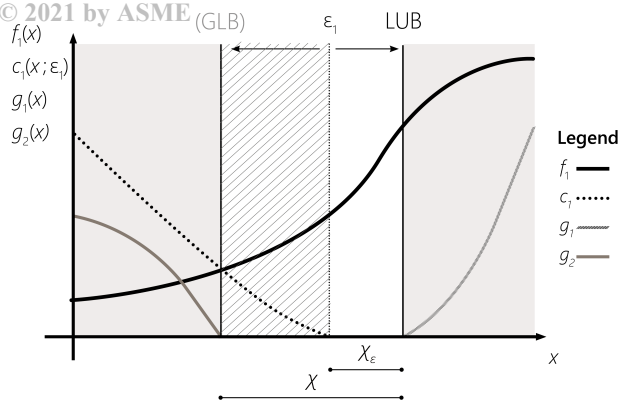


Fig. 1: MOMA allows the partial identification of the Pareto set, by identifying the values of ε where the bound objectives are active, semi-active, violated, and inconsistent

important relationships between the objectives and the constraints g_i that affect the Pareto set. Here, it is important to consider the the influence of ε on the activity of $\mathbf{g}(\mathbf{x})$:

Definition 2 Global Activity
In the monotonicity analysis of an upper-bound problem, a constraint $g_i(\mathbf{x})$ is said to be globally active if and only if $f(\mathcal{X}_i) < f(\mathcal{X}_*)$ for any $\{\varepsilon \in \mathbb{P} \,|\, \varepsilon_L \leq \varepsilon \leq \varepsilon_U\}$.

Trade-off variables can only be optimized out if an active bound objective is used to eliminate it or if the bound objective can be determined to be dominated w.r.t. said trade-off variable by another globally active constraint. This notion of global activity is central to multiobjective monotonicity analysis. A reduced model would potentially only identify parts of the Pareto set if we were to optimize variables out with constraints that are not globally active.

The final extension to MA that is necessary in order to deal with multiobjective problems is the question of how to partially minimise several objectives concurrently:

Definition 3 Partial minimisation of bound objectives
In a well-constrained multiobjective, upper-bound minimization problem, any increasing objective variable not in the primary objective, is bounded below by at least one non-increasing active constraint.

Modelling objectives as constraints is merely a route to identifying Pareto points. It is still desirable to identify partial minima for bound objectives. By simply extending MP1 into multiobjective problems, we can reduce multiple objectives, i.e., identify partial minima for f_{i+1} in $c_i(\mathbf{x}, \varepsilon_i) = f_{i+1}(\mathbf{x}) - \varepsilon_i \leq 0$. Nevertheless, it is necessary to take particular care in this process. Unless it is certain that the optimal value of a given variable is the same for all objectives, i.e., $\arg\min f_i(x) = \arg\min f_j(x)$ for any i and j, optimizing the variable out would result in a model that does not describe the entire Pareto set. When a globally active constraint can be identified, the bound objectives can always be partially minimized. This is relatively straightforward to do in situations where the condition $\arg\min f_i(x) = \arg\min f_j(x)$ for any i and j is upheld by definition. Following MP1, harmonious variables and critically constrained variables [36] will always meet this condition. As will variables that are bound by constraints that only depend on harmonious variables or on variables that only influence one objective, because constraint activity will be unaffected by the values of ε.

4.1.2 Impact of constraint activity in multiobjective problems

With these definitions, we can apply MA to multiobjective problems and, in doing so, identify trade-off variables that may be *hidden* in constraints. Here, it is beneficial to note the impact on the objective functions. There are two situations of relevance to trade-off analysis; when an objective changes monotonicity w.r.t a variable, or when it becomes dependant on new variables. Consider an example:

$$\begin{aligned}
\text{min.} \quad & f_1(x_1,x_2,x_3) = x_1^2 - x_2 + x_3 && (12)\\
& f_2(x_2,x_4,x_5) = \frac{1}{x_2} - x_4^2 + 2x_5 && (13)\\
\text{s.j.t} \quad & 2x_4 - x_1 \leq 0 && (14)\\
& x_2^2 + 4x_2 - 2x_3 \leq 0 && (15)\\
& x_2^3 + 2x_4 \leq P_1 && (16)\\
& 10 - 3x_5 \leq x_5^2 && (17)\\
& x \in \mathbb{P} && (18)
\end{aligned}$$

Without inspection of the influence of the constraints, it would seem there is no trade-off between f_1 and f_2, as they are both monotonically decreasing w.r.t the only shared variable, x_2. Yet, when converted into an upper-bound formulation, monotonicity analysis reveals hidden dependencies:

$$\begin{aligned}
\text{min.} \quad & f_1(x_1^+,x_2^-,x_3^+) = x_1^2 - x_2 + x_3 && (19)\\
\text{s.j.t} \quad & c_1(x_2^-,x_4^-,x_5^+;\varepsilon_1^-) = \frac{1}{x_2} - x_4^2 + 2x_5 - \varepsilon_1 \leq 0 && (20)\\
& g_1(x_1^-,x_4^+) = 2x_4 - x_1 \leq 0 && (21)\\
& g_2(x_2^+,x_3^-) = x_2^2 + 4x_2 - 2x_3 \leq 0 && (22)\\
& g_3(x_2^+,x_4^+) = x_2^3 + 2x_4 - P_1 \leq 0 && (23)\\
& g_4(x_5^-) = 10 - x_5^2 - 3x_5 \leq 0 && (24)
\end{aligned}$$

where f_2 has been converted into a bound objective $c_1(\mathbf{x}, \varepsilon_1)$. Following MP1, it is clear that g_1 and g_2 are critical w.r.t. x_1 and x_3, respectively, for any value of ε_1, and are therefore active. Following Definition 3, we also conclude that g_4 is active as it is critical for x_5, meaning we partially minimize f_2 in c_1 by optimizing x_5 out. Solving for the minimizers yields $x_1^* = 2x_4$, $x_3^* = \frac{1}{2}x_2^2 + 2x_2$, and $x_5^* = 2$. With back-substitution, a reduced problem is reached:

$$\begin{aligned}
\text{min.} \quad & f_1(x_2^+,x_4^+) = 4x_4^2 + \frac{1}{2}x_2^2 + x_2 && (25)\\
\text{s.j.t} \quad & c_1(x_2^-,x_4^-;\varepsilon_1^-) = \frac{1}{x_2} - x_4^2 + 4 - \varepsilon_1 \leq 0 && (26)
\end{aligned}$$

Journal of Mechanical Design. Received December 17, 2020;
Accepted manuscript posted August 31, 2021. doi:10.1115/1.4052444

$$g_3(x_2^+, x_4^+) = x_2^3 + 2x_4 - P_1 \leq 0 \quad (27)$$

Here, f_1 has changed monotonicity w.r.t. x_2 and now depends on x_4, being oppositely monotonic to the bound objective c_1. Following Theorem 1, both x_2 and x_4 are trade-off variables, meaning that there is no single solution to the optimization problem but rather a Pareto set. Considering Corollary 1.1 the problem is, in fact, asymptotically bounded, as x_2 and x_4 are unbounded from below unless a well defined upper limit is imposed on ε_1. Hence, c_1 is globally active.

While this example may seem simplistic, it demonstrates the shifts in dependency between objectives that occur in the presence of active constraints. Such relationships are not necessarily easy to spot in non-reduced optimization models, nor is it given that the designer is aware of them. As such, monotonicity analysis can be used to identify trade-off variables, and in doing so, reveal what constraints in a design cause a lack of objective alignment - in this case, g_1 and g_2, as they introduce trade-off variables into the problem.

4.2 ε-Monotonicity Analysis

With the theoretical developments introduced so far, one can apply monotonicity analysis to systematically reduce multiobjective models, gradually converging towards an explicit description of the Pareto set while identifying trade-off variables in the process. When all globally active constraints have been identified, one can optimize the active bound objectives out of the model. If one determines that $c_j(\mathbf{x};\varepsilon_j) \equiv 0$, and subsequently optimizes a trade-off variable $\underline{x_i}$ out, then $f(\mathbf{x})$ and $g(\mathbf{x}), c_i(\mathbf{x};\varepsilon) \in D_s(x_i), i \neq j$ become dependent on ε_j through back-substitution. A parameter from an eliminated bound objective will be denoted as $\tilde{\varepsilon}_j$ and treated as a variable, referred to as the *reduced-objective variable*.

The reasoning behind treating $\tilde{\varepsilon}_j$ as a variable is twofold. Firstly, the primary objective function has been transformed into a bi-objective function, $f(\mathbf{x}, \tilde{\varepsilon}_j)$, describing the trade-off between the primary objective, $f(\mathbf{x})$ and $\tilde{\varepsilon}_j$. Secondly, the feasible values of $\tilde{\varepsilon}_j$ are now determined by a set of constraints $\mathbf{g}(\mathbf{x}, \tilde{\varepsilon}_j)$. The bi-objective Pareto front between f_1 and f_{j+1} will thus be defined by $f(\mathbf{x}, \tilde{\varepsilon}_j)$ and $\mathbf{g}(\mathbf{x}, \tilde{\varepsilon}_j)$. Meanwhile, the trade-offs amongst the eliminated objectives themselves are expressed through $\mathbf{g}(\mathbf{x}, \tilde{\varepsilon})$, henceforth referred to as *Pareto-constraints*. This means that if we treat $\tilde{\varepsilon}_j$ as a variable, identifying the constraints that bound it can be used to better understand the cause of the shape of the Pareto set.

In principle, all active bound objectives can be eliminated from the model. This will result in a multiobjective expression $f(\mathbf{x}, \tilde{\varepsilon})$ describing the trade-off between the primary objective and all others, while all the Pareto-constraints $\mathbf{g}(\mathbf{x}, \tilde{\varepsilon})$ describe the trade-offs between the eliminated objectives. However, it may not always be beneficial to do so, for instance, when elimination results in a loss of monotonic properties or when explicit elimination becomes too time-consuming. To allow the furthest reduction of the model, it is beneficial to attempt to eliminate the trade-off variables that are shared between the largest number of constraints.

What remains after objective reduction is:

$$\begin{aligned} \text{min.} \quad & f_1(\mathbf{x}, \tilde{\varepsilon_1}, ..., \tilde{\varepsilon}_{k-1}) && (28)\\ \text{s.j.t} \quad & \mathbf{g}(\mathbf{x}, \tilde{\varepsilon}) \leq 0 && (29)\\ & \mathbf{h}(\mathbf{x}) = 0 && (30) \end{aligned}$$

where $f_1(\mathbf{x}, \tilde{\varepsilon_i}^+)$ or when $\tilde{\varepsilon_i}$ is a maximisation objective, and $f_1(\mathbf{x}, \tilde{\varepsilon_i}^-)$ when $\tilde{\varepsilon_i}$ is a minimisation objective. Applying monotonicity analysis to this formulation thus allows the identification of active Pareto-constraints at the single objective optimum, f_1^*. Solving for $\tilde{\varepsilon_i}^*$ would then yield an explicit description of the relationship between the remaining design variables, and $\tilde{\varepsilon_i}$ at a single Pareto point. Subsequent back-substitution reveals how influential the trade-off with $\tilde{\varepsilon_i}$ is upon f_1^*. To study the whole Pareto set, however, a symbolic cost function $U(f_1, \tilde{\varepsilon})$ is introduced; $U(f_1, \tilde{\varepsilon})$ is monotonically increasing w.r.t. minimization objectives and decreasing w.r.t. maximization objectives:

$$\begin{aligned} \text{min.} \quad & \mathrm{U}(f_1^+, \tilde{\varepsilon_1}^+, ..., \tilde{\varepsilon}_{k-1}^-) && (31)\\ & f_1(\mathbf{x}, \tilde{\varepsilon_1}^-, ..., \tilde{\varepsilon}_{k-1}^+) && (32)\\ \text{s.j.t} \quad & \mathbf{g}(\mathbf{x}, \tilde{\varepsilon}) \leq 0 && (33)\\ & \mathbf{h}(\mathbf{x}) = 0 && (34) \end{aligned}$$

In minimizing cost, we can exploit its inherent monotonicity w.r.t. the objectives to identify the constraints that *bound* $\tilde{\varepsilon}$, and hence affect the topology of the Pareto set. Thus MP1 can be employed to derive the following theorem:

Theorem 3 Boundedness of $\tilde{\varepsilon_i}$
In a reduced multiobjective problem, the single objective optimum of a minimisation objective, $\tilde{\varepsilon_i}$, is determined by its greatest lower bound. Correspondingly, the lowest upper bound determines the nadir of $\tilde{\varepsilon_i}$. As such, the span of the Pareto set is in part determined by $\mathcal{X}(\tilde{\varepsilon})$.

Essentially, each reduced-objective variable is bounded by one or more Pareto-constraints across the objective space. Beyond simple optimization models, they are not necessarily critically constrained. Rather, the optimization of one $\tilde{\varepsilon}_i$ will affect the constraints of another, $\tilde{\varepsilon}_j$, if their respective glb/lub share variables, or depend on multiple $\tilde{\varepsilon}$.

Theorem 4 Conditional Activity of Pareto Constraints
In a set of Pareto-constraints that are conditionally critical for $\tilde{\varepsilon_i}$, any constraint, $g_i(\mathbf{x}, \tilde{\varepsilon})$, will at least be semi-active w.r.t. $\tilde{\varepsilon_i}$ somewhere in the objective space, if it is dependant on $\underline{x}$ or more than one reduced-objective variable. That is, unless there exists a Pareto constraint g_j such that $g_i(\mathbf{x}, \tilde{\varepsilon}) < g_j(\mathbf{x}, \tilde{\varepsilon}) \leq 0$ for any feasible value of $\tilde{\varepsilon}$.

The implication here is that changes in constraint activity can occur across the Pareto set if no $\tilde{\varepsilon_i}$ is critically constrained, and no Pareto-constraint is dominant. Identifying these changes in activity reveals how the objectives interact, as exemplified in Figure 2. Pareto-constraints can take on several forms, that shape the Pareto set in different ways:

- *Bound shift*: A Pareto constraint can for example shift the extremum of a monotonic variable, in effect making it a trade-off variable. Consider a problem where $f_1(x_1^+, x_2^-, \tilde{\varepsilon}_1^-)$, and one of the constraints is $g_i(x_2^+, \tilde{\varepsilon}_1^-) \equiv 0$. As $\tilde{\varepsilon}_1 \to 0$, the lub of x_1 shifts downward, worsening the optimum of f_1. Thus, g_i makes x_1 a trade-off variable w.r.t. f_1 and $\tilde{\varepsilon_1}$, with $\text{argmin}\{\tilde{\varepsilon_1}, x_2 \in \mathcal{X}\} = \underline{x_2}$.
- *Inconsistency by* ε: Pareto constraints can narrow the feasible domain of design variables that are bounded from above and below. Consider a problem with $U(f_1^+, \tilde{\varepsilon_1}^+, \tilde{\varepsilon_2}^-)$ where a variable x_1 is bounded from above by $g_1(x_1^+, \tilde{\varepsilon}_1^-) \leq 0$ and from below by $g_2(x_1^-, \tilde{\varepsilon}_2^+) \leq 0$. As $\tilde{\varepsilon} \to \tilde{\varepsilon}^*$, the feasible domain for x is reduced, meaning g_1 and g_2 become inconsistent beyond the Pareto set. Hence, g_1 and g_2 reduce objective alignment between $\tilde{\varepsilon_1}$ and $\tilde{\varepsilon_2}$, with one becoming semi-active at the resulting bi-objective Pareto frontier.
- *Multiple objectives*: Pareto constraints that depend on multiple $\tilde{\varepsilon_i}$ drastically reduce objective alignment, for instance if a constraint takes the form $g_1(\mathbf{x}, \tilde{\varepsilon}_1^+, \tilde{\varepsilon}_2^-)$.

Hence, trade-offs between the reduced-objectives are apparent in the Pareto-constraints themselves. An objective pair, $\tilde{\varepsilon_i}$ and $\tilde{\varepsilon_j}$, is in trade-off if they share a constraint of the form $g(\mathbf{x}, \tilde{\varepsilon_i}, \tilde{\varepsilon_j})$ or if their constraints become inconsistent w.r.t. to a shared variable, x, when $\tilde{\varepsilon} \to \tilde{\varepsilon}^*$. Such constraints therefore require special attention.

4.3 Analysis Procedure

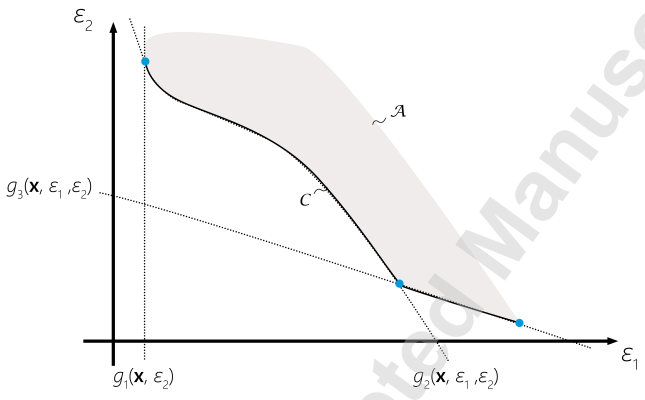


Fig. 2: An example of how the topology of a Pareto set is affected by constraints. Here the optima of ε_1 and ε_2 are determined by g_1 and g_3 respectively, with the multiobjective Pareto constraint, g_2 further reducing objective alignment

Applying the MOMA and ε-monotonicity theorems to multiobjective optimization problems allows systematic reduction down to a point where the dependencies that exist in the Pareto set are revealed. The root causes of these dependencies are, from a design perspective, the constraints and shared variables that create said dependencies. Thus, if we systematically reduce multiobjective problems and make a note of trade-off variables, the constraints that introduce them, and the constraints that bound the Pareto set, we find the relationships that in effect create, shape, and position the Pareto set. The steps in the required analysis process, which builds upon monotonicity analysis as developed by Papalambros and Wilde [36], are as follows:

1. Model the multiobjective problem as an upper-bound formulation in negative-null form.
2. Set up a monotonicity table (see e.g. [36, 41]) and assess the monotonicity of the objectives and constraints w.r.t. their variables. Make a note of any trade-off variables.
3. Use monotonicity analysis procedures to assess whether the model is well bounded [36], with the addition of the special case of the well-boundedness of trade-off variables. If the model is not well bounded, add constraints.
4. Identify constraints that are active w.r.t the primary objective and use them to reduce the model. Make a note of constraints that introduce new trade-off variables. If possible, identify the conditions under which the bound objectives become active, following Theorems 1 and 2.
5. Partially minimize the bound objectives when no further reductions to the primary objective can be made. Take care not to use constraints that potentially bound other variables regionally in the objective space. Make a note of constraints that introduce new trade-off variables.
6. When the remaining variables are either trade-off variables, non-monotonic or bounded by a conditionally critical set of constraints, run the optimization model.
7. If the numerical results reveal further globally active constraints, make further model reductions.
8. If any bound objectives are globally active, optimize said objectives out, eliminating trade-off variables in the process. The ε parameters will now appear in the remaining constraints and objective functions.
9. Treat ε parameters of the eliminated bound objectives as variables and identify the constraints that bound them. In the presence of conditional critical Pareto constraints, decompose the problem into *Pareto-Optimal Activity Cases* (see Table 1). Identify the values of ε that cause change in constraint activity or make specific constraints inconsistent. Verify this against the numerical results.

Following Theorem 4, the bounds of $\tilde{\varepsilon}$ can be interdependent, meaning that the minimisation of $\tilde{\varepsilon_i}$ affects the bounds of the remaining $\tilde{\varepsilon_j}, \forall j \neq i$, and $\underline{x}$, causing changes in activity across the Pareto set. Each change in activity implies regional dependencies between the objectives in regions of the Pareto set, as illustrated in Fig. 2. Each potential combination of active Pareto constraints hence represents a unique *Pareto Efficient Activity Case*. One can either exhaustively study all cases or focus the analysis procedure upon cases of interest. The case analysis procedure, demonstrated on a problem with minimisation objectives, is shown in Table 1. It closely resembles the parametric solution procedure developed by Wilde [36], albeit for objectives instead of design parameters. The analysis of three Pareto optimal activity case is demonstrated in Section 5.4 as a part of the case study.

Table 1 - Analysis of Pareto-optimal Activity Cases

Step 1 - Case identification

To minimise $\mathrm{U}(f_1^+, \tilde{\boldsymbol{\varepsilon}})$, identify the conditionally critical set of Pareto constraints for each $\tilde{\varepsilon}_i$. For each $g_j(\mathbf{x}, \tilde{\boldsymbol{\varepsilon}}_i)$ that is conditionally critical w.r.t. $\tilde{\varepsilon}_i$:

1. Assume $g_j(\mathbf{x}, \tilde{\boldsymbol{\varepsilon}}_i) \equiv 0$, and solve w.r.t. $\tilde{\varepsilon}_i$.
2. Identify the constraints that become active as a consequence of $\tilde{\varepsilon}_i \to \tilde{\varepsilon}_i{}^* \wedge g_j(\mathbf{x}, \tilde{\boldsymbol{\varepsilon}}_i) \equiv 0$ and use this to reduce the expression $\tilde{\varepsilon}_i{}^*(\mathbf{x}, \tilde{\boldsymbol{\varepsilon}})$.
3. Back substitute the eliminated variables into the remaining constraints, including the Pareto constraints that bound other reduced objective variables. If possible, identify the glb and lub of $\tilde{\varepsilon}_l, \forall l \neq i$ and use it to solve for $\underline{\tilde{\varepsilon}_l}$.

Step 2 - Case elimination

Compare the terms for $\tilde{\varepsilon}_i{}^*$ from each case:

1. If any case j is dominant, i.e. $\tilde{\varepsilon_{i,j}}{}^* > \tilde{\varepsilon_{i,k}}{}^*$ for any feasible value of $\mathbf{x}$ and $\tilde{\varepsilon}_l, \forall l \neq i$, then $g_k(\mathbf{x}, \tilde{\boldsymbol{\varepsilon}})$ is either inactive or bounds another variable.
2. If any variable is revealed to be unbounded as a consequence of $g_j(\mathbf{x}, \tilde{\boldsymbol{\varepsilon}}_i) \equiv 0$, then the problem is either not well-constrained, or g_j is never critical w.r.t. $\tilde{\varepsilon}_i$, meaning the case can be disregarded.
3. Identify the conditions under which the remaining cases become active. If feasible values of $\mathbf{x}$ and $\tilde{\varepsilon}_l, \forall l \neq i$ exist such that two cases become equivalent, i.e. $\tilde{\varepsilon_{i,j}}{}^* = \tilde{\varepsilon_{i,k}}{}^*$ then g_j and g_k are regionally active in the objective space, with a change in activity occurring at $\tilde{\varepsilon_{i,j}}{}^* = \tilde{\varepsilon_{i,k}}{}^*$. Such points are vertices of the Pareto set.

Step 3 - Case reduction

Reduce the remaining cases further to identify the extrema of the Pareto set:

1. Further minimise $\tilde{\varepsilon}_i{}^*(\mathbf{x}, \tilde{\boldsymbol{\varepsilon}})$ by optimizing trade-off variables out, letting $\overline{\underline{x}} \to \{\underline{x} \text{ if } \tilde{\varepsilon}_i(x^+), \overline{x} \text{ if } \tilde{\varepsilon}_i(x^-)\}$. If the glb and lub of $\overline{\underline{x}}$ cannot be determined, the problem case can be split into sub-cases.
2. If possible, identify the cases that yield utopia and nadir points for each objective

Beyond potentially deriving single-objective optima, this procedure can be used to explicitly derive trade-off functions of the forms $f_1(\mathbf{x}, \tilde{\boldsymbol{\varepsilon}})$ and $\tilde{\varepsilon}_i^*(\mathbf{x}, \tilde{\boldsymbol{\varepsilon}})$. As a consequence of Theorem 1, these equations actually describe the Pareto set prior to the elimination of $\overline{\underline{\mathbf{x}}}$, as any feasible value of a monotonic trade-off variable yields a Pareto point. If an objective pair $\tilde{\varepsilon}_i$ and $\tilde{\varepsilon}_j$ is in trade-off, then these reduction steps will inevitably yield minima of the form $\tilde{\varepsilon}_i{}^*(\mathbf{x}, \tilde{\varepsilon}_j{}^-)$ and $\tilde{\varepsilon}_j{}^*(\mathbf{x}, \tilde{\varepsilon}_i{}^-)$, or of the form $\tilde{\varepsilon}_i{}^*(\mathbf{x})$ and $\tilde{\varepsilon}_j{}^*(\mathbf{x})$, where $\overline{\underline{\mathbf{x}}} \subset \mathbf{x}$.

Pareto-constraints that become inconsistent beyond the Pareto set are revealed as the bound objectives are optimized out. In simple problems, this degree of reduction might be reached through algebraic manipulations alone. For complex problems, however, full reduction might not be worthwhile due to the algebraic effort. Here, one can employ a more pragmatic approach by utilizing numerical results to identify additional active constraints that can be used to reduce the model further post optimality. If numerical solution reveals constraints that fulfill the Global Activity criterion from Definition 2, then such constraints can essentially be dealt with exactly as with constraints that are found to be active through MA. The globally active constraint is used to back-substitute variables post-optimality, thereby reducing the model further and giving a clearer picture of the relationships that exist at the Pareto set. As outlined in Section 3.1, this does require that the Lagrange multipliers are stored for each Pareto point to allow the evaluation of the activity of each constraint across the Pareto set.

5 Case - The Self-orienting Millimeter-scale Applicator

First published by Abramson et al. [42], the SOMA device (**S**elf-**O**rienting **M**illimeter-scale **A**pplicator) is a drug delivery device currently in development. The SOMA is designed for oral delivery of large proteins such as insulin, which cannot otherwise be administered orally, as the stomach breaks them down, and as they have poor permeability across the intestinal barrier. This substantially reduces the efficacy of such drugs, meaning they are mostly delivered via subcutaneous injections today.

Essentially, the SOMA is a pill-sized device designed to be swallowed by the user. Once in the stomach, the SOMA self-orients to a stable position due to a low center of mass and an outer shape inspired by that of leopard tortoises (*S. pardalis*) [42]. Once oriented, the device injects a biodegradable needle loaded with active pharmaceutical ingredient (API) into the *submucosa* tissue-layer of the stomach, which has a high density of blood vessels, allowing systemic uptake. This functionality is currently embodied with a linear spring actuator, held in place by a triggering mechanism (see Fig.3). The API mixture is shaped into a needle-like geometry (6) and is attached to a hub component (2) which is pre-loaded by a compression spring (4). The hub is held in place by two snap features, which are press-fit against the housing (1) by a plug (3), made out of isomalt, a dissoluble solid poly-alcohol. Once in the stomach, the device is submerged in stomach fluid, causing the plug to start dissolving to a point where the spring force pushes the snap features out of engagement. This triggers the device, with the spring pushing the needle into the stomach lining through a hole in the base (8) of the device. Until injection, the needle is kept dry in the hostile environment of the stomach by a silicone O-ring (5) and valve (7) that seal the needle inside the SOMA. The position of the centre of mass is low, as the base (8) is denser than the other parts, which aids self-orientation.

At the time of this study, the SOMA device was in the preliminary phases of design, still in the process of configuration, prototyping, and testing [42], and was yet to be tested on humans. Numerous configurations have been designed and built, with one, shown in Fig.3, showing the most

Journal of Mechanical Design. Received December 17, 2020;
Accepted manuscript posted August 31, 2021. doi:10.1115/1.4052444

	Part name	Suffix
1	Top Housing	*H*
2	Hub	*NH*
3	Plug	*P*
4	Power Spring	*PS*
5	Sealing ring	*S*
6	Needle	*N*
7	Valve	*V*
8	Base housing	*B*

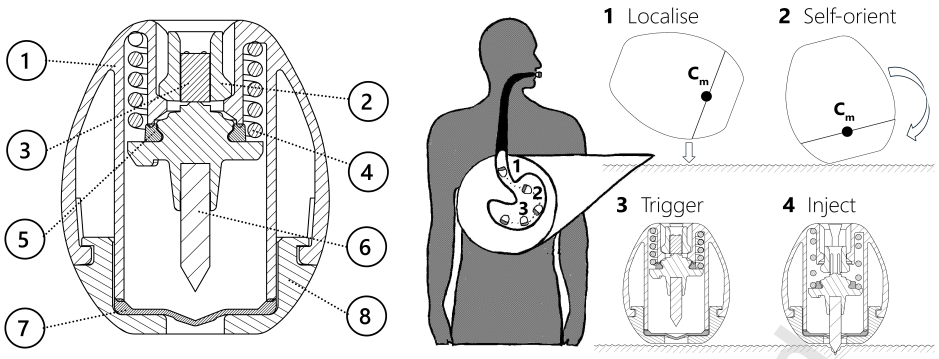


Fig. 3: An overview of the SOMA device (in part) adapted from [42]. The patient swallows the device, which self-orients inside the stomach and injects a needle of pure API into gastric tissue, detaching the needle from the device. Here the needle dissolves, resulting in systemic uptake while the device passes through the gastrointestinal tract and out of the body.

promise. Its outer shape was originally derived through optimization [42], but the inner configuration was iteratively developed, within the limits defined by the outer shape. This study focuses on the design of these internal components.

5.1 Optimization Model

The internal configuration design of the SOMA presents several design trade-off challenges. For an oral device to be viable, it needs to deliver an amount of API comparable to dosing with injection devices (e.g., insulin pens). This implies a dose of at least 80 units of insulin, which equates to a payload of approximately 2.8 mg of pure crystalline insulin. At the same time, the needle needs to be delivered reliably into a tissue layer deep enough to enable systemic uptake. The properties of the stomach lining are such that a large injection force is required to deliver the needle at the right depth. Hence, the challenge is to design a device that is small enough to be swallowable while reliably self orienting and injecting a sufficient amount of API deep enough. Furthermore, low cost and robust performance is essential. If only 1% of the world's 400M+ diabetics were treated with long-acting once-daily insulin from a SOMA, the annual production volume would be over 1.46bn devices. Given the potential volume, even slight improvements to the configuration may have a vast financial and societal impact. Understanding what causes trade-offs is hence highly valuable.

Four objectives were modelled for this study: swallowability, the height of the center of mass, API capacity, and injection depth. Given the early stage of development, the goal was to develop the simplest meaningful model, leading to the following simplifications:

1. As the focus is on the internal configuration, the outer shape is kept constant. Hence it is sufficient to optimize the vertical position of the center of mass to improve self-orientation.
2. Swallowability is proportional with the minor diameter [43], which is equivalent to d_{t1} illustrated in Fig. 4.
3. Injection depth is dependant on the mechanical properties of gastric tissue, the velocity at which the needle impacts gastric tissue, the diameter of the needle, and the sharpness of its tip. The needle is made from compacted protein, meaning that the sharper the tip, the more costly and sensitive the production process. Hence it is preferable to achieve a sufficient depth with a large velocity and not rely on sharpness. The impact velocity is therefore modelled as a maximizing objective since the injection depth increases monotonically with it.

The resulting initial optimization model is

$$\min \quad f_1(\mathbf{x}) = -\frac{\sum_{p=1}^{p=8} m_p \cdot (C_p + Z_p)}{(l_{t1} + l_{t2} + l_{b1}) \cdot \sum_{p=1}^{p=8} m_p} \tag{35}$$

$$\text{s.j.t} \quad c_1(\mathbf{x};\varepsilon_1) = d_{t1} - \varepsilon_1 \leq 0 \tag{36}$$

$$c_2(\mathbf{x};\varepsilon_2) = \varepsilon_2 - \rho\frac{\pi}{4} d_{n1}^2 \left(l_{n1} + \frac{1}{3} \cdot l_{n2} \right) \leq 0 \tag{37}$$

$$c_3(\mathbf{x};\varepsilon_3) = \varepsilon_3 - \sqrt{2\left(g + \frac{F_s}{m_{acc}} \right) z_{acc}} \leq 0 \tag{38}$$

$$\mathbf{g}(\mathbf{x}) \leq 0 \tag{39}$$

$$\mathbf{h}(\mathbf{x}) = 0 \tag{40}$$

$$\mathbf{x}, \varepsilon \in \mathbb{P} \tag{41}$$

where:

- f_1 is the self-orientation objective, which maximises the distance, Z_{cm}, between the top of the device and the system centre of mass, C_m, relative to the total height of the device, $l_{t1} + l_{t2} + l_{b1}$. Here, m_p, C_p, and Z_p are intermediate functions, with m_p describing the mass of each part in the device, C_p the centre of mass in each part, and Z_p the axial distance of each part from the top of the device. Expressions for m_p and C_p were derived explicitly using geometric idealisations (e.g. ellipsoids and cylinders) to describe the shape of the parts while accounting for fea-

Journal of Mechanical Design. Received December 17, 2020;
Accepted manuscript posted August 31, 2021. doi:10.1115/1.4052444

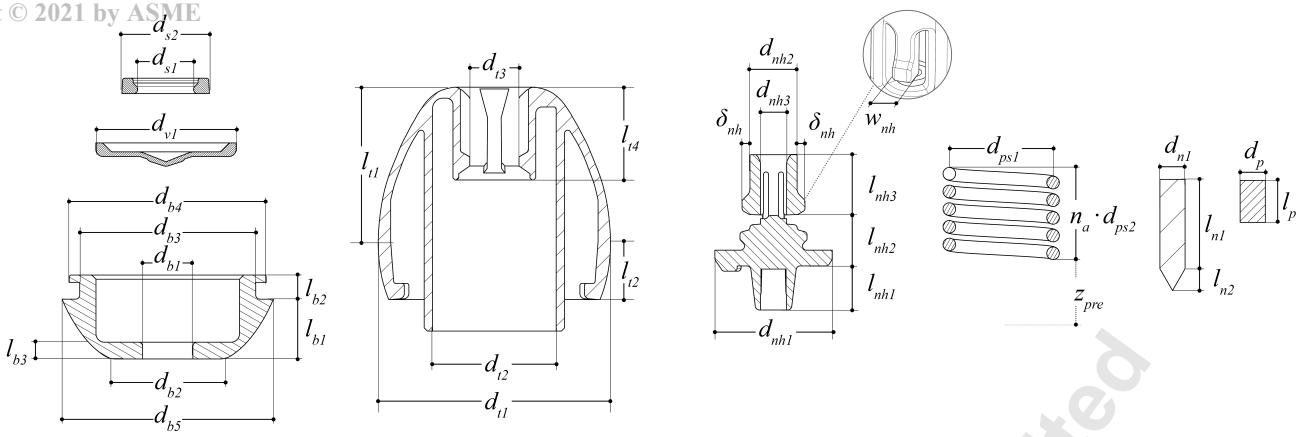


Fig. 4: Design variables of relevance to the constraints used to demonstrate the trade-off root cause analysis. l denotes length variables, d -diameters, δ -deflections, and z denotes vertical positions

tures such as rounds, draft, and snap interfaces. The objective function was then verified against CAD models of the SOMA in several sizes, with a max. deviation of 0.83% compared to the mass distribution analysis done in the CAD system (PTC Creo 4.0). This was deemed to be acceptable for the purposes of the study.

- c_1 is the bound objective for size minimization, where d_{t1} is the diameter of the device.
- c_2 is the bound API capacity objective, which maximises the mass of the needle, and is hence a simple volumetric expression. The design variables are illustrated in fig. 4, while ρ is the density of the compacted API.
- c_3 is the bound impact velocity objective, maximising the velocity of impact between needle and tissue. Here, F_s is the spring force, m_{acc} the mass that is accelerated during injection, z_{acc} the acceleration stroke between the initial position of the needle tip and the gastric tissue), and g the gravitational acceleration. Non-linearities such as device recoil (which counteracts z_{acc},), self-accelerating spring mass, and the reduction of F_s as a function of displacement, are also accounted for.

The inequality constraints, $\mathbf{g}(\mathbf{x})$, encompass criteria such as geometric fits (e.g., radial and axial part fits in use and assembly), manufacturability (e.g. min. wall thicknesses and feature size), and structural load cases (e.g., spring stress, interface stresses, needle buckling). The equality constraints, $\mathbf{h}(\mathbf{x})$, mostly account for the shape of the device. In total, there are 52 inequality constraints, 7 equality constraints, 45 design variables, and multiple parameters. For the present demonstration, further model details are not necessary and are omitted for brevity.

5.2 Monotonicity Analysis

Using MOMA, the model was reduced prior to computation down to 18 design variables, 28 inequality constraints, 2 equality constraints and 4 objectives, 3 of them bound. An important set of constraints - the radial fits between the parts - will be used to demonstrate the application of the methods described in Section 4. These constraints ensure that the parts fit together radially - i.e., that the plug fits into the hub, the hub into the top housing, the spring around the trigger system, and the top housing in the base:

$$h_1(l_{t1}^+, d_{t1}^-) = l_{t1} - d_{t1} C_T = 0 \tag{42}$$
$$h_2(d_{nh3}^+, d_p^-) = d_{nh3} - d_p - 2R_{cl} = 0 \tag{43}$$
$$h_3(d_{b4}^+, d_{b5}^-) = d_{b4} + 2R_{wt} + 2R_{cl} - d_{b5} = 0 \tag{44}$$
$$g_1(d_{t1}^-, l_{t1}^-, l_{t2}^+, d_{b5}^+) = d_{b5} - \sqrt{\frac{2(l_{t1}-l_{t2})d_{t1}^2}{l_{t1}} - \frac{(l_{t1}-l_{t2})^2 d_{t1}^2}{l_{t1}^2}} \leq 0 \tag{45}$$
$$g_2(d_{b3}^+, d_{b4}^-) = d_{b3} + 2R_{ov} - d_{b4} \leq 0 \tag{46}$$
$$g_3(d_{t2}^+, d_{b3}^-) = d_{t2} + 4R_{wt} + 2R_{cl} - d_{b3} \leq 0 \tag{47}$$
$$g_4(d_{t2}^-, d_{ps1}^+, d_{ps2}^+) = d_{ps1} + d_{ps2} + 2R_{cl} - d_{t2} \leq 0 \tag{48}$$
$$g_5(d_{t3}^+, \delta_{nh}^+, d_{ps2}^+, d_{ps1}^-) = d_{t3} + 2(\delta_{nh} + R_{cl} + R_{wt}) + d_{ps2} - d_{ps1} \leq 0 \tag{49}$$
$$g_6(d_{nh2}^+, d_{t3}^-) = d_{nh2} - d_{t3} + 2R_{cl} \leq 0 \tag{50}$$
$$g_7(d_{nh2}^-, d_{nh3}^+) = d_{nh3} + 2R_{wt} - d_{nh2} \leq 0 \tag{51}$$
$$g_8(\delta_{nh}^-) = 0.3\text{mm} - \delta_{nh} \leq 0 \tag{52}$$
$$g_9(d_p^-) = 1\text{mm} - d_p \leq 0 \tag{53}$$
$$g_{10}(\delta_{nh}^+, d_{nh3}^-) = 2\delta_{nh} + R_{cl} - d_{nh3} \leq 0 \tag{54}$$
$$g_{11}(d_{ps1}^-, d_{ps2}^+, n_a^-, z_{pre}^+, \delta_{nh}^-, w_{nh}^-) = \frac{z_{pre}\cos(\Theta_{nh}) G_{st} d_{ps2}^4}{16 d_{ps1}^3 n_a \delta_{nh} w_{nh}} - \sigma_c \leq 0 \tag{55}$$

where C_T is the aspect ratio between the diameter and height of the top housing, R_{wt} is the min. wall thickness, R_{ov} the min. radial interface overlap, R_{cl} the min. radial clearance, Θ_{nh} the contact angle in the trigger interface, G_{st} the shear modulus of the spring steel, and σ_c the allowable stress in the trigger interface. The variables are illustrated in fig. 4.

The monotonicity of the objectives w.r.t these variables is:

$$f(l_{t1}^-, l_{t2}^-, d_{t1}^-, d_{t2}^+, d_{t3}^+, d_{b3}^+, d_{b4}^+, d_{ps1}^+, d_{ps2}^+, d_{nh2}^+, d_{nh3}^-, \delta_{nh}^+, d_p^+) \tag{56}$$

$$c_1(d_{t1}^+) \tag{57}$$

$$c_3(l_{t1}^-, l_{t2}^-, d_{ps1}^+, d_{ps2}^-, d_{nh2}^+, \delta_{nh}^+) \tag{58}$$

c_2 is independent of these variables in the initial model. The only trade-off variables that are visible so far, are the device diameter, d_{t1}, and the spring wire diameter d_{ps2}, as exhibited by their opposite monotonicity in the objectives. To begin reducing the problem, we first use h_1 to eliminate l_{t1}, h_2 to eliminate d_{nh3}, and h_3 to eliminate d_{b5}. Furthermore, g_2, g_3, g_4, and g_6 are critical w.r.t. d_{b4}, d_{b3}, d_{t2} and d_{t3} respectively, meaning MP1 can be applied to eliminate them. After back-substitution, the objectives and constraints have changed:

$$\begin{aligned}
\text{min.} \quad & f(l_{t2}^-, d_{t1}^-, d_{t3}^+, d_{ps1}^+, d_{ps2}^+, d_{nh2}^+, \delta_{nh}^+, d_p^+) && (59)\\
\text{s.j.t.} \quad & c_1(d_{t1}^+; \varepsilon_1) && (60)\\
& c_3(d_{t1}^-, l_{t2}^-, d_{ps1}^+, d_{ps2}^-, d_{nh2}^+, \delta_{nh}^+; \varepsilon_3) && (61)\\
& g_1(d_{t1}^-, l_{t2}^+, d_{ps1}^+, d_{ps2}^+) = d_{ps1} + d_{ps2} \\
& + 6R_{wt]} + 6R_{cl} + 2R_{ov} \\
& - \sqrt{\frac{2(C_T d_{t1} - l_{t2}) d_{t1}}{C_T} - \frac{(C_T d_{t1} - l_{t2})^2}{C_T^2}} \leq 0 && (62)\\
& g_5(d_{nh2}^+, \delta_{nh}^+, d_{ps2}^+, d_{ps1}^-) = d_{nh2} + 2\delta_{nh} + d_{ps2} \\
& - d_{ps1} + 4R_{cl} + 2R_{wt} \leq 0 && (63)\\
& g_7(d_{nh2}^-, d_p^+) = d_p + 2R_{wt} + 2R_{cl} - d_{nh2} \leq 0 && (64)\\
& g_8(\delta_{nh}^-) = 0.3\text{mm} - \delta_{nh} \leq 0 && (65)\\
& g_9(d_p^-) = 1\text{mm} - d_p \leq 0 && (66)\\
& g_{10}(\delta_{nh}^+, d_p^-) = 2\delta_{nh} - R_{cl} - d_p \leq 0 && (67)\\
& g_{11}(d_{ps1}^-, d_{ps2}^+, n_a^-, z_{pre}^+, \delta_{nh}^-, w_{nh}^-) \\
& = \frac{z_{pre} \cos(\Theta_{nh}) G_{st} d_{ps2}^4}{16 d_{ps1}^3 n_a \delta_{nh} w_{nh}} - \sigma_c \leq 0 && (68)
\end{aligned}$$

So far, trade-offs have been revealed between size, c_1, and both impact velocity, c_3, and self-orientation, f_1, through d_{t1}, and between impact velocity and self-orientation through the spring wire diameter d_{ps2}. Increasing the wire diameter increases spring force and hence velocity, but it also increases the spring mass, shifting the system centre of mass upward.

Following Theorems 1 and 2, the boundedness of d_{t1} reveals important information about the SOMA device. Firstly, $c_1(\mathbf{x}; \varepsilon_1) \equiv 0$ for any $\varepsilon_L(1) < \varepsilon_1 < \varepsilon_U(1)$. Secondly, the only *non-objective lower bound* for d_{t1} is g_1, the constraint that ensures that the top and base housings fit together radially. This means that g_1 will be active at the single-objective minimum. Looking at eq. 62, it is evident that all the objectives cannot be minimised simultaneously, without reaching a point where $c_1(d_{t1}^+) < g_1(d_{t1}^-)$ meaning that $\mathcal{X}(d_{t1}) = \emptyset$. Hence, g_1 is at least semi-active in any bi-objective Pareto-front involving the size objective, c_1. As a consequence, l_{t2} is a trade-off variable when g_1 is active, and d_{ps2} also becomes a trade-off variable w.r.t. size. The implication for design is, that the further the mating surface between top and base is moved downward, the less space there is available for the spring mechanism. The only harmonious variable left in g_1, is d_{ps1}; identifying its' glb may reveal additional variables that contribute to the trade-offs between f_1, c_1 and c_3.

The remaining variables, including d_{ps1} have a conditionally critical set of constraints. Specifically, the spring, hub, and plug variables are potentially bound by inequality constraints relating to the yield stress of different parts, while the top housing variables are also involved in the axial fit constraints. Hence they cannot be eliminated without substantial algebraic manipulation to identify the glb or lub of each variable, meaning it is more efficient to identify the remaining active constraints numerically.

5.3 Numerical Results

The upper bound problem was solved 200,000 times using the SQP fmincon routine in MATLAB2019R [44] for different values of ε sampled from a quasi-random set (a leaped Halton set) distributed between $\varepsilon_\mathbf{L} = [8.5\text{mm}; 1.5\text{mg}; 10\text{m/s}]$ and $\varepsilon_\mathbf{U} = [11.5\text{mm}; 4.5\text{mg}; 30\text{m/s}]$. These values were set based on input from the SOMA team in Novo Nordisk. The results are shown in Fig. 5 and Table 1.

Objective	Optimum	Nadir	λ_{min}	λ_{max}
f_1	-0.78h	-0.64h	-	-
ε_1	8.67 mm	11.50 mm	0.0131	2.7708
ε_2	4.50 mg	1.50 mg	0.0016	0.4355
ε_3	28.34 m/s	10 m/s	0.0008	0.3795

Table 2: Numerical results

As the minimum λ values of each bound objectives are positive, they are active in the entire sampling region, and all feasible solutions are Pareto-optimal. As seen in Fig. 5, all four objectives are in trade-off with each other. Furthermore, g_5 and g_7 are globally active. Interestingly, g_1 and g_5 were violated in every infeasible iteration, pointing to inconsistent constraints beyond the Pareto set.

While 42% of the iterations yielded feasible, optimal solutions, the other 58% failed to identify a feasible solution. A few measures were taken to verify the model that led to these results. Firstly, the validity of the MA was assessed by running the original unreduced model over a narrower range of ε values. This led to the same results as with the reduced model. Secondly, a constraint satisfaction problem was run for all the failed iterations. This was used to search for a feasible solution to use as a new initial guess in a re-run with the

Journal of Mechanical Design. Received December 17, 2020;
Accepted manuscript posted August 31, 2021. doi:10.1115/1.4052444

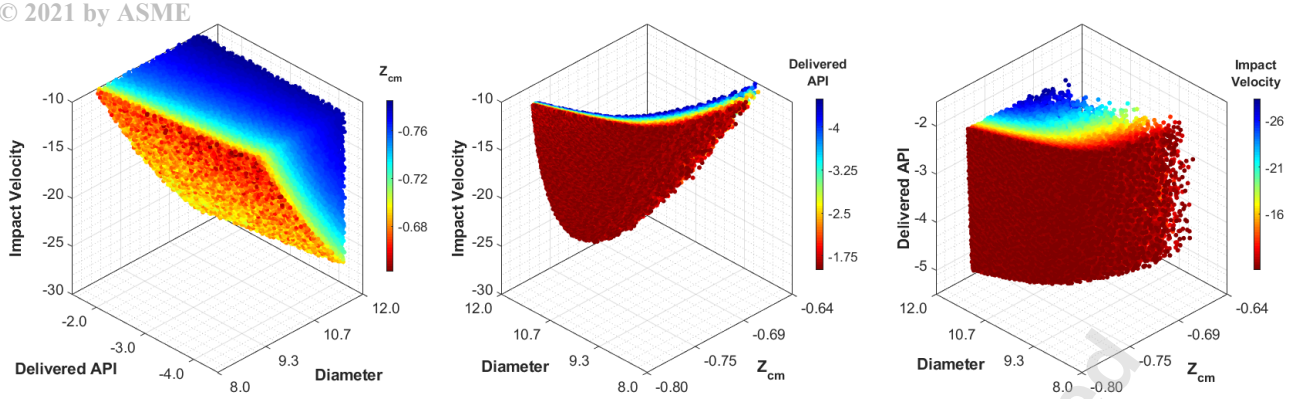


Fig. 5: Different projections of the 4D-Pareto set, where the 4th objective is visualised with a color map

same values of ε. Only 1.8% of these cases identified a new feasible initial guess, and these all yielded a Pareto point subsequently. This indicates that the remaining iterations indeed failed due to a lack of feasible domain caused by inconsistent constraints, meaning that the approximate Pareto-frontier of the sampled objectives had indeed been identified.

We note that the US-FDA generally recommends pills and capsules stay below a standard 00-size [43], which has a 8.35mm diameter, while the largest standard size, 000 capsules, are 9.91mm in diameter. Complications from swallowing pills start at about 8mm dia. and grows substantially beyond a 11mm dia. [45]. Initial work in the SOMA project has revealed that the impact velocity is critically important to the bioavailability of the delivered API (the % of the administered drug that reaches systemic circulation). It is also critical to the robustness and cost of the shaping of the needle geometry, as a low velocity results in a need for a sharper tip. Thus, the trade-off between size and velocity ultimately affects the amount of drug that can be delivered in a swallowable device and the cost of treatment. Model reduction using the numerical results reveals the cause of this trade-off.

5.4 εMA and Pareto Optimal Case Analysis

The global activity of g_5 and g_7 is used to eliminate $d^*_{ps1} = d_p + 2\delta_{nh} + d_{ps2} + 4R_{cl} + 2R_{wt}$ and $d^*_{nh2} = d_p + 2(R_{wt} + R_{cl})$ further reducing equations 59-68. Globally active axial fit and mechanical yield constraints allow the elimination of n_a and w_{nh}, the back substitution of which results in the elimination of z_{pre} from g_{11}, which will be handled implicitly from here on. Subsequently, the globally active bound size objective $c_1(d^+_{t1};\varepsilon_1)$ is used to eliminate d_{t1}, introducing $\tilde{\varepsilon_1}$ into g_1 and two objectives, f_1 and c_3:

$$\text{min.} \quad f(l^-_{t2}, d^+_{t3}, d^+_{ps2}, \delta^+_{nh}, d^+_p, \tilde{\varepsilon_1}^-) \tag{69}$$

$$\text{s.j.t.} \quad c_3(l^-_{t2}, d^-_{ps2}, d^+_p, \delta^+_{nh}, \tilde{\varepsilon_1}^-; \varepsilon_3) \tag{70}$$

$$g_1(\tilde{\varepsilon_1}^-, l^+_{t2}, d^+_{ps2}, d^+_p, \delta^+_{nh}) = 2d_{ps2} + 2\delta_{nh} + d_p + 10R_{wt} + 13R_{cl} + 2R_{ov} - \sqrt{\frac{2(C_T\tilde{\varepsilon_1} - l_{t2})\tilde{\varepsilon_1}}{C_T} - \frac{(C_T\tilde{\varepsilon_1} - l_{t2})^2}{C_T^2}} \leq 0 \tag{71}$$

$$g_8(\delta^-_{nh}) = 0.3\text{mm} - \delta_{nh} \leq 0 \tag{72}$$

$$g_9(d^-_p) = 1\text{mm} - d_p \leq 0 \tag{73}$$

$$g_{10}(\delta^+_{nh}, d^-_p) = 2\delta_{nh} - R_{cl} - d_p \leq 0 \tag{74}$$

$$g_{11}(d^+_{ps2}, d^-_p, \delta^-_{nh}) \leq 0 \tag{75}$$

As expected, g_1 makes l_{t2} a trade-off variable, as $\overline{l_{t2}} \to 0$ as $\tilde{\varepsilon_1} \to 0$ when $g_1 \equiv 0$,and given that $f(l^-_{t2})$ and $c_3(l^-_{t2})$. The velocity objective c_3 has not been optimized out, as there is no closed form solution to $c_3(\mathbf{x}, \tilde{\varepsilon_1}; \varepsilon_3) \equiv 0$ w.r.t any $\underline{x}$. Its elimination would involve solving for d_{ps2}, as it is critically constrained from below by c_3 and is shared with the largest number of constraint functions that remain in the model. This would make g_1 a multiobjective Pareto constraint. Therefore, g_1 is involved in three Pareto-optimal activity cases; when g_1 bounds $\tilde{\varepsilon_1}$, d_{ps2}, and l_{t2}. Looking at these cases in detail using the procedure from Table 1, reveals the root cause of the shape and position of the bi-objective Pareto front between size and velocity.

Activity case 1: Smallest Possible Device, U($\tilde{\varepsilon_1}^+$)
Here g_1 determines $\underline{\tilde{\varepsilon_1}}^*$ and yields the optimal size. Eliminating l_{t2} allows a closed form solution for $\tilde{\varepsilon_1}$ using Eq. 71. Letting $l_{t2} \to 0$, implying that the mating surface between top and base is located at the widest point of the device, allows the smallest $\tilde{\varepsilon_1}$. Inserting this, and the parameter values, $C_t = 0.68$, $R_{wt} = 0.45$mm, $R_{cl} = 0.1$mm, $R_{ov} = 0.6$mm yields a reduced expression:

$$g_1(\tilde{\varepsilon_1}^-, d^+_{ps2}, d^+_p, \delta^+_{nh}) = 2(d_{ps2} + \delta_{nh}) + d_p + 7\text{mm} - \tilde{\varepsilon_1} \tag{76}$$

$$\Rightarrow \underline{\tilde{\varepsilon_1}}(d^+_{ps2}, \delta^+_{nh}, d^+_p) = 2d_{ps2} + 2\delta_{nh} + d_p + 7\text{mm} \tag{77}$$

$$\Rightarrow g_8(\delta_{nh^-}) \equiv 0 \wedge g_9(d_p^-) \equiv 0 \quad (78)$$

$$\Leftrightarrow \underline{\delta_{nh}} = 0.3\text{mm}, \underline{d_p} = 1\text{mm} \quad (79)$$

$$\Leftrightarrow \underline{\tilde{\varepsilon_1}}^* = 2d_{ps2} + 8.6\text{mm} \quad (80)$$

As d_{ps2} is a trade-off variable, minimising $\tilde{\varepsilon_1}$ will lead to a point where $g_{11} < g_8$ and $g_{11} < g_9$ w.r.t. d_p and δ_{nh}. This results in g_8 and g_9 becoming active, leading to the back-substitution performed in eqs. 77-80. As a result, $\mathcal{X}(d_{p2})$ is narrowed at the Pareto frontier between size and velocity, given that these reductions leave $g_{11}(d_{ps2}^+)$ and $c_3(d_{ps2}^-, \tilde{\varepsilon_1}^-; \varepsilon_3)$. Further, c_3 is critical w.r.t. d_{ps2}, meaning that $\varepsilon_L(3)$ ultimately determines the lowest feasible value of d_{ps2}, and hence $\underline{\tilde{\varepsilon_1}}^*$.

Activity case 2: Maximum Impact Velocity, $\mathrm{U}(\tilde{\varepsilon_3}^-)$
Here g_1 determines $\overline{d_{ps2}}$. As c_3 is monotonically decreasing w.r.t. d_{ps2} to the power of 4, its supremum yields the single-objective optimal impact velocity. Thus, the same parameter values and value of l_{t2} can be inserted as in Case 1. Yet, as opposed to Case 1, g_8 and g_9 are inactive, as pushing d_{ps2} to its upper limit makes $g_{11}(d_{ps2}^+, d_p^-, \delta_{nh}^-) \leq 0$ active. g_{11} is a interface stress criterion, and because the spring force grows with the wire diameter, the dimensions that determine the area - d_p and δ_{nh} increase correspondingly. This in turn makes $g_{10}(\delta_{nh}^+, d_p^-)$ active, as $g_9 < g_{10}$ for any $\delta_{nh} > 0.45$mm. These activities yield:

$$g_1(\tilde{\varepsilon_1}^-, d_{ps2}^+, d_p^+, \delta_{nh}^+) = 2(d_{ps2} + \delta_{nh}) + d_p + 7\text{mm} - \tilde{\varepsilon_1} \quad (81)$$

$$\Rightarrow \overline{d_{ps2}}(\tilde{\varepsilon}_1^+, \delta_{nh}^-, d_p^-) = 0.5(\tilde{\varepsilon_1} - d_p - 2\delta_{nh} - 7\text{mm}) \quad (82)$$

$$g_8 \equiv 0 \Rightarrow d_p^* = 2\delta_{nh} - R_{cl} \quad (83)$$

$$g_{11} \equiv 0 \Rightarrow \delta_{nh} = \delta_{nh}(d_{ps2}^+; \sigma_c, \sigma_y) \quad (84)$$

$$\Rightarrow \overline{d_{ps2}} = 0.5(\tilde{\varepsilon_1} - 4\delta_{nh}(d_{ps2}^+; \sigma_c, \sigma_y) - 6.9mm) \quad (85)$$

where g_{11} has been used to implicitly eliminate the overlap δ_{nh} between hub and top housing in the load bearing trigger interface as no closed form solution exists; σ_y is the yield stress of the spring, and σ_c is the allowable static stress in the trigger interface. This substitution reveals a feedback coupling; as the wire diameter d_{ps2} is increased, so does the required load bearing area, reducing the space available for the spring wire in a device of a given size, $\tilde{\varepsilon_1}$.

Activity case 3: Lowest Possible Center of Mass, $f_1(l_{t2}^-)$
Here g_1 determines $\overline{l_{t2}}$. As $f_1(l_{t2}^-)$, this case occurs at the single objective optimal self-orientation. Given the non-linearity of Eq. 71 w.r.t. l_{t2}, the variable is best eliminated implicitly, yielding $\overline{l_{t2}} = l_{t2}(\tilde{\varepsilon_1}^+, d_p^-, d_{ps2}^-, \delta_{nh}^-)$. As a consequence d_{ps2} is bounded by c_3, d_p by either g_9 or g_{10}, and δ_{nh} by either g_8 or g_{11}. Furthermore, $\tilde{\varepsilon_1} = \varepsilon_U(1)$, as no constraint bounds $\tilde{\varepsilon_1}$ from above. As discussed in Section 5.2, g_1 reduces objective alignment between self-orientation and size and impact velocity respectively.

5.5 Design Implications

These Pareto-optimal activity cases demonstrate the root cause of the position and shape of the bi-objective Pareto front between size and impact velocity. The smaller the coiling diameter d_{ps1} of the spring, the more spring force (and hence impact velocity) and the smaller a device. Given that the spring needs to fit inside the diameter of the guiding cylinder d_{t2} and around the trigger system, g_4 and g_5 are active, meaning a harmonious variable is minimised out, $d_{ps1}^* = d_p + d_{ps2} + 2\delta_{nh} + 2R_{wt} + 4R_{cl}$ introducing a trade-off variable, d_{ps2}, and δ_{nh} into g_1.

We can see from g_1 that as the coiling diameter is reduced and the wire diameter increased, the available space left for the trigger system is reduced. The trigger system distributes the spring force over an area equal to $A = 2\delta_{nh} w_{nh} = 2\delta_{nh}(d_p - R_{cl} - R_{wt})$, and stiffening the spring increases the spring force but also reduces the load-bearing area, see Equation 85. With d_{ps2} being the variable with the largest influence on impact velocity (to the power of 4), and d_{ps1} being the second most (to the power of 3), activities of g_5, g_7 and g_{10} are ultimately the main driver of the trade-off. Had the spring and the trigger geometry existed in different cross-sections, the alignment between the two objectives might be drastically improved. This would correspond to d_{ps1}^* being independent of δ_{nh}, not only shifting the glb of d_{ps1} downward, but also removing the contribution of δ_{nh} to the constraint that determines $\underline{\tilde{\varepsilon_1}}$, improving size and impact velocity simultaneously. This also increases $\overline{\delta_{nh}}$. Furthermore, the design of the snap-interface between the top- and base housings also influences objective alignment between self-orientation, size, and impact velocity.

6 Discussion

Pareto set dependency analysis presents an optimization-focused alternative to current techniques for dependency analysis (e.g. DSM and Axiomatic Design [7]), with MOMA allowing the analysis of dependencies unique to the optimum by addressing the impact of constraints directly, and εMA revealing regional dependencies that shape the Pareto set. As an added benefit, the procedure for systematic reduction of multiobjective problems helps reduce computational cost due to the elimination of constraints and variables. In computationally expensive problems, this pre- and post optimality analysis procedure may also help reveal insights that would be too costly to reach computationally, e.g., describing certain relationships that would otherwise only come to light if the Pareto set is exhaustively identified.

From a design optimization perspective, Pareto set dependency analysis is a rigorous approach to exploring the limitations of a given configuration. The definitions and theorems presented allow the systematic identification of trade-off variables, active bound objectives and Pareto-constraints, and the constraints that introduce new trade-off variables. In doing so, one determines what objectives are in trade-off and, even more importantly, the underlying root causes of these trade-offs, clearly exposing the weaknesses in the configura-

tion design. An example from the SOMA case is the trade-off between size and impact velocity, which is in part caused by the spring needing to fit around the trigger.

If ε-monotonicity analysis is applied exhaustively to all Pareto-optimal activity cases and all active Pareto-constraints are be identified, the Pareto set is essentially derived explicitly. This is similar to the approaches developed by Gobbi et al. [30] and Mastinu et al. [31] for the explicit derivation of the Pareto set of lower-dimensional problems using back-substitution of ε. However, ε-monotonicity analysis goes beyond this to allow the identification of the drivers of trade-offs. Furthermore, the analysis is opportunistic; it can be performed partially and still provide useful insights. It is neither necessary that every case is studied nor that all bound objectives are eliminated in order for the analyst to identify some of the dependencies that reduce objective alignment and may guide redesign. From this, it follows that one might view the presented analysis as a way of *checking the design* ahead of computation or *explaining the results* after computation. For example, one can use the theorems to reduce a multiobjective model after computation using numerical activity information, should the model at hand be too large or complex to reduce through algebraic analysis alone. Alternatively, one could skip computation should initial MOMA reveal drivers of trade-off that might be eliminated through a change in configuration design.

The opportunistic nature of monotonicity analysis (MA) also reveals the key limitations of the methodology we have developed. Firstly, not all problems are monotonic or even differentiable. This might be dealt with using techniques for local [46] and regional MA [36] if the expressions are regionally differentiable. This comes at the cost of increased analysis effort, which might be offset using sampling-based computational experiments (e.g., DoE) to reveal regional properties in non-monotonic or non-algebraic problems.

Secondly, MA mostly relies on algebraic manipulations, and some design problems are too complex to be expressed algebraically. Yet, that certain aspects of a design's behaviour can only be expressed numerically does not necessarily imply a lack of monotonicity, e.g. as is often the case with stiffness and deflection. In such situations, implicit MA [36] procedures, and meta-models might be used. This would reveal the variables and constraints that cause trade-offs, albeit without the derivation of explicit expressions of the relationships that exist in the Pareto set.

It is also well accepted that purely algebraic models can play a substantial role in practice [47], in both conceptual and configuration design. These phases are often characterized by a lack of sophisticated quantitative models to support decision making due to requirement uncertainty and the modelling effort involved, compared to how quickly and often the design changes [48]. Configuration design also often involves the combination and arrangement of well-known types of parts and modules, which might be described algebraically, for example as seen in the machine elements, engines, hydraulics, and thermal systems.

Finally, the effort in analysis is proportional to the number of objectives, constraints, and variables in the problem. This effort is amplified by non-monotonicity and by regionally active constraints. Thus the bookkeeping and algebraic effort required to reduce a multiobjective model systematically may be prohibitive if the problem is large.. Here, the use of symbolic solvers can help reduce the effort in back-substitution and model reduction. In that regard, quite some work was done (with some success) on automating MA in the 1980s [46, 49]. In the view of the authors, there is potential in attempting to improve automation of MA (and thus MOMA and εMA) by leveraging the achievements made in computational techniques such as machine learning, AI, and data analysis and clustering, since MA methods where last in vogue. Given the advances in meta-modelling since then, it is also not unlikely that more complicated non-algebraic models might be analysed using the methods described in this paper. Hence the (partial) automation of MOMA and εMA is possible future work.

Ultimately, the value of this methodology comes down to the cost involved in analysis vs. the expected benefit in discovering better configurations. As discussed in the introduction, trade-off knowledge and decision-making are largely experience-driven in early stage design. Finger and Dixon [50] highlighted the dearth of quantitative design analysis and evaluation methods for the early stages, especially those which allow multiobjective analysis and support the identification of alternative configurations and concepts. The presented methodology addresses some of these unmet needs.

When the cost vs. benefit estimate noted above is favorable, the methodology might be used to target iterative configuration redesign efforts, to guide morphological studies to identify alternative solutions, or simply to explain the results of an optimization model from a design perspective. For small, tightly coupled systems such as the SOMA device, the value in discovering the non-obvious influence of certain variables and constraints on design trade-offs, amply justifies the analysis effort. For a larger system, the methodology can be worthwhile if the system is obviously monotonic or if the optimization model is constructed at an architectural level of abstraction that limits the number of design variables and expressions to analyze.

7 Conclusions

Trade-offs between objectives are an inevitable challenge in mechanical design. In multiobjective optimization, most prior work focused on quantifying these trade-offs, but there has been little prior work on their causes. Understanding this causality provides insights for improvements in proportional and, most importantly, in configuration design.

We demonstrated extensions to monotonicity analysis specific to multiobjective problems that allow rigorous identification of the constraints and variables contributing to trade-offs. Using the upper bound formulation for multiobjective problems, we extended monotonicity analysis and its application, proposing a novel procedure, ε-monotonicity analysis, to identify and study the constraints bounding the Pareto set. The methodology leads to deeper insights into the strengths and weaknesses of a design configura-

tion. We demonstrated the methodology on the early-stage design of the SOMA device, finding trade-offs that are in part caused by a load-bearing interface needing to fit inside the spring that exerts said load. Such insights may guide redesign resulting in improvements in performance beyond what is achievable through proportional design, i.e., beyond the Pareto set for the particular embodiment. A systematic redesign procedure to identify such improvements utilizing the output of the presented analysis method will be treated in a subsequent publication.

Acknowledgements

The authors would like to thank the Danish Innovations Fund and the Novo Nordisk STAR-programme for funding this research project (grant nr. 7038-00221B), Novo Nordisk for sharing design information and data, Asst. Prof. Giovanni Traverso of MIT and his colleagues for their helpful comments and input, and the University of Michigan Donald C. Graham Endowed Chair for providing visiting scholar support. The opinions presented here are solely those of the authors.

References

[1] W Brian Arthur. "Why Do Things Become More Complex?" In: *Scientific American* 268.5 (1993), pp. 144–144. DOI: 10.1038/scientificamerican0593-144.

[2] Durward K Sobek, Allen C Ward, and Jeffrey K Liker. "Toyota ' s Principles of Set-Based Concurrent Engineering Toyota ' s Principles of Set-Based Concurrent Engineering". In: *Sloan Management Review* 40.2 (1999), pp. 67–83.

[3] Saeema Ahmed, Ken M. Wallace, and Luciënne T.M. Blessing. "Understanding the differences between how novice and experienced designers approach design tasks". In: *Research in Engineering Design* 14.1 (2003), pp. 1–11. DOI: 10.1007/s00163-002-0023-z.

[4] G Pahl and W Beitz. *Engineering design — A systematic approach*. 1999. DOI: 10.1016/0261-3069(96)84970-3.

[5] Panos Y Papalambros and Kristina Shea. "Creating Structural Configurations". In: *Formal Engineering Design Synthesis*. Cambridge University Press, Nov. 2001, pp. 93–125. DOI: 10.1017/CBO9780511529627.007.

[6] David G. Ullman, Thomas G. Dietterich, and Larry A. Stauffer. "A model of the mechanical design process based on empirical data". In: *Artificial Intelligence for Engineering, Design, Analysis and Manufacturing* 2.1 (1988), pp. 33–52. DOI: 10.1017/S0890060400000536.

[7] Nam P. Suh. "Axiomatic Design Theory for Systems". In: *Research in Engineering Design - Theory, Applications, and Concurrent Engineering* 10.4 (1998), pp. 189–209. DOI: 10.1007/s001639870001.

[8] Hillary G. Sillitto. "On systems architects and systems architecting: Some thoughts on explaining and improving the art and science of systems architecting". In: *19th Annual International Symposium of the International Council on Systems Engineering, INCOSE 2009*. 2009. DOI: 10.1002/j.2334-5837.2009.tb00995.x.

[9] M M Andreasen and T J Howard. "Is Engineering Design Disappearing from Design Research ?" In: *The Future of Design Methodology*. Ed. by H Birkhofer. London: Springer Verlag, 2011. Chap. 2, pp. 21–34. DOI: 10.1007/978-0-85729-615-3.

[10] Ching-Shin Norman Shiau and Jeremy J Michalek. "Should Designers Worry About Market Systems?" In: *Journal of Mechanical Design* 131.1 (Dec. 2008). DOI: 10.1115/1.3013848.

[11] Robin C Purshouse and Peter J Fleming. "Conflict, Harmony, and Independence: Relationships in Evolutionary Multi-criterion Optimisation". In: *Evolutionary Multi-Criterion Optimization*. Ed. by Carlos M Fonseca et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 16–30.

[12] Indraneel Das. "A Preference Ordering Among Various Pareto Optimal Alternatives". In: *Structural Optimization* 18 (1999), pp. 30–35. DOI: 10.1007/BF01210689.

[13] Jarod C Kelly et al. "Incorporating user shape preference in engineering design optimisation". In: *Journal of Engineering Design* 22.9 (2011), pp. 627–650. DOI: 10.1080/09544821003662601.

[14] R. T. Marler and J. S. Arora. "Survey of multi-objective optimization methods for engineering". In: *Structural and Multidisciplinary Optimization* 26.6 (2004), pp. 369–395. DOI: 10.1007/s00158-003-0368-6.

[15] E. M. Kasprzak and K. E. Lewis. "Pareto analysis in multiobjective optimization using the collinearity theorem and scaling method". In: *Structural and Multidisciplinary Optimization* 22.3 (Oct. 2001), pp. 208–218. DOI: 10.1007/s001580100138.

[16] Kevin N. Otto and Erik K. Antonsson. "Trade-off strategies in engineering design". In: *Research in Engineering Design* 3.2 (1991), pp. 87–103. DOI: 10.1007/BF01581342.

[17] S Gunawan and S Azarm. "Multi-objective robust optimization using a sensitivity region concept". In: *Structural and Multidisciplinary Optimization* 29.1 (2005), pp. 50–60. DOI: 10.1007/s00158-004-0450-8.

[18] Christopher A Mattson and Achille Messac. *Pareto Frontier Based Concept Selection Under Uncertainty, with Visualization*. Tech. rep. 2005, pp. 85–115.

[19] Carlos M. Fonseca and Peter J. Fleming. "Multiobjective optimization and multiple constraint handling with evolutionary algorithms - Part I: A unified formulation". In: *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans.* (1998). DOI: 10.1109/3468.650319.

[20] Mehmet Unal, Gordon P. Warn, and Timothy W. Simpson. "Quantifying tradeoffs to reduce the dimensionality of complex design optimization problems and expedite trade space exploration". In: *Structural and Multidisciplinary Optimization* 54.2 (2016), pp. 233–248. DOI: 10.1007/s00158-015-1389-7.

[21] Martin Philip Bendsøe and Noboru Kikuchi. "Generating optimal topologies in structural design using a homogenization method". In: *Computer Methods in Applied Mechanics and Engineering* (1988). DOI: 10.1016/0045-7825(88)90086-2.

[22] Mehmet Unal, Gordon P. Warn, and Timothy W. Simpson. "Quantifying the shape of pareto fronts during multi-objective trade space exploration". In: *Journal of Mechanical Design, Transactions of the ASME* 140.2 (2018), pp. 1–13. DOI: 10.1115/1.4038005.

[23] Indraneel Das. "On characterizing the "knee" of the Pareto curve based on Normal-Boundary Intersection". In: *Structural Optimization* 18 (1999), pp. 107–115. DOI: 10.1007/BF01195985.

[24] Bart Frischknecht and Panos Papalambros. "A Pareto Approach To Aligning Public and Private Objectives in Vehicle Design". In: 2008.

[25] Bart D. Frischknecht, Diane L. Peters, and Panos Y. Papalambros. "Pareto set analysis: Local measures of objective coupling in multiobjective design optimization". In: *Structural and Multidisciplinary Optimization* 43.5 (May 2011), pp. 617–630. DOI: 10.1007/s00158-010-0599-2.

[26] Jin Wu and Shapour Azarm. "Metrics for quality assessment of a multiobjective design optimization solution set". In: *Journal of Mechanical Design, Transactions of the ASME* (2001). DOI: 10.1115/1.1329875.

[27] Timothy Ward Athan and Panos Y. Papalambros. "A quasi-Monte Carlo method for multicriteria design optimization". In: *Engineering Optimization* 27.3 (1996), pp. 177–198. DOI: 10.1080/03052159608941405.

[28] P. Papalambros and D. J. Wilde. "Global Non-iterative Design Optimization Using Monotonicity Analysis". In: *Journal of Mechanical Design, Transactions of the ASME* 78 -WA/DE-17 (1978).

[29] Nestor F Michelena and Alice M Agogino. "Multiobjective Hydraulic Cylinder Design". In: *Journal of Mechanisms, Transmissions, and Automation in Design* 110.1 (Mar. 1988), pp. 81–87. DOI: 10.1115/1.3258910.

[30] M. Gobbi et al. "On the analytical derivation of the Pareto-optimal set with applications to structural design". In: *Structural and Multidisciplinary Optimization* 51.3 (2015), pp. 645–657. DOI: 10.1007/s00158-014-1152-5.

[31] Giampiero Mastinu, Massimiliano Gobbi, and Carlo Miano. *Optimal design of complex mechanical systems: With applications to vehicle engineering*. 2006, pp. 1–359. DOI: 10.1007/978-3-540-34355-4.

[32] Pramod Jain and Alice M. Agogino. "Theory of design: An optimization perspective". In: *Mechanism and Machine Theory* 25.3 (1990), pp. 287–303. DOI: 10.1016/0094-114X(90)90030-N.

[33] K. Ishii and B. Parkan. "Active Constraint Deduction - A Framework for Expert Systems in Mechanical Systems Design". In: *Advances in Design Automation - ASME Design Technology Conferences - The Design Automation Conference*. Vol. 10. 1987.

[34] Jonathan Cagan and Alice M. Agogino. *Innovative design of mechanical structures from first principles*. 1987. DOI: 10.1017/S0890060400000275.

[35] Kalyanmoy Deb and Aravind Srinivasan. "Innovization: Innovating design principles through optimization". In: *GECCO 2006 - Genetic and Evolutionary Computation Conference* 2 (2006), pp. 1629–1636.

[36] Panos Y. Papalambros and Douglass J. Wilde. *Principles of Optimal Design*. Cambridge University Press, Jan. 2017. DOI: 10.1017/9781316451038.

[37] D.G. Carmichael. "Computation of Pareto Optima in Structural Design". In: *International Journal for Numerical Methods in Engineering* 15 (1980), pp. 925–952. DOI: 10.1017/S0022029900029393.

[38] J G Lin. "Maximal Vectors and Multi-Objective Optimization". In: *Journal of Optimization Theory and Applications* 18.01 (1976).

[39] George Mavrotas. "Effective implementation of the ε-constraint method in Multi-Objective Mathematical Programming problems". In: *Applied Mathematics and Computation* 213.2 (2009), pp. 455–465. DOI: https://doi.org/10.1016/j.amc.2009.03.037.

[40] Yacov Y. Haimes and Warren A. Hall. "Multiobjectives in water resource systems analysis: The Surrogate Worth Trade Off Method". In: *Water Resources Research* 10.4 (1974), pp. 615–624. DOI: 10.1029/WR010i004p00615.

[41] Panos Y. Papalambros. "Model Reduction and Verification Techniques". In: *Advances in Design Optimization*. Ed. by H Adeli. New York: Chapman and Hall, 1994, pp. 109–138.

[42] Alex Abramson et al. "An ingestible self-orienting system for oral delivery of macromolecules". In: *Science* 363.6427 (2019). DOI: 10.1126/science.aau2277.

[43] U.S. Department of Health and Human Services Food and Drug Administration (CDER). "Guidance for Industry: Size, Shape and Other Physical Attributes of Generic Tablets and Capsules". In: *Pharmaceutical Quality/CMC* December (2013), pp. 1–11.

[44] Mathworks. *Optimization Toolbox ™ Users Guide R2020b, retrieved November 27, 2020*. 2020.

[45] K. S. Channer and J. P. Virjee. "The effect of size and shape of tablets on their esophageal transit". In: *Journal of Clinical Pharmacology* 26.2 (1986), pp. 141–

146. DOI: 10.1002/j.1552-4604.1986.tb02922.x.

[46] S Azarm and P Papalambros. "An Automated Procedure for Local Monotonicity Analysis". In: *Journal of Mechanisms, Transmissions, and Automation in Design* 106.1 (Mar. 1984), pp. 82–89. DOI: 10.1115/1.3258566.

[47] G A Hazelrigg. "On the Role and Use of Mathematical Models in Engineering Design". In: *Journal of Mechanical Design* 121.3 (Sept. 1999), pp. 336–341. DOI: 10.1115/1.2829465.

[48] Rajesh Radhakrishnan and Daniel A McAdams. "A Methodology for Model Selection in Engineering Design". In: *Journal of Mechanical Design* 127.3 (2005), pp. 378–387. DOI: 10.1115/1.1830048.

[49] J Zhou and R W Mayne. "Interactive Computing in the Application of Monotonicity Analysis to Design Optimization". In: *Journal of Mechanisms, Transmissions, and Automation in Design* 105.2 (June 1983), pp. 181–186. DOI: 10.1115/1.3258506.

[50] Susan Finger and John R Dixon. "A review of research in mechanical engineering design. Part II: Representations, analysis, and design for the life cycle". In: *Research in Engineering Design* 1.2 (1989), pp. 121–137. DOI: 10.1007/BF01580205.

Appendix 3: Paper B

Title: A Novel Approach to Configuration Redesign: Using Multiobjective Monotonicity Analysis to Alter the Pareto-set (2021)
Authors: Sigurdarson, N.S.; Eifler, T.; Ebro, M.; Papalambros, P.Y.
Publication: Submitted and approved pending revisions in the ASME Journal of Mechanical Design. The appended manuscript is the revision currently under review.

A Novel Approach to Configuration Redesign: Using Multiobjective Monotonicity Analysis to Alter the Pareto-set

Nökkvi S. Sigurdarson*
Mechanical Engineering,
Technical University of Denmark,
Kgs. Lyngby, Denmark,
noksig@mek.dtu.dk

Tobias Eifler
Mechanical Engineering,
Technical University of Denmark,
Kgs. Lyngby, Denmark,
tobeif@mek.dtu.dk

Martin Ebro
Device R&D,
Novo Nordisk A/S,
Hillerød, Denmark,
mixe@novonordisk.com

Panos Y. Papalambros
Mechanical Engineering,
University of Michigan,
Ann Arbor, MI 48109,
pyp@umich.edu

Configuration (or topology or embodiment) design remains a ubiquitous challenge in product design optimization and in design automation, meaning configuration design is largely driven by experience in industrial practice. In this article, we introduce a novel configuration redesign process founded on the interaction of the designer with results from rigorous multiobjective monotonicity analysis. Guided by Pareto-set dependencies, the designer seeks to reduce trade-offs among objectives or improve optimality overall, deriving redesigns that eliminate dependencies or relax active constraints. The method is demonstrated on an ingestible medical device for oral drug delivery, currently in early concept development.

Nomenclature

$\mathcal{A}$ Attainable set
$\mathcal{C}$ Pareto Set
$\mathbf{c}$ Vector of bound objectives in the upper bound problem
D_s Indices of the constraint functions that depend on a shared variable x_i
f Primary objective function in the upper bound problem
$f(x^+)$ A function increasing monotonically w.r.t. x
$f(x^-)$ A function decreasing monotonically w.r.t. x
F^0 The utopia point
$\mathbf{g}(\mathbf{x})$ Vector of inequality constraints of the design problem
$\mathbf{h}(\mathbf{x})$ Vector of equality constraints of the design problem
j Number of computational iterations ε is sampled over
k Number of objectives
p Number of redesign iterations
$\mathcal{X}$ The set constraint
$\mathbf{x}$ vector of design variables
$\underline{x}$ Argument of the infimum of the design problem
$\overline{x}$ Argument of the supremum of the design problem
$\overline{\underline{x}}$ Trade-off variable
$\overline{\overline{x}}$ A monotonically decreasing harmonious variable
$\underline{\underline{x}}$ A monotonically increasing harmonious variable
$\overline{\varepsilon}$ A k-1 dimensional vector of upper-bound parameters
ε_i Upper-bound parameter for the *ith* bound objective
$\varepsilon_{\mathbf{L}}$ Lower limit of objective bounds
$\varepsilon_{\mathbf{U}}$ Upper limit of objective bounds
$\tilde{\varepsilon}_i$ Reduced-objective variable

1 Introduction

In the 'double-diamond' model of the design process [1], the first diamond results in the generation of a design concept in a functional form as starting point for the generation of a particular embodiment design in the second diamond. Product design optimization typically iterates on the particular embodiment instantiation to achieve optimality with respect to given objectives while satisfying a set of given constraints. In practice, starting with a particular design and

*Corresponding Author

finding an optimum iteratively is in fact a design improvement or redesign process. Such improvements are associated with (i) proportional changes, where design variables are resized; (ii) parametric changes, where parameters, e.g. material properties or production processes are modified allowing for relaxation of design constraints; and (iii) configuration changes where the embodiment of functions or distribution of sub-functions is modified.

Proportional (or size) design has been facilitated by advancements in computational modeling, computing power, and optimization techniques. More robust and higher fidelity numerical solutions in size optimization have reduced risk in constraint relaxation for increasingly larger and more complex problems. Parametric design with gradual constraint relaxation has been a primary approach for product performance improvement in industry. Examples of such include new or improved materials, new production processes, and a deeper understanding of failure phenomena leading to reduced design margins, as seen in the design of engines [2].

Configuration design, also referred to as embodiment design [3] and layout design [4], is the process of creating actual 'buildable' instantiations of a design concept, transitioning from an initial abstract functionality to a more geometric realisation. Conceptual functionality may be implemented in various ways, and so many alternative configurations may be created, one of which is selected as "the best" for further development. Critical decisions that determine how a system fits together, how its parts interact with each other, and how the system interacts with its surroundings are made in the embodiment/configuration design stage. Examples include the layout and distribution of parts and sub-functions, force paths, mechanism design, and assembly sequence [3, 5].

In design optimization, configuration design remains elusive. The fundamental challenge is the lack of appropriate mathematical modeling capabilities: different configurations require different mathematical problem formulations. A notable exception is topology optimization [6], the success of which lies in the introduction of a unified mathematical model across configurations using a tensor field representation. Some success has also been achieved through combinatorial methods such as grammars [7] and graph-based models [8, 9]. These use a predefined set of system elements that are combined by some algorithm. Both approaches have limitations in the context of early iterations in configuration design. Topology optimization relies on a predefined set of boundary conditions and loads, while combinatorial techniques can only capture configurations accounted for in the model. For further reference on computational synthesis see Antonsson & Cagan [10] or Chakrabarti et al [11].

In product design, configuration methods mostly involve heuristics (e.g., axiomatic design [12] and TRiZ [13]), error avoidance [3, 5], or prescription of specific characteristics under the 'Design for X' moniker [14–16]. Configuration improvement is often prescribed by simply avoiding dependencies among design objectives [3, 12]. Such prescriptions tend to disregard the effect of active (binding) constraints, which often link objectives indirectly through their activity, a situation common in mechanical design.

Dependencies and active constraints are inevitable in practice. Organizations generally remain competitive by integrating more features while improving performance in each new product generation [17], preferably with as little production complexity as possible. Increased dependency comes with increased integration, leading to more trade-offs between objectives [3] and a need for more design iterations [18], ultimately influencing development time and cost.

Formal design optimization techniques offer a rigorous route to the study of trade-offs. Their utility in the design process beyond the identification of the optimum is often touted [19–21]. Yet, they provide little systematic support in identifying when and how to improve the problem formulation by changing the configuration design, as they mostly deal with proportional or parametric changes. In this context, the application of Monotonicity Analysis (MA) [22] might further such understanding. MA is an opportunistic approach used to identify active constraints and allow model reduction by revealing dependencies in the design that are unique to the optimum. Although some work has been done on the application of MA in aiding configuration design (c.f. [21, 23]), its potential remains relatively unexplored.

In view of the above limitations, successful configuration design in practice is largely dependent on the proficiency and experience of the designer [24, 25]. Mnagement of trade-offs [24] and constraints [26, 27] is a key differentiator between novice and experienced designers. Ahmed et al. [24] found that experienced designers were intuitively aware of the trade-offs they need to deal with and were focused on doing so upfront, while novice designers were more prone to trial and error. Without an understanding of how to configure a system in a way that limits trade-offs, inexperienced designers are thus left at a disadvantage. We hypothesize that the methodical approach presented here can provide such insights and guide configuration redesign.

This article presents a systematic, analytical foundation for configuration redesign using *Pareto-set Dependency Analysis* [28]. This analysis relies on multiobjective monotonicity analysis to identify dependencies in the Pareto-set that cause trade-offs while also accounting for the constraints. It derives relationships necessary at optimality and thus helps the designer identify configuration design changes that yield performance improvements beyond what could be reached by mere size and parameter changes. Section 2 presents the theoretical foundations of this work, followed by a description of the methodology in Section 3. Section 4 presents the case study; the Self-Orienting Millimeter-Scale Applicator (SOMA), which is a mechanical pill for the oral delivery of insulin and other drugs [29]. Finally, a discussion and conclusion is offered in Sections 5 and 6, respectively.

2 Theoretical Foundation

Multiobjective design optimization problems are stated in negative-null form as ([19]:

$$\min \quad \mathbf{f}(\mathbf{x}) \tag{1}$$

$$\text{subject to} \quad \mathbf{g}(\mathbf{x}) \leq 0 \tag{2}$$
$$\mathbf{h}(\mathbf{x}) = 0 \tag{3}$$
$$\mathbf{x} \in \mathcal{X} \tag{4}$$

where $\mathbf{f}(\mathbf{x})$ is a vector of objectives f_i, $i = [1,2,...,k]^T$ to be minimised, $\mathbf{x}$ is a vector of real-valued design variables, $\mathbf{h}(\mathbf{x})$, $\mathbf{g}(\mathbf{x})$ are the equality and inequality constraints respectively, and $\mathcal{X}$ is the set constraint that may include additional restrictions besides those of Eq. (2) and (3). If the set constraint is in just the real space, then $\mathcal{X}$ denotes the feasible domain consisting of the $\mathbf{x}$ values fulfilling(2) and (3). The attainable set $\mathcal{A}$ contains all feasible values of $\mathbf{f}(\mathbf{x})$. A point $\mathbf{f}(\mathbf{x}^*) \in \mathcal{A}$ is said to be Pareto-optimal if and only if there exists no other point $\mathbf{f}(\mathbf{x}) \in \mathcal{A}$ such that $\mathbf{f}(\mathbf{x}) \leq \mathbf{f}(\mathbf{x}^*) \wedge f_i(\mathbf{x}) < f_i(\mathbf{x}^*)$. The Pareto set $\mathcal{C}$ containing all Pareto-optimal points lies on the boundary of $\mathcal{A}$ facing the origin, hence it is also referred to as the Pareto frontier. The utopia point F^0 is a k-dimensional point consisting of all the single-objective minima, which lies outside of $\mathcal{A}$ and is often used as a reference for evaluating Pareto points.

The *upper-bound formulation* is one way of identifying the Pareto set (see e.g. [30]). It creates a scalar substitute problem by splitting $\mathbf{f}(\mathbf{x})$ into a single objective function $f(\mathbf{x})$ and a vector of (upper) bound objectives, $\mathbf{c}(\mathbf{x};\boldsymbol{\varepsilon})$:

$$\min \quad f(\mathbf{x}) \tag{5}$$
$$\text{s.j.t} \quad \mathbf{c}(\mathbf{x};\boldsymbol{\varepsilon}) \leq 0 \tag{6}$$
$$\mathbf{g}(\mathbf{x}) \leq 0 \tag{7}$$
$$\mathbf{h}(\mathbf{x}) = 0 \tag{8}$$
$$\mathbf{x} \in \mathcal{X}, \boldsymbol{\varepsilon} \in \mathcal{R}^{k-1}, \tag{9}$$

where $\mathbf{c}$ is a $k-1$ dimensional vector of bound objectives expressed in the form $c_i(\mathbf{x};\varepsilon_i) = f_{i+1}(\mathbf{x}) - \varepsilon_i \leq 0$ or $c_i(\mathbf{x};\varepsilon_i) = \varepsilon_i - f_{i+1}(\mathbf{x}) \leq 0$, $i = [1,2,..,(k-1)]$; $\boldsymbol{\varepsilon}$ is a vector of parameters ε_i in the real space $\mathcal{R}^{k-1}$ for the bound objectives. When $f(\mathbf{x})$ is minimised for given values of ε_i, then the solution $\mathbf{x}^*$ is Pareto optimal if all of the bound objectives are active with non-zero Lagrange multipliers. Pareto points are thus identified by varying $\boldsymbol{\varepsilon}$ systematically between lower and upper limits $\boldsymbol{\varepsilon}_\mathbf{L}$. See [30, 31] for an overview of the upper bound formulation (also known as ε-constraint), the underlying mathematics, and how to define limits for $\boldsymbol{\varepsilon}$. As discussed in [28], this formulation has some computational limitations but its use here benefits the proposed analysis.

Monotonicity Analysis [22] leverages any monotonic behaviour in objective- and constraint functions to reduce optimization models and check their boundedness. A scalar function is monotonically increasing with respect to a variable x, if it holds that $f(x_2) > f(x_1)$ for any $x_2 > x_1$, denoted as $f(x^+)$, and is said to be monotonically decreasing w.r.t. x, if it holds that $f(x_2) < f(x_1)$, denoted as $f(x^-)$. In the presence of monotonicity, the following principles [19] can be used to find active inequality constraints without needing to find the optimum first and prior to any computation:

First Monotonicity Principle (MP1): In a well-constrained minimization problem, every increasing variable is bounded below by at least one non-increasing active constraint.

Second Monotonicity Principle (MP2): In a well-constrained minimization problem every nonobjective variable is bounded both below by at least one non-increasing semi-active constraint and above by at least one non-decreasing semi-active constraint.

If an active constraint can be identified, it can be used to 'partially minimize' the model; namely, solve the constraint function with respect to one of its dependent variables and back-substitute the solution for that variable into the objective and remaining constraint functions, thus reducing the model dimensionality. This process reveals the relationships that *necessarily* exist at the optimum as a consequence of the constraint activity. For simplification and consistency with typical design situations, it is customary in monotonicity analysis to assume that the set constraint is the strictly positive real space $\mathcal{P}$. Hence, from here on, we assume that the feasible domain $\mathcal{X}$ is a subset of $\mathcal{P}$.

In [28], these principles were applied to derive the Multiobjective Monotonicity Analysis (MOMA) process for rigorous identification of the dependencies that cause trade-offs between objectives in problems of the form shown in Eqs. 6-9. In well-bounded problems, a set of theorems and definitions specific to MOMA can be used to simultaneously reduce multiple objectives so that the degrees of freedom remaining in the Pareto set, and the constraints that shape it, are revealed. For the analysis in the present paper, we need the following definition and theorem:

Definition 1 Trade-off and Harmonious Variables
If an objective pair f and c_i have a variable x_1 in common, but differ in monotonicity w.r.t. x_1, e.g., $f(x_1^+)$ and $c_i(x_1^-)$, then x_1 is said to be a trade-off variable, denoted $\underline{\bar{x}}_1$. Correspondingly, an objective pair of like monotonicity w.r.t. a common variable, indicates that the variable is harmonious and can be used to partially minimise both simultaneously.

Theorem 1 Influence of Monotonic Trade-off Variables
In the presence of monotonic trade-off variables, no dominant minimum exists, resulting in a Pareto set. The proof for this is a corollary to MP1.

Proof. Let f_1 be monotonically increasing w.r.t. $x \in \mathcal{P}$ and f_2 monotonically decreasing, and let x be well bounded from above and below. Then by MP1, $\arg\min \mathrm{f}_1(\mathrm{x}) = \underline{x}$, and $\arg\min \mathrm{f}_2(\mathrm{x}) = \bar{x}$, meaning that the minimizers for the two objectives are respectively defined by the *greatest lower bound* (glb)and the *least upper bound* (lub) of x . Hence any feasible value of x will yield a unique Pareto point. ■

Using this basic insight, along with other theorems that ensure correct model reduction [28], MOMA allows identification of the conditions under which the bound objectives are active, i.e., the values of ε that affect the feasible domain of $\mathbf{x}$, as illustrated in Fig. 1. This in turn allows reduction of multiobjective problems to reveal dependencies existing at the Pareto set. Furthermore, determining

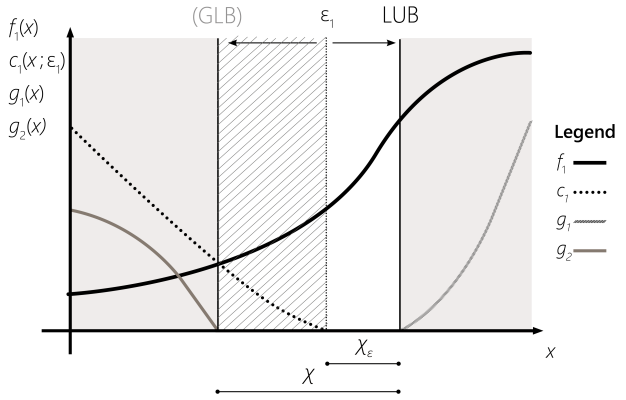


Fig. 1: MOMA allows the partial identification of the Pareto set, finding the values of ε where the bound objectives are active, semi-active, and inconsistent [28].

a bound objective to be active, i.e., $c_j(\mathbf{x},\varepsilon_j)=0$, can be used to optimize out a trade-off variable $\underline{x_n}$, meaning $f(\mathbf{x})$ and $g(\mathbf{x}), c_i(\mathbf{x};\varepsilon)\in D_s(x_n), i\neq j$ become dependent on ε_j through back-substitution. If all $c_i(\mathbf{x},\varepsilon_i)$ are found to be active, i.e., in trade-off with f_1, then the problem is reduced to:

$$\begin{aligned} \text{min.} \quad & \mathrm{U}(f_1^+,\tilde{\varepsilon}) && (10)\\ & f_1(\mathbf{x},\tilde{\varepsilon}) && (11)\\ \text{s.j.t} \quad & \mathbf{g}(\mathbf{x},\tilde{\varepsilon})\leq 0 && (12)\\ & \mathbf{h}(\mathbf{x})=0 && (13)\end{aligned}$$

Here, the bound objectives have been optimized out, causing the back-substitution of ε into the primary objective function f_1 and the inequality constraints $\mathbf{g}$. Treating these parameters as variables, denoted $\tilde{\varepsilon}$, we can apply the ε-monotonicity analysis (εMA) procedure [28] to further our understanding of the Pareto set. The εMA process involves introducing the symbolic cost function $U(f_1,\tilde{\varepsilon})$, which is monotonically increasing w.r.t. the minimization objectives and decreasing w.r.t. maximization objectives, meaning that we can study the bounds of $\tilde{\varepsilon}$ across the Pareto set.

The multiobjective expression $f(\mathbf{x},\tilde{\varepsilon})$ describes the trade-off between $f(\mathbf{x})$ and $\tilde{\varepsilon}$. The constraints dependent on $\tilde{\varepsilon}$, $\mathbf{g}(\mathbf{x},\tilde{\varepsilon})$, referred to as *Pareto constraints*, determine the feasible values of each $\tilde{\varepsilon}_i$. When active, the Pareto constraints bound the Pareto set (hence their name) and they express the trade-offs between the eliminated bound objectives, e.g., in a multiobjective form $g(\mathbf{x},\tilde{\varepsilon_i},\tilde{\varepsilon_j})$, or if they share any of the remaining design variables $\mathbf{x}$.

Given that the activity of $g(\mathbf{x},\tilde{\varepsilon})$ can be conditional on values of $\tilde{\varepsilon}$ relative to each other, these constraints can be regionally or globally active. Thus, a case-based approach [19] can be applied to look at the local frontiers and vertices of the Pareto set, where the implications of constraint activity can be assessed. Solving $g_i(\mathbf{x},\tilde{\varepsilon})=0$ w.r.t $\tilde{\varepsilon_j}$ yields trade-off expressions revealing the dependencies that exist between the objective specific to different extremities of the Pareto-set. For an objective pair in trade-off, these can be of the form $\tilde{\varepsilon_i}^*(\mathbf{x},\tilde{\varepsilon_j}^-)$ and $\tilde{\varepsilon_j}^*(\mathbf{x},\tilde{\varepsilon_i}^-)$, or $\tilde{\varepsilon_i}^*(\mathbf{x})$ and $\tilde{\varepsilon_j}^*(\mathbf{x})$ where $\overline{\mathbf{x}}\subset\mathbf{x}$, thus revealing the dependencies that cause trade-off between the objectives. In this article, we use these theorems and developments of monotonicity analysis to derive a novel configuration redesign methodology.

3 Configuration Redesign Methodology

We introduce an opportunistic configuration redesign methodology where the designer uses Pareto set dependency analysis to identify directions for improvement. We start with a discussion on how the formal treatment of Pareto set dependency results [28] gives rise to a set of design principles. We then present the redesign methodology as the systematic application of these principles to eliminate dependencies and relax the constraints that limit optimality.

First, a formal definition of *design improvement* is necessary. We employ the notion of *meta-Pareto optimality* [32]:

Definition 2 Meta-Pareto Set
Given Pareto sets $\mathcal{C}_1,\mathcal{C}_2,...,\mathcal{C}_p$ for p configuration solutions for a given design problem, the meta-Pareto set $\breve{\mathcal{C}}$ consists of points within the union of these sets, $\mathcal{C}_\mathcal{U}=\mathcal{C}_1\cup\mathcal{C}_2,\cup...\cup\mathcal{C}_p$, that are Pareto-optimal with respect to the set $\breve{\mathcal{C}}$. A point $\mathbf{f}_*$ is meta-Pareto-optimal if and only if there exists no point $\mathbf{f}\in\mathcal{C}_U$ such that $f_i\leq f_i*$ for all i and that $f_i<f_i*$ for at least one i.

Definition 3 Design Improvement
If a configuration with Pareto set $\mathcal{C}_0$ is redesigned, resulting in a new Pareto set $\mathcal{C}_1$, the redesign is said to be an improvement, if and only if the meta-Pareto set of $\mathcal{C}_0$ and $\mathcal{C}_1$ is identical to $\mathcal{C}_1$, namely, $\breve{\mathcal{C}}=\mathcal{C}_1$, which implies that all of the Pareto points of the original design are at least weakly dominated by the Pareto points of the redesign.

The definition implies that the achievable performance in the new design is at least equal to or better than that of the previous design, w.r.t. all criteria, exemplified in Fig. 2. This formal definition is independent of the design context and the relative importance of the objectives, and it uses quantifiable properties we can employ in deriving rigorous redesign principles. Since optimality is defined only in the context of the particular optimization model [19], there is an implicit assumption that improvement comparisons are made for designs derived from models of similar fidelity.

3.1 Implications of Pareto Set Dependency Analysis

Application of MOMA and εMA uncovers the relationships in the Pareto set that limit optimality and drive trade-offs. Looking more deeply, we can uncover causalities for the shape and position of Pareto sets through key outputs of Pareto-set dependency analysis, as illustrated in Fig.3.

We start by noting that the trade-off variables defined earlier stem either from an inherent dependency between the objective functions or a dependency that exists at the optimum due to active constraints. They have a substantial influence on the Pareto set, which can be understood by considering the effect of their absence in a design problem:

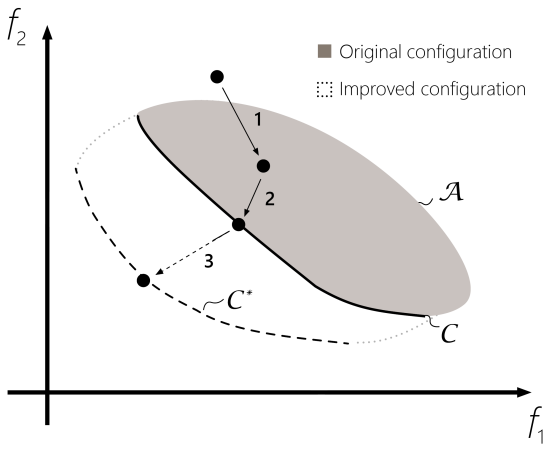


Fig. 2: The difference between constraint satisfaction (1), optimization (2), and configuration improvement (3). In this example the meta Pareto set is identical to $\mathcal{C}^*$

Theorem 2 Existence of Pareto set
If no trade-off variables exist (globally or regionally) after back-substitution of active constraints, then the optimum is a point F∗, rather than a set. Therefore, a Pareto set cannot exist without trade-off variables.

Proof. If $x_i \notin \underline{\bar{x}}$, for any design variable i, then following Theorem 1 and MP1, argmin $f_i(\mathbf{x})$ = argmin$f_j(\mathbf{x})$ for $i \wedge j = 1..k, i \neq j$. Hence, $\dim(\mathbf{x}^*) = [n,1]$, meaning a single dominant optimum exists. ■

Thus, the more trade-off variables exist in a design problem, the larger the distance between the utopia point F^0, and $\mathcal{C}$. Furthermore, recall from Theorem 1 that the optimum of each objective dependent on a trade-off variable $\underline{\bar{x}_i}$, exists at either $\underline{x_i}$ or $\overline{x_i}$ depending on the objective's monotonicity, with any other feasible value of $\underline{\bar{x}_i}$ yielding a Pareto point. As a result, the size of the feasible domain of trade-off variables contributes to the size of the Pareto set.

Finally, being oppositely monotonic, the partial derivatives of an objective pair in trade-off w.r.t $\underline{\bar{x}_i}$ will have an opposite sign across the entire Pareto-set. The larger the difference between these, the larger the slope of the frontier. Thus, the impact of $\underline{\bar{x}_i}$ on the trade-offs between objectives can be worsened by multipliers and divisors.

In summary, the trade-off variables in an optimization problem cause the existence of the Pareto set, determining its span and to some extent its shape. From this observation, we can derive several corollaries which reveal design principles to guide generation of an improved Pareto set.

Corollary 2.1 Separation of Trade-off Variables
If the trade-off variable $\underline{\bar{x}_1}$ affecting the objectives $f_i(x_1^+)$ and $f_j(x_1^-)$, is substituted through design change in one objective by a new variable x_2, such that $\hat{f}_i(x_2^+), x_2 \notin f_j$, and $\min \hat{f}_i \leq \min f_i$, *then* $\operatorname{argmin}\hat{\mathbf{f}}(\mathbf{x}) = \{\overline{x_1}, \underline{x_2}\}$ *while* $\operatorname{argmin}\mathbf{f}(x_1) = x_1 \in \mathcal{X}$. *As a result* $\hat{\mathbf{f}}(\mathbf{x}) < \mathbf{f}(\underline{\bar{x}}_1)$ *for any value of* x_1.

The same would also apply if x_2 were substituted into the problem such that $\hat{f}_j(x_2^-)$, or if the influence of x_1 upon one of the objectives in a multivariate problem was simply eliminated without the introduction of another variable. In such problems, it follows that such changes would result in a new Pareto set $\mathcal{C}_2$ that at least weakly dominates the original Pareto set, $\mathcal{C}_1$. The term separation here is used in the same spirit as in TRiZ [13], reflecting the removal of a dependency.

Correspondingly, the same would occur if a design change is introduced that makes x_1 a harmonious variable without otherwise affecting the objective functions. In fact, any modification of the design problem which eliminates or reduces the influence of the trade-off variable on one objective may result in a new, weakly dominant Pareto set. This is stated formally with the following corollaries.

Corollary 2.2 Flipping Trade-off Variables
If the monotonicity of a trade-off variable $\underline{\bar{x}_n}$ affecting the objectives $f_i(x_n^+)$ and $f_j(x_n^-)$, is flipped in one objective through design change, such that $\hat{f}_i(x_n^-)$, and $\min \hat{f}_i \leq \min f_i$, *then* $\operatorname{argmin}\hat{\mathbf{f}}(\mathbf{x}) = \overline{x_n}$ *whereas* $\operatorname{argmin}\mathbf{f}(x_n) = x_n \in \mathcal{X}$. *As a result* $\hat{\mathbf{f}}(x_n) < \mathbf{f}(\underline{\bar{x}}_n)$.

Corollary 2.3 Scaling Trade-off Variables
If the influence of a trade-off variable $\underline{\bar{x}_n}$ affecting the objectives $f_i(x_n^+)$ and $f_j(x_n^-)$ is scaled through the introduction of an independent variable x_{n1} in f_i such that $\partial \hat{f}_i/\partial x_n < \partial f_i/\partial x_n$ then $\min \hat{f}_i < \min f_i$ *for any value of x_n, reducing the trade-off between f_i and f_j. Correspondingly, if $\partial \hat{f}_j/\partial x_n > \partial f_j/\partial x_n$ then* $\min \hat{f}_j < \min f_j$.

As mentioned, harmonious variables are shared between objectives of like monotonicity, denoted $\underline{\bar{\bar{x}}}$ when the objectives are monotonically decreasing, and $\underline{\underline{x}}$ when they are increasing. For such variables, the glb (for $\underline{\underline{x}}$) or lub (for $\underline{\bar{\bar{x}}}$) is active at all Pareto points. They might be optimized out using MOMA if globally active constraints are identified, or remain in the model if there are regionally active constraints. While trade-off variables create the Pareto set, identifying harmonious variables reveals other useful information:

Theorem 3 Position of Pareto Set $\mathcal{C}$
Harmonious variables, $\underline{\underline{x}}$ and $\underline{\bar{\bar{x}}}$, affect the position of the Pareto set $\mathcal{C}$ relative to the origin. Thus design changes that widen their feasible domains in an improving direction, yield a new strongly dominant Pareto set, $\mathcal{C}_{i+1} < \mathcal{C}_i$.

Proof. Let f_i and f_j depend on $\underline{\bar{\bar{x}}}_1, \underline{\underline{x}}_2, \overline{x_3}$, i.e., $f_i(x_1^-, x_2^+, x_3^+)$, $f_j(x_1^-, x_2^+, x_3^-)$ where $x_1, x_2, x_3 \in \mathcal{P}$. If the problem is well bounded, then by MP1 and Theorem 1, $\arg\min f_i(\mathbf{x}) = \{\overline{x_1}, \underline{x_2}, \underline{x_3}\}$ and $\arg\min f_j(\mathbf{x}) = \{\overline{x_1}, \underline{x_2}, \overline{x_3}\}$. If the active constraints are modified or relaxed, such that $\overline{x_1} < \overline{\hat{x}_1}$ and/or $\underline{\hat{x}_2} < \underline{x_2}$, then $\hat{\mathbf{f}}^*(\mathbf{x}) < \mathbf{f}^*(\mathbf{x})$ for any value of $\overline{x_3}$, given the monotonicity of f_i and f_j. Hence, $\check{\mathcal{C}} = \mathcal{C}^*$. The reverse is true if the active constraints are tightened, such that $\overline{x_1} > \overline{\hat{x}_1}$ and/or $\underline{\hat{x}_1} > \underline{x_1}$. Hence, the harmonious variables and their bounds influence the position of $\mathcal{C}$. ■

Whereas changing the bounds of trade-off variables only enlarges the Pareto set and moves its utopia point, relax-

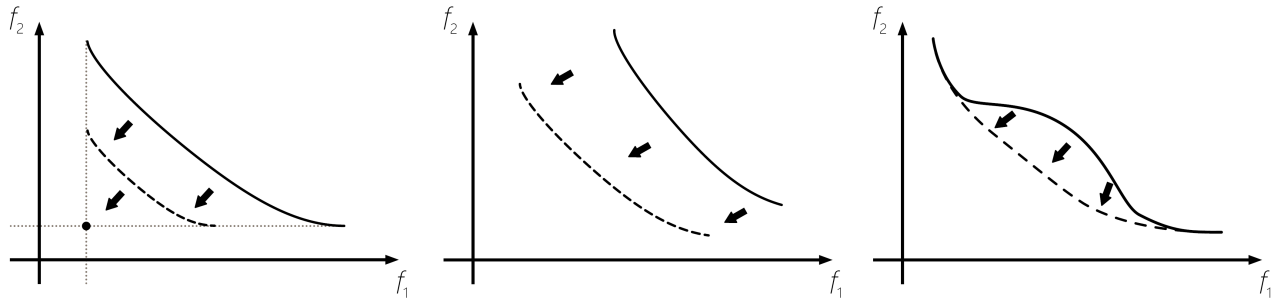


Fig. 3: The relation between analysis outputs and the Pareto set. (*Left*): Trade-off variables result in an optimum that is a set rather than a point; eliminating the underlying dependency brings the set closer to utopia. (*Center*): Harmonious Variables affect the position of the Pareto set; relaxing their constraints shift the set. (*Right:*) Pareto constraints introduce regional relationships that affect the Pareto set; eliminating or relaxing them changes the shape of the set

ing the constraints of harmonious variables results in an improved Pareto set. Furthermore the slope of C will be affected as well, unless x_1 and x_2 influence f_i and f_j equally.

Lastly, Pareto-constraints are a consequence of the systematic reduction of multiobjective problems modelled in an upper bound formulation. When globally active, i.e., for any $\tilde{\varepsilon}_{\mathbf{L}} \leq \tilde{\varepsilon} \leq \tilde{\varepsilon}_{\mathbf{U}}$, Pareto constraints allow the derivation of terms of the form $\tilde{\varepsilon}_i(\mathbf{x}, \tilde{\varepsilon})$, revealing additional trade-off variables while describing the relationship that exist between the objectives at the Pareto-set. In this case, they are merely a representation of trade-off variables, albeit one which may significantly impact the Pareto set. When they are regionally active, however, Pareto constraints reveal regional trade-off variables. This occurs when some of the constraints in the non-reduced model become active for specific values of ε, causing discontinuous trade-offs. In higher dimensional problems, $k \geq 3$, regionally active Pareto constraints might cause a Pareto frontier between an objective pair. Such situations can be studied through a case analysis procedure described in [28]. Thus, Pareto constraints may, when studied, reveal additional trade-off variables or discontinuous relationships such as variables that are in trade-off in specific regions of the Pareto-set, thereby affecting its shape.

In summary, Pareto-set dependency analysis helps explain the relationship between the design problem and the shape of the Pareto set. Thus, we can utilise these theorems and corollaries to derive a set of redesign principles.

3.2 Configuration Redesign Principles

Insights into the relationship between the design problem and the shape of the Pareto set is of substantial value in the synthesis and improvement of configuration designs. As discussed in [28], the Pareto set is created by variables and constraints that are shared between objectives. Some shared variables can be used to improve upon several objectives simultaneously, while others cannot. To a large extent, these relationships are determined by decisions made in conceptual and configuration design. Designers hence need to identify and manage global (i.e. shared variables) and regional dependencies (i.e., shared active constraints) at an early stage to reach *good* configuration designs. We posit that experienced designers apply tacit knowledge of constraints [26] and trade-offs [24] to synthesise and improve configurations. They use this knowledge to configure the components of a system in a way that leverages harmonious variables to achieve a high performance, e.g., placing rotating components as far inside an assembly as possible and load-bearing components as far outside. Similarly, they will attempt to avoid trade-off variables or obviate them through design changes.

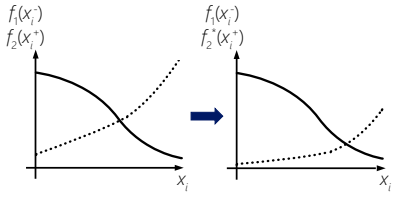
Reaching the required insights is not trivial, especially in highly interdependent systems. Pareto set dependency analysis bridges this gap, providing a causal link between optimality and configuration design limitations. This understanding allows more informed and deliberate identification, prioritisation, and handling of the dependencies that cause trade-offs. The introduced theorems, proofs, and corollaries demonstrate how certain types of model transformation based on the results of this analysis lead to an improved Pareto set. Translating these transformations into specific design changes would mitigate the dependencies that create the Pareto set and relax the constraints that position it, just as experienced designers do through tacit knowledge.

In this spirit, we state four reconfiguration design principles, illustrated in Figs. 4-6, stemming from the theorems and corollaries presented in Section 3.1. When employed in configuration redesign, these principles lead to an associated improvement of the Pareto set. Within each principle, we state a number of more specific strategies stemming from basic model transformations that result in an improved Pareto set. Each represents an alternative way of implementing the principles and corresponds to certain forms of design change. The figures illustrate each of the four principles and the available strategies within each principle:

1. **Align Trade-off Variables**. Reduce or eliminate the effect of trade-off variables on the objectives without impacting their single-objective optima, thereby improving their alignment and the Pareto set, c.f. Theorem 2. This involves eliminating the dependency, making the variable harmonious, or scaling it, c.f. Corollaries 2.1-2.3.
2. **Leverage Harmonious Variables**.Widening the feasi-

Align Trade-off Variables

SCALE

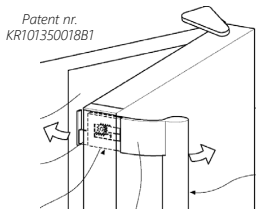


Mathematical transformation
Any design change adding a multiplier or divisor (a parameter or variable) to a trade-off variable in one objective allows scaling of the trade-off. Examples of such transformations include:

$$f_1(x_i^+) = x_i^2 \quad \rightarrow \quad f_1^*(x_i^+, x_j^-) = x_i^2/x_j$$
$$f_2(x_i^-) = -x_i \quad \rightarrow \quad f_2(x_i^-) = -x_i$$
$$f_1(x_i^-) = -x_i \quad \rightarrow \quad f_1^*(x_i^-, x_j^-) = -x_i x_j$$
$$f_2(x_i^+, x_j^-) = x_i - x_j \quad \rightarrow \quad f_2(x_i^+, x_j^-) = x_i - x_j$$

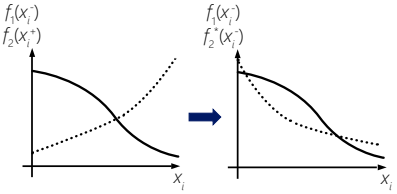
Typical design changes
Trade-offs can be reduced through a wide range of design changes that scale one objective but not the other, ranging from the addition of lubrication to new subsystems, features, and interfaces. The introduction of gearing and mechanical leverage in general, load balancing, lubrication, and intermittent kinematic constraints (e.g. for nonlinear stiffness), are all examples of scaling solutions.

Related heuristics: Amplification and filtering [32], manage friction [3,5,13], local quality in TRIZ [12], decoupling [11], and leverage/gearing [2,5].



Example: *Fridge door mechanisms*
To ensure efficient cooling, refrigerator doors are tightly sealed when closed, which is achieved with pretension of the door with a rubber seal. Combined with negative pressure inside the fridge due to cooling, this results in a high opening force. Several designs scale down this *efficiency vs. opening force* trade-off e.g. with auxiliary opening mechanisms and pivoting lever handles.

FLIP MONOTONICITY



Mathematical transformation

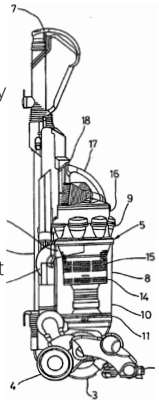
$$\begin{matrix} f_1(x_i^-) \\ f_2(x_i^+) \end{matrix} \longrightarrow \begin{matrix} f_1(x_i^-) \\ f_2^*(x_i^-) \end{matrix}$$

Any design change that inverts the monotonicity of one objective w.r.t. a trade-off variable, while the rest are unchanged, effectively makes the variable harmonious. In nonlinear terms this might be achieved by changing the bounds of other variables that act as multipliers or divisors to the trade-off variable.

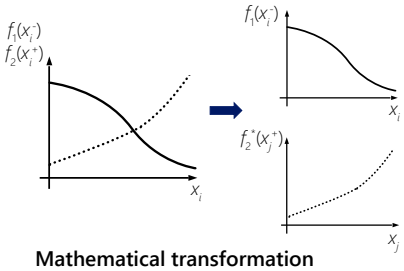
Typical design changes
While somewhat challenging, making a trade-off variable harmonious can be achieved in certain circumstances, especially if the dependency stems from an active constraint. Changes such as the inversion of components and interfaces, *"self-helping"* systems, redistribution of sub-functions, changes in working directions and load paths, or the use of a different working principle, can result in a change in monotonicity.

Related heuristics: "The other way round", nested doll, and self-help [12] Principles of self-help and force transmission [2],

Example: *Dyson Vacuum*
Bag-based vacuum cleaners are generally affected by a trade-off between filtration quality and suction pressure; the tighter the filter the larger the pressure loss. Vacuums that rely on cyclonic separation where filtration increases with the pressure, get around this issue. While the example is conceptual, as it relates to a change in filtration principle, it illustrates the general idea.



SEPARATE



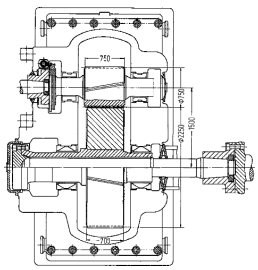
Mathematical transformation

$$\begin{matrix} f_1(x_i^-) \\ f_2(x_i^+) \end{matrix} \longrightarrow \begin{matrix} f_1(x_i^-) \\ f_2^*(x_j^+) \end{matrix}$$

Any design change that makes an objective independent of a trade-off variable - either through substitution or elimination, mitigates the trade-off, unless the objectives share additional trade-off variables.

Typical design changes
Separation is a widely used principle, involving changes such as the splitting parts, change in working axis and load direction, parallel subsystems, asymmetry, or the avoidance of "unintended" dependencies through exact constraint design. It may result in an increased number of parts, but can also involve the redistribution of functionality amongst the parts of the system. Unlike in other frameworks, the approach here is to only apply separation to trade-off variables.

Related heuristics: Independence axiom [11], division of tasks [2], separation in space, time, or condition [12].

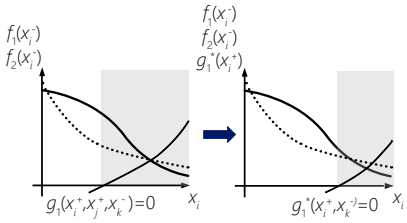


Example: *Siemens-Maag Gearbox* [2]
This gearbox drive shaft design described by Pahl and Beitz [2], is a prime example of this principle. The drive shaft has been split in two to eliminate a trade-off between efficiency and wear; the stiff outer shaft transmits the torque from the gears, while the flexible inner shaft is free to absorb oscillations, protecting the gears.

Fig. 4: The strategies within Principle 1: Reduce or eliminate the impact of a trade-off variable upon an objective pair

Leverage Harmonious Variables

SHIFTED BOUNDS



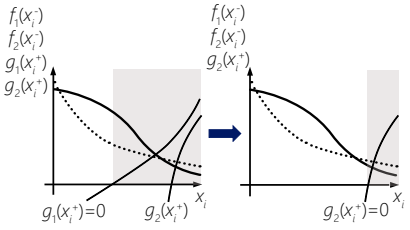
Mathematical transformation
Any design change that shifts the *glb*/*lub* of a harmonious variable, improves the optimum of its dependant objectives. This involves eliminating increasing contributors, introducing additional decreasing contributors, or scaling parts of the constraint. This can also scale a trade-off, when the harmonious variable is a multiplier or divisor of a trade-off variable.

Typical design change
In configuration design terms, these changes are specific to the type of constraint. Generally speaking this is oft matter of positioning components in an assembly in the most beneficial way - e.g. locating parts with decreasing variables as far *inside* an assembly as possible and increasing variables on the outside. Further, it involves designing to avoid unnecessary contributors to the active constraint, e.g. stress concentrations and associated loads in structural constraints.

Related heuristics: Reduce information content [11], Principle of balanced forces [2], Minimise tolerance paths [5]

Example: *Mazda Skyactiv-G* [33]
In the design of combustion engines, thermal efficiency increases with the compression ratio. Yet, this ratio cannot be increased beyond a point where knocking occurs, which is in part driven by residual gas after combustion. Most petrol engines hence have a ratio between 8:1-12:1. In the Skyactiv engine, Mazda pushed this ratio 14:1, using a longer exhaust manifold, increasing gas scavenging, and shifting the knocking constraint.

CONSTRAINT RELAXATION

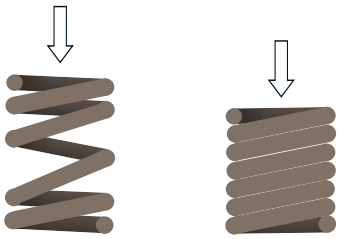


Mathematical transformation
When a harmonious variable is actively (but not critically) constrained, we might try to change the configuration design in a way that eliminates the active constraint. This shifts the *glb* or *lub* of the variable to the next constraint, improving the optimum of all its dependent objectives.

Typical design changes
As with shifted bounds, the changes required to eliminate a constraint, are contextual. Examples include changes aimed at redirection of force paths to eliminate a load case, a new part structure to avoid certain parts being bound by limiting geometric constrains (e.g. one part inside another), a change in assembly sequence to avoid some alignment constraint.

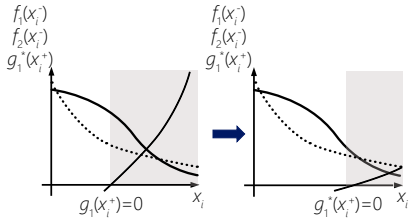
Related heuristics: Vary the structure of main elements [14,], redirect load path [5], merge parts [5,13], shielding [33].

Example: *Spring strength at block*
A well known example of load path redirection, compression springs are self-



reinforcing when deformed to their block length. In design applications where a maximum load resistance is desired, a spring design that is deformed to its' block length rather than to its' elastic yield limit is far stronger. Introducing such a change to a design, is equivalent to eliminating the yield constraint driven by shear stress.

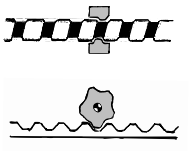
NEW FUNCTIONAL FORM



Mathematical transformation
A complete change in functional form of an active constraint, may yield a widened feasible domain. This is distinct from *shifted bounds*, in that it involves the entire function, and may hence result in changed constraints, monotonicity, exponents, and so on.

Typical design changes
Such a drastic model change will most likely probably require substantial design change, e.g. a change in components, working principles, and/or the physics of the problem. Examples of such include a change in production process, the separation or combination of parts, a change in the realisation of a given sub-function, a change in load type and distribution, and so on.

Related heuristics: Design for pure compression and tension [5], Select rotary over linear motion [3,5] Self-help [3,12]

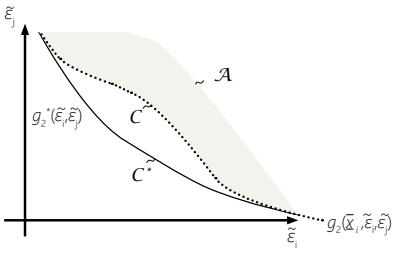


Example: *Rotary to linear movement*
A rack and pinion and a lead screw fundamentally meet the same functional purpose - to convert rotation into linear motion, or vice versa. Yet, what is superior, depends on the objectives, primarily due to quite different constraints involved in their design. For instance, the rack for instance only slides, and as a result the mechanical stress expressions are quite different, compared to the rotating screw, which is why they are commonly used for high load applications.

Fig. 5: The strategies within Principle 2: Increase the influence of harmonious variables

Relax Pareto Constraints

SHIFTED PARETO CONSTRAINT



Mathematical transformation
The elimination of increasing contributions to Pareto constraints shifts the frontier regionally or globally. Transformations include the elimination of parametric-, variable-, or objective contributions, e.g.:

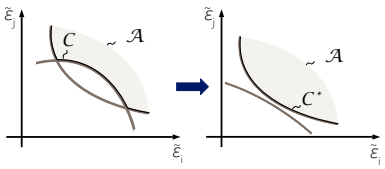
$$g(\mathbf{x}, \tilde{\epsilon_i}^+, \tilde{\epsilon_j}^-) \longrightarrow g^*(\mathbf{x}, \tilde{\epsilon_j}^-)$$
$$g(x_1^+, x_2^-, \tilde{\epsilon_i}^+) \longrightarrow g^*(x_2^-, \tilde{\epsilon_i}^+)$$
$$g(\mathbf{x}, \tilde{\epsilon_i}^+; P^+) \longrightarrow g^*(\mathbf{x}, \tilde{\epsilon_i}^+)$$

Typical design changes
In many ways, the design changes and model transformations involved here, resemble those of *Leverage Harmonious variables*. Shifting Pareto constraints is tantamount to reducing the equilibrium that exists between the objectives in certain (or all) regions of the Pareto set, due to the activity of constraints. Thus, this might involve rearrangement of parts, change in load distribution, and so forth.

Related heuristics: Reduce information content [11], principle of balanced forces [3], vary the structure of elements [14], redirect load path [5]

Example - *Additive Manufacturing and Topology Optimization (TO)*
In industrial practice, TO efforts are usually actively constrained by material and the manufacturing constraints. In this context, the utility of additive manufacturing is broadly cited, as it essentially shifts several manufacturing constraints, e.g. allowing hollow geometry and undercuts, and shaping not being limited by tooling directions. This has allows increasingly light load bearing components, reducing the trade-off between stiffness and mass. While this is more a process change than a design change, it serves to illustrate the model transformation.

DEPENDENCY REDUCTION



Mathematical transformation
Reduction of the dependencies between Pareto constraints, that either result in trade-off variables, variables with empty feasible domains beyond the Pareto set (i.e. *two-sided failure*), or regional bounds for $\tilde{\boldsymbol{\epsilon}}$, will change the optimal set. An example of such a transformation is:

$$\begin{matrix} g_1(x_1^+, \tilde{\epsilon_i}^-) \\ g_2(x_1^-, \tilde{\epsilon_j}^+) \end{matrix} \longrightarrow \begin{matrix} g_1(x_1^+, \tilde{\epsilon_i}^+) \\ g_2^*(x_2^-, \tilde{\epsilon_j}^-) \end{matrix}$$

Typical design changes
The design changes and model transformations involved here, resemble those of *Align trade-off variables*. The difference is that these might be regional trade-off variables. Hence, it is equally impactful to introduce changes to the eliminated active constraints that contribute to the Pareto constraint, creating the dependencies. Examples include inverting components and interfaces, eliminating load cases, change in working axis and load direction.

Example: *FlexTouch Safety Mechanism*
Insulin pens cannot be dialled to a dose setting beyond the amount of insulin left. An "*end of content*" locking mechanism prevents the user from receiving a smaller dose than has been set. Such locks need

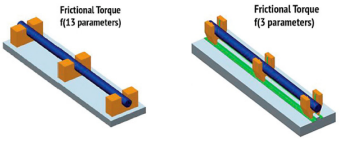
to withstand substantial loads when users unknowingly attempt to get beyond this limit. Ultimately, this affects the achievable combination of device size and dose setting torque (which is important to users with limited dexterity). In the FlexTouch® device, the dial is connected to the dose setting mechanism via a flexible spline connection, which disengages if the user attempts to set a dose beyond what is left. No load is transferred, protecting the pen from overloading, and eliminating a dependency between the size and torque caused by the yield constraint.

Eliminate Parasitic Contributions

Mathematical transformation
This category is too broad to provide a universal mathematical transformation, but all the sub-types of transformations are well known. These are the removal of parametric and scalar contributions which increase the optimal value of one or more objectives (without decreasing any), and the elimination of harmonious variables that are bound in such a way that they cannot be *leveraged*. These may be unintended contributions (i.e. from design error), or contributions involving active constraints with little room to introduce further relaxation (e.g. a wall thickness constraint).

Typical design changes
There are many types of contributions that are parasitic. Examples the negative impact of undesired vibrations, electromagnetic fields, heat, parasitic loads, friction, unintended contact points, and manufacturing and assembly features. Design changes mostly involve efforts to remove these effects from performance critical part geometries or locations in the assembly. E.g relocating assembly features to another cross section.

Related heuristics: Exact constraint design [3], reduce information content [11], avoid associated loads [13], shielding [34].



Example: *Overconstrained axle* [34]
An oft cited example in kinematic design, over-constrained axles cause major issues w.r.t. production and efficiency. The typical design error is to increase system stiffness by introducing more radial bearings, resulting in statical indeterminacy. This issue can be reduced or resolved entirely by using fewer or different bearings.

Fig. 6: The strategies within Principles 3 & 4: Reduce regional trade-offs and eliminate parasitic contributions

ble domains of harmonious variables in the improving direction, per Theorem 3. This involves design changes that modify or delete the constraints that bound harmonious variables, striving towards letting $\overline{\overline{x}} \to \infty \wedge \underline{\underline{x}} \to 0$.

3. **Relax Pareto Constraints**. Relax globally active Pareto constraints, thereby aligning trade-off variables. Relax regionally active Pareto constraints, or eliminate the inconsistencies that exist between them beyond the Pareto set (i.e., in the infeasible region). This might change or eliminate the Pareto frontiers between certain objectives.
4. **Eliminate Parasitic Contributors**. Consider situations where it is not possible to widen the feasible domain of harmonious and independent variables; e.g. when their bounds represent unmodelled objectives or physical phenomena that cannot be circumvented. Such situations can introduce parametric or scalar contributions to the objectives that worsen their optima. Therefore, it may be better to eliminate the influence of these variables on the objectives rather than relax the constraint.

These principles and associated strategies relate to specific variables and constraints. They can be applied recursively to improve a configuration beyond the identified optima. Recall that Pareto constraints are representations of trade-off variables. Furthermore, parasitic contributions are harmonious variables that are bound in a manner that prevents them from being leveraged to move the Pareto set. As such, Principles 3 and 4 are special cases of Principles 1 and 2 respectively.

The strategies are very general in that they apply to any design that has been studied through Pareto set dependency analysis. Their specificity becomes evident when considering a specific design problem. Just as design 'goodness' is contextual to the objectives at hand, so are the design improvements. Hence, the strategies are not intended for use in initial configuration design; rather, they motivate designers to identify improvements after careful analysis. Thus, the process of optimization becomes a driver for redesign.

As Figs.4-6 illustrate, the forms of design change involved are common in product design. The redesign strategies are related to well-known design heuristics, albeit with key differences. First, they are opportunistic but have a rigorous foundation and are hence valid independent of context. Second, they are applied following Pareto-set dependency analysis, letting designers rely on analysis results rather than intuition to identify which heuristic to apply where. Heuristics such as *separation*[13] and *independence* [12] or *division of tasks* [3] for instance, prescribe avoidance of dependency. As Section 3.1. shows, this is actually only relevant for trade-off variables when aiming to improve performance.

3.2.1 Sample Problem

In [28], we used a sample problem to demonstrate the application of MOMA to reveal hidden trade-off variables:

$$\text{min.} \quad f_1(x_1^+, x_2^-, x_3^+) = x_1^2 - x_2 + x_3 \tag{14}$$

$$\text{s.j.t} \quad c_1(x_2^-, x_4^-, x_5^+; \varepsilon_1^-) = \frac{1}{x_2} - x_4^2 + 2x_5 - \varepsilon_1 \leq 0 \tag{15}$$

$$g_1(x_1^-, x_4^+) = 2x_4 - x_1 \leq 0 \tag{16}$$

$$g_2(x_2^+, x_3^-) = x_2^2 + 4x_2 - 2x_3 \leq 0 \tag{17}$$

$$g_3(x_2^+, x_4^+) = x_2^3 + 2x_4 - P_1 \leq 0 \tag{18}$$

$$g_4(x_5^-) = 10 - x_5^2 - 3x_5 \leq 0 \tag{19}$$

$$\varepsilon_L \leq \varepsilon_1 \leq \varepsilon_U \tag{20}$$

$$\mathbf{x}, \varepsilon_1 \in \mathbf{P} \tag{21}$$

This problem was reduced using MP1 and Theorem 1, revealing that all of the degrees of freedom are trade-off variables, due to g_1, g_2, and g_4 being critical. The resulting back-substitution of $x_1^* = 2x_4$, $x_3^* = \frac{1}{2}x_2^2 + 2x_2$, and $x_5^* = 2$, yields:

$$\text{min.} \quad f_1(x_2^+, x_4^+) = 4x_4^2 + \frac{1}{2}x_2^2 + x_2 \tag{22}$$

$$\text{s.j.t} \quad c_1(x_2^-, x_4^-; \varepsilon_1^-) = \frac{1}{x_2} - x_4^2 + 4 - \varepsilon_1 \leq 0 \tag{23}$$

$$g_3(x_2^+, x_4^+) = x_2^3 + 2x_4 - P_1 \leq 0 \tag{24}$$

$$\varepsilon_L \leq \varepsilon_1 \leq \varepsilon_U \tag{25}$$

We can use this same problem to illustrate some of the underlying model transformations involved in the redesign principles. In this problem it is clear that the span of the Pareto set between will be defined by $\overline{x_2}$, $\overline{x_4}$ g_3 and ε_L. While x_2 and x_4 are trade-off variables, x_1, x_3 and x_5 are harmonious variables, albeit not shared between objectives. The activity of g_1 makes x_4 a trade-off variable, the activity of g_2 makes x_2 a trade-off variable, while the activity of g_4 introduces a parasitic contribution of +4 to c_1. Thus, following the optimality principles, there are different routes improvement to be explored in a design change process:

- $\overline{x_4}$ As it is caused by g_1, we could either try to apply scale, flip and separate principles to x_4 in g_1, or to x_1 in f_1.
- $\overline{x_2}$ As it is caused by g_2, we could either try to apply the scale, flip and separate principles to x_2 in g_2, or to x_3 and x_2 in f_1.
- $\underline{\underline{x_5}}$ We can either remove its influence on c_1 or relax g_4.

There are many options available and, in a design context, some of these would be more practical than others. If we imagine that we were able to identify design changes that substitute x_2 in g_2 with a new variable, x_6, and substitute x_4 in g_1 with x_5, the resulting reduced problem becomes:

$$\text{min.} \quad f_1(x_2^-) = 16 - x_2 + \frac{1}{2}x_6^2 + 2x_6 \tag{26}$$

$$\text{s.j.t} \quad c_1(x_2^-, x_4^-; \varepsilon_1^-) = \frac{1}{x_2} - x_4^2 + 4 - \varepsilon_1 \leq 0 \tag{27}$$

$$g_3(x_2^+, x_4^+) = x_2^3 + 2x_4 - P_1 \leq 0 \tag{28}$$

$$\varepsilon_L \leq \varepsilon_1 \leq \varepsilon_U \tag{29}$$

With these changes, the problem is now poorly bounded, meaning a constraint bounding x_6 needs to be introduced. When this is done, it is simple to evaluate whether the

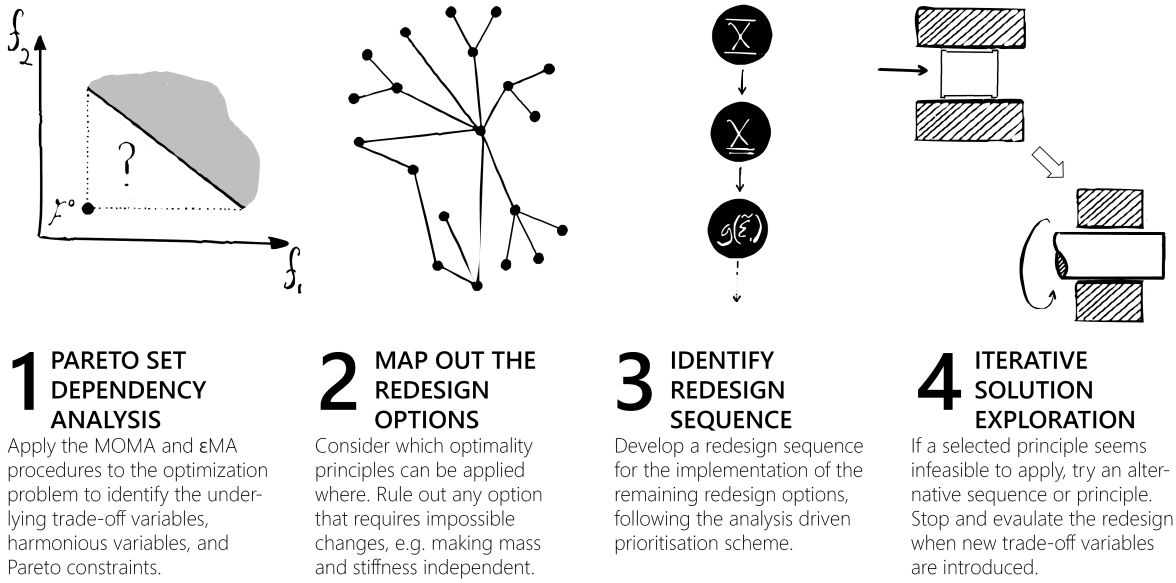


Fig. 7: Configuration Redesign Process - A redesign procedure supported rigorous analysis

changes are an improvement by inspecting the bounds. So long as $16 \leq 4\underline{x_4}^2$, and $\underline{x_6} \leq \underline{x_2}$, not only is the optimum of the new design a dominant minimum rather than a set, it also dominates the Pareto set of the old design. This illustrates how we utilise the configuration redesign principles to steer the exploration of design changes and the output of the original MOMA to evaluate the effect of said changes, namely, the concurrent evaluation and exploration of design changes.

3.3 Configuration Redesign Process

The previous sections may seem excessively formal compared to many design frameworks. However, without the insights MOMA and εMA provide, one might introduce changes to eliminate dependencies or relax constraints that have no bearing on the Pareto set or even accidentally worsen the set. Ultimately, the above principles come down to a more targeted approach for dependency reduction and constraint relaxation. These are design practices that are already widely advocated [2, 3, 12, 13, 18, 26].

As alluded to, the strategies within the same principle are mutually exclusive. For example, we cannot make an objective function independent of a trade-off variable through separation while also scaling the same variable. Depending on the problem, some design changes are also more influential or easier to implement than others. Thus, it is beneficial to map out all options for improvement after analysis and select the most promising ones, rather than randomly applying the principles. While the strategies and underlying principles have a quantitative foundation, the designer must still determine which principle to apply to each variable and constraint, and in which sequence. As summarized in Fig. 7, we thus propose a systematic configuration redesign procedure involving said mapping and prioritization steps between analysis and design change. A critical element in this is the use of a prioritization scheme in Step 3 to identify a redesign sequence. We suggest the following scheme, which is determined by two factors; the magnitude of the potential influence of the change and the ease of implementation:

1. Eliminate parasitic influences.
2. Leverage the harmonious variables, attempting *relaxation* rather than *shifted bounds* when possible. Only leverage the variables that:
   - influence multiple objectives,
   - have a multiplying effect on a trade-off variable.
   - are bound by a constraint with a comparatively high Lagrange multiplier.
   - are actively constrained in a manner that introduces a new trade-off variable or a contribution to a Pareto constraint
3. Relax Pareto constraints that depend on more than one $\tilde{\varepsilon}$ variable and/or are globally active
4. Align trade-off variables in an order based on the number of influenced objectives and on the relationship between F^* and $\underline{\mathbf{x}}^*$. To avoid increasing system complexity, apply flipped monotonicity over the other strategies, and separate over scale unless separation only is possible through the introduction of new variables.
5. Leverage remaining harmonious variables and relax remaining Pareto constraints

The underlying logic behind this scheme is to ensure that independent issues and design errors (i.e., parasitic contributions) are handled first, followed by the changes that result in the largest improvement to the Pareto set. The step order is defined based on the observation that the globally active Pareto constraints and the harmonious variables in Step 2 will, in most cases, exceed the influence of single trade-off variables. Alternatively, one could base the sequence on

	Part name	Suffix
1	Top Housing	*H*
2	Hub	*NH*
3	Plug	*P*
4	Power Spring	*PS*
5	Sealing ring	*S*
6	Needle	*N*
7	Valve	*V*
8	Base housing	*B*

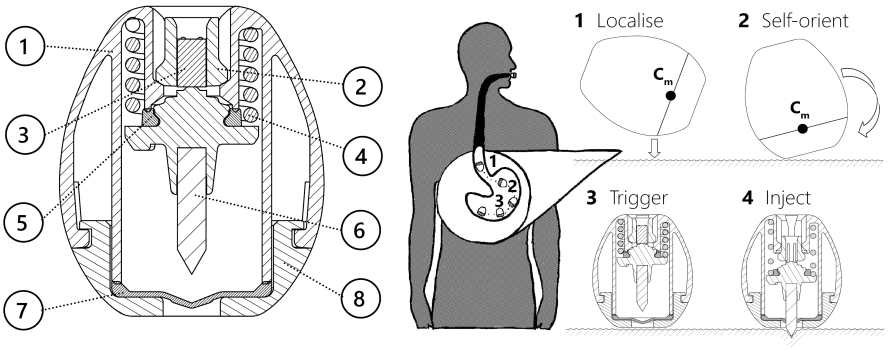


Fig. 8: An overview of the components and functionality of the SOMA device (adapted from [29])

objective weighting. The overall process is suggested to increase the likelihood of successful redesigns. It can be used iteratively and requires a combination of analysis, qualitative reasoning, engineering judgement, and creativity.

4 Case - Design of the SOMA Device

The **S**elf **O**rienting **M**ilimeter-Scale **A**pplicator (SOMA), is a medical device for oral delivery of pharmaceutical protein compounds such as insulin. Such compounds cannot otherwise be administered orally as the gastric system breaks down large proteins by its nature, and are thus administered using needle-based injection devices today. First described by Abramson et al. [29], SOMA was still in early development at the time of the present study. When swallowed, SOMA falls into the stomach, where it self-orients into a stable position on the lining of the stomach thanks to its shape and mass distribution. Once oriented, a compression spring (4 in Fig. 8) injects a milipost of pure insulin (6) or another Active Pharmaceutical Ingredient (API), penetrating into a deep enough tissue layer to reach capillaries, resulting in systemic uptake as the milipost dissolves. This injection is triggered by a plug (3) which dissolves upon contact with liquid, allowing the compliant snap features on the hub component (2) to pass through a ratchet interface on the top housing (1).

The SOMA device presents a number of interesting design challenges. It must reliably deliver a large enough amount of API to meet the dosage needs of patients without compromising the self-orientation performance or injection depth, while being small enough to be swallowed without discomfort. Hence, SOMA was used in [28] to demonstrate the use of Pareto set dependency analysis, applying it to a 4-objective optimization model in upper-bound form:

$$\min \quad f_1(\mathbf{x}) = -\frac{\sum_{p=1}^{p=8} m_p \cdot (C_p + Z_p)}{(l_{t1} + l_{t2} + l_{b1}) \cdot \sum_{p=1}^{p=8} m_p} \tag{30}$$

$$\text{s.j.t} \quad c_1(\mathbf{x};\varepsilon_1) = d_{t1} - \varepsilon_1 \leq 0 \tag{31}$$

$$c_2(\mathbf{x};\varepsilon_2) = \varepsilon_2 - \rho\frac{\pi}{4} d_{n1}^2 \left(l_{n1} + \frac{1}{3} \cdot l_{n2} \right) \leq 0 \tag{32}$$

$$c_3(\mathbf{x};\varepsilon_3) = \varepsilon_3 - \sqrt{2\left(g + \frac{F_s}{m_{acc}} \right) z_{acc}} \leq 0 \tag{33}$$

$$\mathbf{g}(\mathbf{x}) \leq 0 \tag{34}$$

$$\mathbf{h}(\mathbf{x}) = 0 \tag{35}$$

$$\mathbf{x}, \varepsilon \in \mathbb{P} \tag{36}$$

where f_1 is a self-orientation objective, maximising the normalized distance, Z_{cm}, between the top of the device and the system centre of mass, C_m. This contains intermediate functions; m_p describing the mass of each part in the device, C_p the centre of mass in each part, and Z_p the vertical position of each part. c_1 is the bound size objective, minimizing d_{t1}, as pill swallowability is proportional with their minor diameter [36]. c_2 is the bound API capacity objective, maximizing the mass of the needle. Finally, c_3 is the bound velocity objective, maximising the velocity of impact between needle and tissue. Here, F_s is a nonlinear expression for the accelerating force, m_{acc} the mass that is accelerated, z_{acc} the stroke between the initial position of the needle tip and the gastric tissue, and g is gravity. Constraints such as geometric fits, manufacturability, and structural load cases, are represented by $\mathbf{g}(\mathbf{x})$, while $\mathbf{h}(\mathbf{x})$ mostly accounts for the shape of the device. In this paper, we demonstrate the application of the redesign methodology, using the analysis results presented in [28], which are described briefly for the sake of exposition.

4.1 Results of Pareto-set Dependency Analysis

The optimization results [28] (shown in Fig. 11) revealed the trade-offs involved in the design of SOMA. The size-related trade-offs are of special importance, as the US FDA recommends that the minor diameters of capsules do not exceed 8.35mm to avoid medical complications [36]. While this is infeasible in the current design, it is possible to stay below Ø9.91mm, the largest standard capsule size. However, this comes at the cost of self-orientation and impact velocity. Applying MOMA revealed several trade-off variables, see Table 1, such as the spring wire diameter (d_{ps2}), needle length (l_{n1}), position of the split between

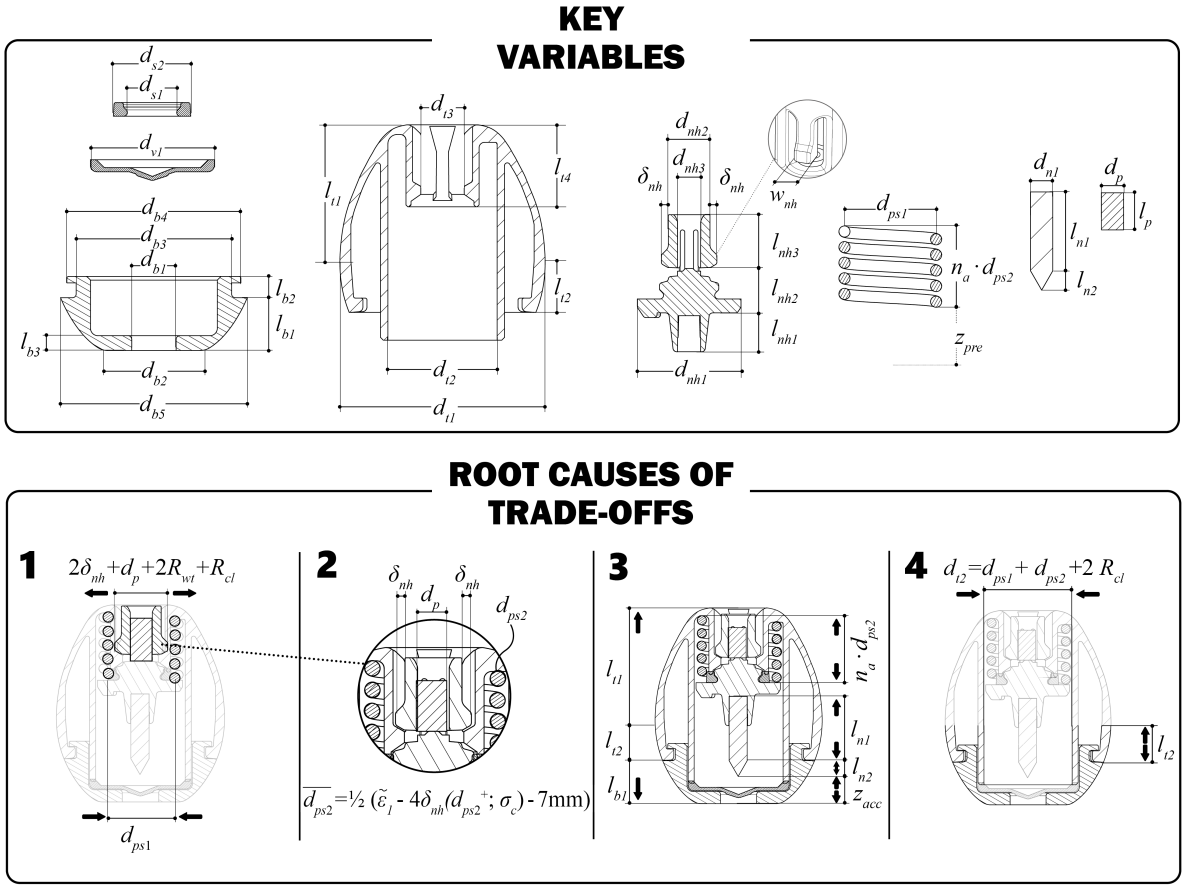


Fig. 9: An overview of the design variables in the SOMA and the key drivers of trade-off identified through analysis in [28]

top housing and base (l_{t2}), and device diameter (d_{t1}). Variables such as spring coiling diameter (d_{ps1}), number of active windings (n_a), and trigger arm overlap (δ_{nh}) were found to be harmonious. Combined with εMA post computation, the causes of the trade-offs (also shown in Fig. 9) become clear:

1. **Radial fit**: The glb of spring coiling diameter, d_{ps1} is determined by a constraint stemming from that the spring needs to fit around the trigger, $g_5 = d_p + 2\delta_{nh} + d_{ps2} - d_{ps1} + 6R_{cl} + 4R_{wt} \leq 0$. Here, R_{cl} and R_{wt} are radial clearance and wall thickness parameters respectively. Back-substituting this glb and the bound size objective, $c_1 = d_{t1} - \varepsilon_1 \leq 0$, yields:

$$\text{min.} \quad f_1(l_{t2}^-, d_{t3}^+, d_{ps2}^+, \delta_{nh}^+, d_p^+, \tilde{\varepsilon_1}^-) \tag{37}$$

$$\text{s.j.t.} \quad c_3(l_{t2}^-, d_{ps2}^-, d_p^+, \delta_{nh}^+, \tilde{\varepsilon_1}^-; \varepsilon_3) \tag{38}$$

$$g_1(\tilde{\varepsilon_1}^-, l_{t2}^+, d_{ps2}^+, d_p^+, \delta_{nh}^+) = 2d_{ps2} + 2\delta_{nh} + d_p$$

$$7mm - \sqrt{\frac{2(C_T\tilde{\varepsilon_1} - l_{t2})\tilde{\varepsilon_1}}{C_T} - \frac{(C_T\tilde{\varepsilon_1} - l_{t2})^2}{C_T^2}} \leq 0 \tag{39}$$

where $\tilde{\varepsilon_1}$ is the device diameter objective-variable and C_t is the device height-width ratio parameter. The radial fit constraint g_1 which stems from the fits between the upper and lower housing, is now a Pareto constraint. When g_1 is active, any increase in spring wire diameter results in an increase in device size, a reduction in trigger overlap, or a reduction of l_{t2}. In [28], g_1 was indeed found active at the 3-objective Pareto frontier between f_1, c_1, and c_3, and violated in the region between the frontier and the utopia point. The constraint essentially prevents the device size from being minimized without reducing the space available for the spring and trigger, or shifting the housing snap upward to the widest point of the device meaning $l_{t2} = \overline{l_{t2}} = 0$, which in turn moves the centre of mass upwards, worsening self-orientation.

2. **Boundedness of the trigger**: Prior to computation, the harmonious variables d_p and δ_{nh} were bound by a conditionally critical set of constraints. Looking at constraint activity at the bi-objective Pareto frontier between size and velocity, revealed a locally active glb of d_p, $g_{10} = 2\delta_{nh} - d_p - R_{cl} \leq 0$, which prevents the trigger arms from colliding with each other thereby jamming the device. Further, a stress criterion for the trigger interface, $g_{11}(d_{ps2}^+, \delta_{nh}^-)$, is globally active and is now critical w.r.t. δ_{nh}. As it has no closed-form solution, its implicit solution and back-substitution into Eq.39 yields

$$g_1(\tilde{\varepsilon_1}^-, l_{t2}^+, d_{ps2}^+) = 2d_{ps2} + 4\delta_{nh}(d_{ps2}^+; \sigma_{IF}) + 7\text{mm} - \sqrt{\frac{2(C_T\tilde{\varepsilon_1} - l_{t2})\tilde{\varepsilon_1}}{C_T} - \frac{(C_T\tilde{\varepsilon_1} - l_{t2})^2}{C_T^2}} \leq 0 \tag{40}$$

where σ_{IF} is the allowable interface stress in the trigger. Inserting this into the glb of the spring coiling diameter yields $\underline{d_{ps1}} = d_{ps2} + 4\delta_{nh}(d_{ps2}^{+};\sigma_{IF}) + 6R_{cl} + 4R_{wt}$. This implies that the activity of g_{10} and g_{11} multiplies the influence of d_{ps2} upon the trade-off of velocity against size and self-orientation, as any increase in d_{ps2} results in an increase in both d_p and δ_{nh}, all contributing to size and mass. Further, any decrease in d_{ps1} also decreases the size of the load-bearing area in the trigger. These dependencies, specific to the Pareto-set, mean that the spring force can only be increased to a certain point for a given device size. Beyond this point, the device would fail due to high static interface stress or simply not fit together radially.

3. **Vertical fit of internal components**: The impact velocity is determined by the force profile exerted by the spring, system mass and frictional resistance, and the acceleration stroke distance between the tip and tissue (z_{acc}). Given that the internal parts in the SOMA device are mounted in a vertical series, z_{acc} is involved in the following constraints:

$$\begin{aligned} h_8 = R_{wt} + (n_a + n_d)d_{ps2} + l_{nh2} + l_{nh1} + l_{n1} + l_{n2} \\ + z_{acc} - l_{b1} - l_{t2} - l_{t1} = 0 \\ g_{20} = l_{b3} + l_{v1} + P_{tol} - z_{acc} \leq 0 \end{aligned} \quad (41)$$

Through MOMA and computation, several constraints were found to be active, meaning that $l_{nh1} = 1.5\text{mm}$, $l_{nh2} = 5/2R_{wt}$, $l_{t1} = C_t d_{t1}$, $n_d = 1.5$. Following the reductions introduced previously, n_a is bound from below by a spring yield limit criterion, meaning that $n_a = n_a(d_{ps2}^{+}, z_{pre}^{+}; \sigma_{ps})$, where σ_{ps} is the spring's yield limit. Solving h_8 for z_{acc} and back-substituting these terms into the expression yields:

$$\begin{aligned} z_{acc} = C_t d_{t1} + l_{t2} + l_{b1} - (n_a(d_{ps2}^{+}, z_{pre}^{+}) + 1.5)d_{ps2} \\ - l_{n1} - l_{n2} - 1.5\text{mm} - 7/2R_{wt} \end{aligned} \quad (42)$$

If one were to disregard the effect of constraints, z_{acc} might have seemed to be an independent variable to be used to optimize the impact velocity. However, these constraint activities have resulted in an expression that introduces trade-off variables into c_3 upon back substitution, namely l_{n1} and l_{n2}, and increases the trade-off through d_{t1} and d_{ps2}. This also means that g_{20}, a locally active constraint that prevents the needle from protruding through the valve before injection, contributes to the trade-offs involving all four objectives.

4. **Assembly Features**: The housing snap ($d_{b4} - d_{b3} = 2R_{ov} = 1.2\text{mm}$) and the cylinder on the top housing which seals the valve against the base, are assembly features that result in parasitic contributions that detrimentally affect the Pareto set. These contribute to the trade-offs between self orientation, velocity, and device diameter, given that they affect the relationship between the achievable device size, and the position of the housing split l_{t2} (through g_1). Similarly, the needle attachment (l_{nh1}) affects the set through the axial constraints (g_{20} and h_8). Hence, any design change that eliminates these contributions would improve the Pareto-set.

	d_{t1}	l_{t2}	l_{n1}	d_{n1}	d_{ps1}	d_{ps2}	n_a	d_p	δ_{nh}
f_1	-	-	+	+	+	+	+	+	+
c_1	+	(+)	(+)		(+)	(+)	(+)	(+)	(+)
c_2	(-)		-	-		(+)	(+)		
c_3	-		+	+	+	-	+	+	+

Table 1: The monotonicities of the objectives w.r.t. variables of interest, before εMA. Parentheses indicate a local dependency caused by constraint activity.

4.2 Redesign Mapping, Sequencing, and Exploration

Clearly, several dependencies cause trade-offs and worsen the Pareto set, most notably that the spring fits around the trigger. Any increase in spring force via d_{ps2} results in a larger trigger system to increase the load-bearing area, increasing the device's size. Further, d_{ps1}, an otherwise influential harmonious variable, affects the trade-offs given that reducing the coiling diameter reduces the space available for load-bearing geometry. This worsens the influence of d_{ps2} on size and self-orientation. As the optimal device diameter, $\underline{\tilde{\varepsilon}_1}$, is determined by a radial fit constraint, it also seems inopportune that the spring force is absorbed over an area in the radial direction. As such, a trade-off will always exist between velocity and size unless we find a more space-efficient way of distributing the load while making the d_{ps1} less dependant on the trigger design and vice versa.

The serial vertical arrangement of the internal components results in several trade-off variables; it also causes a trade-off between velocity and API payload, as the needle length l_{n1} cannot be increased at the optimum without reducing the acceleration stroke z_{acc} or the spring length (and thereby d_{ps2}). Further, $n_a(d_{ps2}^{+}, z_{pre}^{+}; \sigma_{ps})$ multiplies the negative influence of d_{ps2} on self orientation, as the spring mass is mounted at the top of the device. Finally, the parasitic contributions introduced by the assembly features in crucial cross-sections detrimentally affects the Pareto set. After mapping out the redesign options to solving these issues, we used the redesign sequence procedure along with the Lagrange multipliers and variable-objective plots to identify the following sequence of changes:

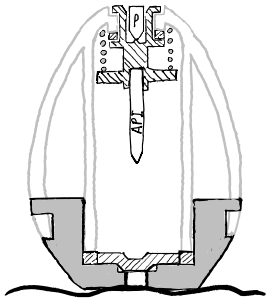
1. *Eliminate parasitic contributions*: Reduce or eliminate the parasitic contributions of the housing snap and needle-hub interface upon the radial and axial fits, and upon the objective functions. Reduce the volume/mass of the plastic components when possible. Explore alternative linear guides for the needle hub and sealing principles for the valve component.
2. *Shifted bounds*: Shift $\underline{d_{ps1}}$, leveraging its harmonious influence, which is to the third power w.r.t. velocity.
3. *Pareto constraint dependency reduction*: In some activity cases, the trigger interface stress becomes a Pareto constraint of the form $g_{11}(\tilde{\varepsilon_1}^{-}, \tilde{\varepsilon_3}^{+}; \sigma_{IF})$. Reduce the geometric dependency between the spring and trigger,

1
FLIPPED TRIGGER

Design changes

Trigger: The ratchet arms have been inverted to work in tension rather than compression.

Assembly: The interface between top and base housings has been changed to a female-male snap feature. The interface between needle and hub has also been changed from a shaft to a hole.



Optimality Strategy

d_{ps1} - The plug diameter, d_p, and two wall thicknesses $2R_{wt}$, have been removed from its GLB *(bound shift*).

w_{nh} - With the trigger arms flipped, the LUB of their width is no longer determined by a mold tool constraint (*constraint relaxation*)

$g_1^*(\tilde{\varepsilon_1}, \overline{d_{ps2}}^+, \overline{l_{T2}}^+, \delta_{nh}^+)$ - The contribution of d_p, $2R_{wt}$, $2R_{ov}$ has been removed (*Shift Pareto constraint*)

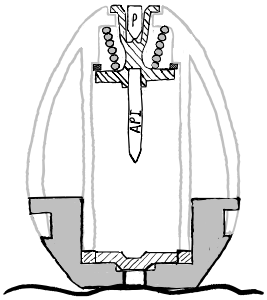
Impact

d_{ps2} and l_{t2} are some of the driving trade-off variables between self-orientation, impact velocity, and size. By relaxing the multiobjective Pareto constraint, and by shifting the bound of their shared harmonious variable, d_{ps1}, the optimum of all three is improved. However, the sealing ring diameter, d_{s1}, is now a part of the GLB of d_{ps1} and ε_1.

2
FLIPPED SEAL

Design changes

Sealing ring: The layout of the top housing, sealing ring, and spring, has been changed. Instead of creating a seal against the top housing inside the spring, the sealing ring fits around the spring, allowing a stiffer conical spring.



Optimality Strategy

d_{ps1} - The sealing ring diameter, d_{s1}, has been removed from the GLB (*bound shift*). The new conical spring coil has the smallest possible coiling diameter beyond placing the trigger arms around the spring, which would cause molding issues.

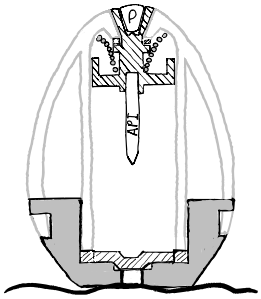
Impact

Combined with the changes made in the first iteration, this iteration has reduced the trade-off between velocity and size, and between velocity and self-orientation. The bounds on two important harmonious variables, d_{ps1} and Z_{acc} have been shifted. In doing so, two wall layers became redundant, meaning two Pareto constraints - g_1 (radial fit) and g_{20}(axial fit) - are shifted by the removal of parametric contributions. The reduced trade-offs still exist however, especially due to the activity of the trigger interface load constraint, which increases the size of the trigger overlap, affecting device size as the spring force is increased.

3
WEDGE TRIGGER

Design changes

Trigger redesign: Building on the changes made in iteration 1, the ratchet based triggering mechanism has been replaced with a wedge design which is also loaded in tension. Hence the spring force is distributed over a larger surface than before.



Optimality Strategy

$g_1^*(\tilde{\varepsilon_1}^-, \overline{d_{ps2}}^+, \overline{l_{T2}}^+)$ - The radial fit constraint is now independent of δ_{nh}, which is critically constrained by the load constraint $g_{42}^*(\overline{d_{ps2}}^+, \delta_{nh}^-)$ at velocity Pareto frontiers. g_{42} is in effect a Pareto constraint, as it depends on $\overline{d_{ps2}}$. Unless another radial fit constraint actively bounds the outer diameter, $\overline{d_{t1}}$, the spring force and the load bearing surface in the trigger can now be increased, without increasing the size of the device beyond the contribution of d_{ps2} (*Pareto constraint dependency reduction)*

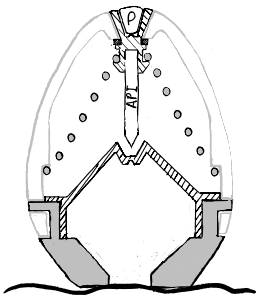
Impact

After the 2nd iteration, the combination of size and velocity comes down to how much stress the loaded plastic components can withstand under long term static loading. Here, this issue has drastically been reduced, as the load is distributed over an additional dimension. In introducing a new degree of freedom affecting the size of the this load bearing area, we can effectively decrease the size of the device further, without compromising the shelf life of the device.

4
FLIPPED ACTUATOR

Design changes

Actuation: The spring has been replaced with a telescopic tension spring to allow the spring coil to pass through itself. As this spring is self-centering, the cylindrical guide can be removed, by using the valve to prevent tilt, ensuring that the needle exits the device.



Optimality Strategy

d_{ps2} - The wire's influence on f_1 is multiplied by n_a and d_{ps1} and its mounting height. Inverting the spring *scales* its' negative influence and made it geometrically independent of the trigger. (*scale trade-off variable* and *bound shift*).

z_{acc} - This inversion also allows the elimination of the contribution of $n_a \cdot d_{ps2}$ from h_8, which determined the achievable acceleration stroke, z_{acc}. Hence, we have leveraged a harmonious variable and shifted g_{20}, which causes a trade-off between velocity and API payload. (*bound shift)*

Impact

This iteration addresses the trade-off between self orientation and velocity. Previously, we have primarily leveraged harmonious variables. Here, we aligned a trade-off variable, by shifting the spring mass downward. We have also radically shifted the upper limit of the acceleration stroke which is mass-less. Yet, this comes at the cost of a new dependency between the amount of spring material and the device diameter.

Fig. 10: Redesign iterations supported by the systematic application of the Principles of Optimality Improvement. Note that these are only principle sketches and do not reflect relative sizing.

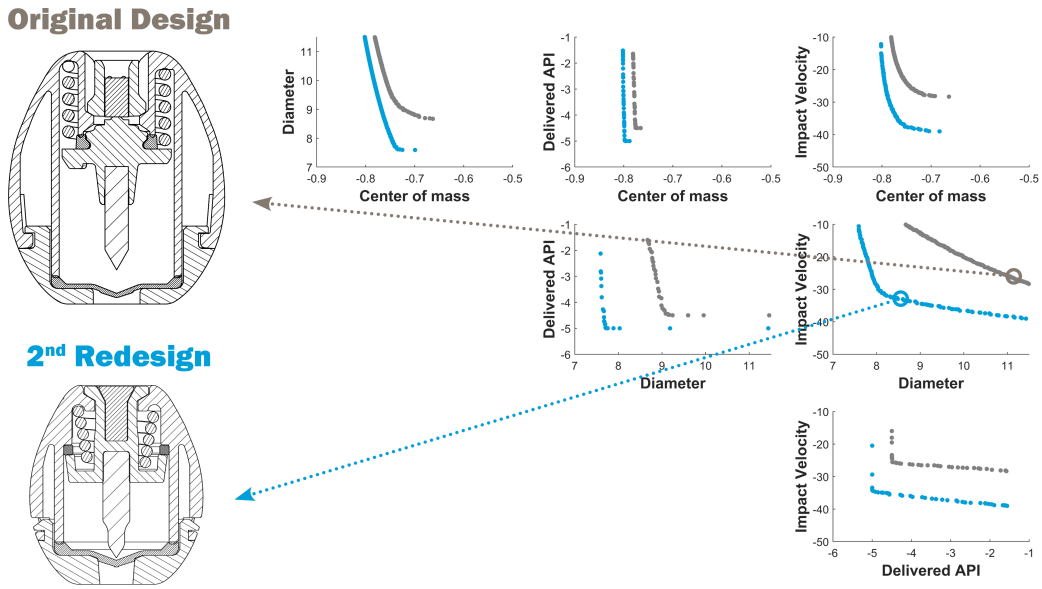


Fig. 11: A head-to-head comparison of the original configuration (grey) against the 2nd redesign (blue), the *Flipped Seal*. The 4D Pareto-set is shown with a 2D projection showing the bi-objective Pareto frontiers between each objective pair, which shows how the redesign is a clear improvement on all accounts. The relative size of the two designs is to scale.

making the radial fit constraint $g_1(\tilde{\varepsilon_1}^-, \overline{d_{ps2}}^+, \underline{l_{t2}}^+)$, and the g_{11} less interdependent. Attempt this by changing the working direction of the trigger interface to add additional degrees of freedom, resulting in a new interface stress criterion (which is currently globally active).

4. *Scale Trade-off Variable*: Reduce the influence of d_{ps2} upon size and self-orientation by to moving the spring closer to the centre of mass - e.g. using a tension spring.
5. *Shifted bounds*: Eliminate the contributors to h_8 that reduce $\overline{z_{acc}}$.
6. *Eliminate Parasitic Contributions*: Reduce the volume/mass of the plastic components when possible. Explore alternative linear guides for the needle hub and sealing principles for the valve component.

The application of this sequence of redesign principles led to a series of redesigns shown and explained in Figure 10. The design changes are relatively simple, and are essentially analogous to well known redesign heuristics, namely, *inversion* [3, 5, 13, 15] and *change in working direction/load path* [3, 5, 13], and *contributor reduction* [3, 12].

In the 1st iteration shown in Fig 9.1, the trigger arms are inverted to work in tension. The plug can now move upward, meaning the spring coil can in part be shaped independently of the plug as the trigger arms are flexible. This allows a slimmer, stiffer spring and an increased $\overline{z_{acc}}$, and eliminates a mold tool constraint which limits the achievable trigger arm width. The housing snap and needle-hub interfaces have also been changed from overlaps to holes. The 2nd iteration shifts the glb of d_{ps1} further by moving the seal outside the spring. A stiffer conical spring can thus be used. Now, $\overline{\delta_{nh}}$ is determined by the outer shape of the device and the length of the trigger arms, as they can flex inside the spring during injection. The 3rd iteration replaces the trigger with a conical wedge-like interface. This change in working direction adds a degree of freedom to the trigger interface design, changing the primary loading direction from radial to axial, allowing the spring to be stiffened without reducing the load-bearing area. In the 4th iteration, the spring is replaced by a tension spring, making the trigger and spring geometrically independent. This also increases $\overline{z_{acc}}$ by eliminating the dependency between spring length and needle length, all the while reducing the impact of the spring's mass upon self-orientation. Interestingly, none of these redesign iterations involve drastic changes, such as additional components or a change in working principle. The use of analysis has guided the identification of simple changes that will substantially impact the optimization model and hence the Pareto set.

4.3 Redesign Evaluation

We built an optimization model to compare a redesign with the original SOMA as a form of validation of improvement according to Definition 3. For brevity, we limited the comparison to a single design selecting the 2nd redesign. It required relatively few modelling changes while still embodying influential changes of the configuration design.

The original optimization model was rebuilt with new constraint functions to reflect the new part fits, updated expressions for mass distribution to reflect the changes in geometry, and changes in spring equations to reflect the conical shape. The model structure, governing equations, and level of detail remained unchanged. This model was run with 200.000 iterations, with $\varepsilon_{\mathbf{L}} = [7\text{mm}; 1.5\text{mg}; 10\text{m/s}]$ and $\varepsilon_{\mathbf{U}} = [11.5\text{mm}; 5\text{mg}; 45\text{m/s}]$. The results in Fig. 11 show the new Pareto set lying beyond the original one. For the union

of the Pareto sets, $\mathcal{C}_{\mathcal{U}} = \mathcal{C}_0 \cup \mathcal{C}_2$ the meta Pareto-set was found to only consist of solutions from the 2nd redesign, i.e., $\breve{\mathcal{C}} = \mathcal{C}_2$, and the single-objective optima of self-orientation has been improved by 2.63%, the size by 12.41 %, API payload by 11.11%, and velocity by 37.68%. We can thus conclude that the redesign is, in fact, a design improvement, as it meets the criteria in Definition 3. For the subsequent redesigns, it is likely that the achievable combination of impact velocity and self-orientation is improved even further, as the design changes are aimed at increasing the load-bearing area in the trigger system and shifting the centre of mass downward while increasing the acceleration stroke.

5 Discussion

With the methodology presented here, a degree of rigour is brought into the iterative design process, allowing the designer to utilize optimization to qualify the introduction of design changes. Thus, it is not singularly a configuration design or design optimization methodology - it is both. As seen with the SOMA device, the actual changes required to achieve an improved Pareto set can be relatively simple. Inversion, change of working direction, and changes to how the components fit together. Still, the impact on the Pareto set is substantial, as seen in the dominance of the *Flipped Seal* redesign over the original design.

One could argue that the presented methodology is a formalization of how experienced designers work, with their decisions largely being based on knowledge of trade-offs [13, 24] and active constraints [2, 26]. They often exhibit a degree of opportunism in identifying and solving issues by obviating dependencies that negatively affect performance and feasibility. This may allow them to synthesize designs that are easier to optimize, merging compatible functionality in certain subsystems/parts while separating functionality that is not. The proposed method shows its benefits by forcing designers to identify, understand, and mitigate the weaknesses of their configuration designs, potentially breaking fixation in the process. This might be especially useful for design tasks not met previously.

The methodology has its limitations. The most obvious one is that the analysis involved would seem onerous to most designers. Here a cost-benefit mindset comes in: if the benefit gained through redesign is accrued over a production volume counted in millions or billions (as is the potential with SOMA), then the cost of analysis becomes almost trivial. Another limitation is that the methodology's success depends entirely on whether all objectives and constraints of importance have been taken into account in the model. Therefore, the importance of a restrained approach to applying the redesign principles cannot be understated. If there is some tacit constraint or objective involved or one which the model simply does not consider, we must keep those in mind when introducing design changes.

The redesign procedure is opportunistic, as it is based on monotonicity analysis. Hence, it may not always be applicable. For designs with non-monotonic objective and constraint functions, the effort involved in understanding the changes in monotonicity across the design space can be prohibitively time consuming or inconclusive. As noted already, cases such as SOMA may warrant the effort. Occasionally, one might perform the model reductions in MOMA and εMA using numerical data (i.e., post-optimality) instead of formal monotonicity analysis, as discussed in [28]. One might also handle more complex problems by applying the redesign principles on a different level of abstraction, e.g., at the functional architectural level, looking for redesign opportunities that redistribute functionality across the subsystems and parts. Alternatively, one might also use the methodology to explore a part of the system, e.g., to understand a trade-off issue between a pair of essential objectives.

When introducing design changes, situations might arise where new trade-off variables are introduced, such as in the flipped actuator redesign of SOMA. Here, the new spring design introduces new trade-off variables between impact velocity and device size. Certain changes might also be incompatible, resulting in trade-offs between design changes. In such situations, it would be necessary to quantify the influence of these changes.

Finally, when introducing configuration design changes iteratively, there will likely be a degree of path dependency. Despite it relying on the outputs of analysis and optimization, our sequencing approach is heuristic, and the final redesigns will depend on the early iterations. This can be overcome in part by exploring solutions before implementing them, but this still does not guarantee compatibility. A pragmatic approach of letting the relative importance of the objectives affect the sequence can be worthwhile. For example, if there were no swallowable designs in the Pareto set of the original SOMA, then it would have made little sense to start mitigating trade-offs between self-orientation and impact velocity.

6 Conclusion

The question of systematic configuration design is a challenge for design and optimization research. In practice, it is mainly driven by the skill and experience of the designer rather than the application of a clear design theory. In this contribution, we have expanded upon previous work on multiobjective monotonicity analysis to demonstrate its application in configuration redesign. The result is a rigorously founded methodology that enables the designer to identify design changes to improve on all objectives and reduce or eliminate trade-offs represented in the Pareto set. When applied systematically, the result is performance beyond what is achievable through proportional or parametric optimization alone. We demonstrated this capability the SOMA device. A new optimization model comparing the 2nd redesign iteration with the original design was built to show the method's validity. This revealed a substantial size reduction and increase in impact velocity without worsening self-orientation or API payload. Such analysis and redesign methodology empowers designers to explore better configuration designs systematically.

Acknowledgements

The authors would like to thank the support of the Danish Innovations Fund and the Novo Nordisk STAR-programme (grant no. 7038-00221B), and the University of Michigan Donald C. Graham Endowment. We are grateful to Novo Nordisk for sharing design information and data, to Chris McMahon of the University of Bristol, for his advice and input, and Giovanni Traverso and his colleagues at MIT for their helpful comments and input. The opinions presented here are solely those of the authors.

References

[1] Design Council. *Eleven lessons: managing design in eleven global companies*. Tech. rep. 272099. 2007.

[2] Mcmahon, C. A. "Observations on Modes of Incremental Change in Design". In: 5.3 (1994). DOI: 10.1080/09544829408907883.

[3] Pahl, G. and Beitz, W. *Engineering design — A systematic approach*. 1999. DOI: 10.1016/0261-3069(96)84970-3.

[4] Ullman, D. G., Dietterich, T. G., and Stauffer, L. A. "A model of the mechanical design process based on empirical data". In: *Artificial Intelligence for Engineering, Design, Analysis and Manufacturing* 2.1 (1988), pp. 33–52. DOI: 10.1017/S0890060400000536.

[5] French, M. J. *Conceptual Design for Engineers*. 1985. DOI: 10.1007/978-3-662-11364-6.

[6] Bendsøe, M. P. and Kikuchi, N. "Generating optimal topologies in structural design using a homogenization method". In: *Computer Methods in Applied Mechanics and Engineering* (1988). DOI: 10.1016/0045-7825(88)90086-2.

[7] Schmidt, L. C. and Cagan, J. "Optimal Configuration Design: An Integrated Approach Using Grammars". In: *Journal of Mechanical Design* 120.1 (Mar. 1998), pp. 2–9. DOI: 10.1115/1.2826672.

[8] Bayrak, A. E., Kang, N., and Papalambros, P. Y. "Decomposition-Based Design Optimization of Hybrid Electric Powertrain Architectures: Simultaneous Configuration and Sizing Design". In: *Journal of Mechanical Design, Transactions of the ASME* 138.7 (2016), pp. 1–9. DOI: 10.1115/1.4033655.

[9] Bayrak, A. E., Ren, Y., and Papalambros, P. Y. "Topology Generation for Hybrid Electric Vehicle Architecture Design". In: *Journal of Mechanical Design, Transactions of the ASME* 138.8 (2016). DOI: 10.1115/1.4033656.

[10] Antonsson, E. K. and Cagan, J., eds. *Formal Engineering Design Synthesis*. Cambridge University Press, Nov. 2001. DOI: 10.1017/CBO9780511529627.

[11] Chakrabarti, A. et al. "Computer-based design synthesis research: An overview". In: *Journal of Computing and Information Science in Engineering* 11.2 (2011). DOI: 10.1115/1.3593409.

[12] Suh, N. P. "Axiomatic Design Theory for Systems". In: *Research in Engineering Design - Theory, Applications, and Concurrent Engineering* 10.4 (1998), pp. 189–209. DOI: 10.1007/s001639870001.

[13] Altshüller, G. *Creativity As an Exact Science*. Gordon and Breach, 1984. DOI: 10.1201/9781466593442.

[14] Matthiassen, B. "Design for Robustness and Reliability - Improving the Quality Conciousness in Engineering Design". PhD thesis. Technical University of Denmark, 1997.

[15] Tjalve, E. "Form Design - a Systematic Approach." In: *Schriftenreihe WDK (Workshop Design - Konstruktion)*. 1981, pp. 559–571.

[16] Boothroyd, G., Dewhurst, P., and Knight, W. A. *Product Design for Manufacture and Assembly*. 2010. DOI: 10.1201/9781420089288.

[17] Arthur, W. B. "Why Do Things Become More Complex?" In: *Scientific American* 268.5 (1993), pp. 144–144. DOI: 10.1038/scientificamerican0593-144.

[18] Wynn, D. C. and Eckert, C. M. *Perspectives on iteration in design and development*. Vol. 28. 2. Springer London, 2017, pp. 153–184. DOI: 10.1007/s00163-016-0226-3.

[19] Papalambros, P. Y. and Wilde, D. J. *Principles of Optimal Design*. Cambridge University Press, Jan. 2017. DOI: 10.1017/9781316451038.

[20] Sobek, D. K., Ward, A. C., and Liker, J. K. "Toyota ' s Principles of Set-Based Concurrent Engineering Toyota ' s Principles of Set-Based Concurrent Engineering". In: *Sloan Management Review* 40.2 (1999), pp. 67–83.

[21] Cagan, J. and Agogino, A. M. *Innovative design of mechanical structures from first principles*. 1987. DOI: 10.1017/S0890060400000275.

[22] Papalambros, P. and Wilde, D. J. "Global Non-iterative Design Optimization Using Monotonicity Analysis". In: *Journal of Mechanical Design, Transactions of the ASME* 78 -WA/DE-17 (1978).

[23] Jain, P. and Agogino, A. M. "Theory of design: An optimization perspective". In: *Mechanism and Machine Theory* 25.3 (1990), pp. 287–303. DOI: 10.1016/0094-114X(90)90030-N.

[24] Ahmed, S., Wallace, K. M., and Blessing, L. T. "Understanding the differences between how novice and experienced designers approach design tasks". In: *Research in Engineering Design* 14.1 (2003), pp. 1–11. DOI: 10.1007/s00163-002-0023-z.

[25] Cross, N. "Expertise in design: An overview". In: *Design Studies* 25.5 (2004), pp. 427–441. DOI: 10.1016/j.destud.2004.06.002.

[26] Eckert, C. M. and Stacey, M. K. "Constraints and Conditions: Drivers for Design Processes". In: *An Anthology of Theories and Models of Design: Philosophy, Approaches and Empirical Explorations*. Ed. by Chakrabarti, A. and Blessing, L. T. M. London: Springer London, 2014, pp. 395–415. DOI: 10.1007/978-1-4471-6338-1{_}19.

[27] Onarheim, B. “Creativity from constraints in engineering design: Lessons learned at Coloplast”. In: *Journal of Engineering Design* 23.4 (2012), pp. 323–336. DOI: 10.1080/09544828.2011.631904.

[28] Sigurdarson, N. S., Eifler, T., Ebro, M., and Papalambros, P. Y. “Multiobjective Monotonicity Analysis: Pareto Set Dependency and Tradeoffs Causality in Configuration Design”. In: *Journal of Mechanical Design* (2021), pp. 1–18. DOI: 10.1115/1.4052444.

[29] Abramson, A. et al. “An ingestible self-orienting system for oral delivery of macromolecules”. In: *Science* 363.6427 (2019). DOI: 10.1126/science.aau2277.

[30] Marler, R. T. and Arora, J. S. “Survey of multi-objective optimization methods for engineering”. In: *Structural and Multidisciplinary Optimization* 26.6 (2004), pp. 369–395. DOI: 10.1007/s00158-003-0368-6.

[31] Carmichael, D. “Computation of Pareto Optima in Structural Design”. In: *International Journal for Numerical Methods in Engineering* 15 (1980), pp. 925–952. DOI: 10.1017/S0022029900029393.

[32] Athan, T. W. and Papalambros, P. Y. “A quasi-Monte Carlo method for multicriteria design optimization”. In: *Engineering Optimization* 27.3 (1996), pp. 177–198. DOI: 10.1080/03052159608941405.

[33] Jugulum, R. and Frey, D. D. “Toward a taxonomy of concept designs for improved robustness”. In: *Journal of Engineering Design* 18.2 (2007), pp. 139–156. DOI: 10.1080/09544820600731496.

[34] Goto, T., Isobe, R., Yamakawa, M., and Nishida, M. “The New Mazda Gasoline Engine Skyactiv-G”. In: *MTZ worldwide eMagazine* 72.6 (2011), pp. 40–47. DOI: 10.1365/s38313-011-0063-8.

[35] Ebro, M. and Howard, T. J. “Robust design principles for reducing variation in functional performance”. In: *Journal of Engineering Design* 27.1-3 (2016), pp. 75–92. DOI: 10.1080/09544828.2015.1103844.

[36] U.S. Department of Health and Human Services Food and Drug Administration (CDER). “Guidance for Industry: Size, Shape and Other Physical Attributes of Generic Tablets and Capsules”. In: *Pharmaceutical Quality/CMC* December (2013), pp. 1–11.

Appendix 4: Paper C (Supplementary)

Title: Functional Trade-offs in the Mechanical Design of Integrated Products - Impact on Robustness and Optimisability
Authors: Sigurdarson, N.S.; Eifler, T.; Ebro, M.
Publication: The Proceedings of the 22nd International Conference on Engineering Design (ICED 19) held in Delft, Netherlands.

FUNCTIONAL TRADE-OFFS IN THE MECHANICAL DESIGN OF INTEGRATED PRODUCTS - IMPACT ON ROBUSTNESS AND OPTIMISABILITY

Sigurdarson, Nökkvi S. (1,2); Eifler, Tobias (1); Ebro, Martin (2)

1: Technical University of Denmark (DTU); 2: Novo Nordisk A/S

ABSTRACT

It is generally accepted in industry and academia that trade-offs between functional design objectives are an inevitable factor in the development of mechanical systems. These trade-offs can have a large influence on the achievable robustness and performance of the final design, with many products only functioning in narrow sweet-spots between different objectives. As a result, the design process of multi-functional products can be prolonged when designers concurrently attempt to find sweet-spots between a number of potentially interdependent trade-offs.This paper will show that designers only have six different approaches available when attempting to manage a trade-off while trying to ensure robustness and a sufficient performance. These fall within one of three categories; accept, optimise, or redesign. Selecting the wrong approach, can result in consequences downstream which can be difficult to predict, amongst others a lack of robustness to geometric variation, constrained performance, and long development lead time.This points to a substantial potential in the synthesis of design methods that support the identification and management of trade-offs in early product development.

Keywords: Robust design, Embodiment design, Optimisation, Design trade-offs

Contact:
Sigurdarson, Nökkvi S,
Danish Technical University (DTU)
Department of Mechanical Engineering
Denmark
noksig@mek.dtu.dk

Cite this article: Sigurdarson, N.S., Eifler, T., Ebro, M. (2019) 'Functional Trade-offs in the Mechanical Design of Integrated Products - Impact on Robustness and Optimisability', in *Proceedings of the 22nd International Conference on Engineering Design (ICED19),* Delft, The Netherlands, 5-8 August 2019. DOI:10.1017/dsi.2019.356

1 INTRODUCTION

In mechanical engineering design it is generally accepted that trade-offs between design-objectives will inevitably need to be made in the development process. Most systems are integrated, meaning that their individual components and subsystems each contribute to several functions and requirements, and often do so simultaneously. This tendency, which seems to be growing, is driven by two factors:

1. Competitive pressure - the success of new products is generally reliant on increased functionality and/or improved performance, and product developers will therefore commonly strive to add new features and sub-functions that bring added value to the user or optimise the existing design.
2. Associated cost - production companies aim for driving as much functionality with as few components as possible, in order to ensure cost efficient and reliable products.

The result is products that are becoming increasingly complex to develop (Arthur, 1993), given a growing amount of constraints and interdependent design objectives. No product can be infinitely accurate, durable, efficient, robust, user friendly, manufacturable, cost effective, etc. In order to find the correct balance between the many objectives a product is designed toward, numerous trade-offs need to be either solved or managed systematically throughout the design process. As a consequence, product development can be a highly iterative process, where the design is gradually refined until the required functionality has been achieved and an acceptable compromise between all design objectives is reached. The aim of this contribution is to show that there is a lack of a comprehensive way for a design engineer, faced with a design trade-off, to decide between *accepting* the trade-off, *optimising* the design, or making a *design change* to avoid the trade-off. As a consequence, realising the design may require tighter tolerances and lesser performance than originally foreseen in the preliminary phases, with a design process that is correspondingly more prone to loopbacks that lead to delays.

2 THEORETICAL BACKGROUND

The development of new products often follows a structured process (Pahl & Beitz, 1996) with a defined set of phases, each becoming more specific and detailed until the product is complete. Decisions made early in this process are decisive for the downstream workload. The process of geometrically realising the functional intent and layout of the final product - often coined embodiment design - determines the limits of its achievable performance (Papalambros & Wilde, 2017), and robustness (Andersson, 1997). Yet as discussed by Andreasen & Howard (2012), this process is ill supported by existing methodology. As such, it seems that most late design changes and development lead-time can be traced back to decisions made during embodiment (Vianello *et al*., 2012).

Most optimisation and robust design methods are highly dependent on knowledge that is not necessarily available at an early stage of development and are therefore often applied at least after a preliminary embodiment has been developed (Papalambros & Wilde, 2017), (Ebro & Howard, 2016). As a result, a lot of existing design methodology aimed at the embodiment phase is based on heuristics e.g. (French, 1971), (Pahl & Beitz, 1996), (Matthiassen, 1996), (Anderson, 1997), and DfX (Olesen, 1992). Yet the challenge in most of these is that their application is mostly limited to single functions, single domains e.g. structural design, hydraulics, and single dispositions, (c.f. DfX). In the context of the embodiment of mechanical systems, there are however several domains (especially in multi-physical design situations) and dispositions to take into account, and as Mattiassen (1997) remarked, no universal design principle can meet all design requirements. This is then further complicated when the designer has to realise multiple functions and sub-functions in the same system.

Few specific embodiment heuristics exist within the question of functional integration, perhaps due to the infinite amount of combinations of working principles and design requirements that could feasibly exist in the same system. Yet the importance of allocation of functionality amongst the components in a system is covered by numerous sources from different perspectives, e.g. division of tasks (Pahl and Beitz, 1996), integration and differentiation (Matthiassen, 1997), merging and segmentation (Altshüller, 1984). However, none of these answer the question of how or when to integrate functionality into fewer components, and when to differentiate. Yet from a production perspective, integration can have significant benefits; cf. design for manufacture and assembly (Boothroyd, 2002).

Functional allocation in a system can have a significant detrimental impact, if the wrong functions are integrated into the same subsystems. Integration implicitly involves designing geometrical features or system properties that contribute to multiple design objectives - either simultaneously or in different

functional states. Increasing integration in other words creates dependencies. A natural consequence of integration in design, dependencies can both have a positive and negative impact on a design:

- Positive in the sense that they allow more functionality to be fulfilled without increasing the amount of sub-systems, parts, or even geometrical features.
- Negative in situations where the dependencies leads to a design trade-off, where two or more objectives have conflicting relationships to same geometry

Herein lies the reason as to why no universal design principle exists to meet all design requirements; as argued Pahl and Beitz (1996), it can be difficult or impossible to optimise the "carrier of several combined functions". Implicitly, multifunctional products will always require trade-offs to be made. Given that design principles are generally made with few objectives in mind, following these otherwise useful recommendations will inevitably come with a cost in form of a trade-off with other objectives within the system. Take the commonly cited "*Provide short direct force paths*" principle (French, 1971/Pahl & Beitz, 1996). While it definitely has benefits from a structural design perspective, it directly contradicts Matthiassens' (1997) principle *integrate for coordinated outputs* which aims at ensuring accuracy and coordination between subsystems in machine by using a shared power input, which of course results in longer force paths than otherwise. As a result, designers are left to rely on experience and intuition to create good designs, as shown by Ahmed *et al*. (2003), who found that experienced designers are much more likely to be aware of trade-offs.

Several existing frameworks address dependencies between requirements in design. All involve some form of dependency identification and assessment, with some aimed at supporting synthesis that aims at reducing the negative impacts of dependency. Dependency modelling methods are widely applied in design. An example of such, is the design structure matrix (DSM), which is used to qualitatively identify dependencies in complex system development, addressing aspects such as modularity, task coordination, system optimisation, and system integration (Eppinger & Browning, 2012).

More quantitative approaches to trade-offs are however widespread in the context of optimization, decision support and design space exploration. Examples include hybrid trade-off strategies (Otto *et al*., 1991), the compromise decision support (Mistree *et al*., 1993), and trade-space exploration (Ross *et al*., 2004). These however all aim at finding the best solution in a system influenced by trade-offs, rather than changing the system itself. The field of optimal design is aimed at identifying the best combination of parameter values given a set of objectives, constraints, and dependencies. Dependency mapping - for instance functional dependence trees (Wagner, 1993) - is used for model partitioning, and decomposition in optimisation, in an effort to simplify model definition and reduce computational cost. Optimisation often involves finding the best trade-off between two or more objectives within a set of constraints. Yet at the same time, the optimisation field does not necessarily question whether the design itself is worth optimising - whether it is sufficiently optimisable, or whether the objectives involved are so conflicting and hence limiting, that alternative embodiments might be worth exploring.

Methods related to design dependency aimed at the support of synthesis also exist. An example of such, is Axiomatic Design (AD) (Suh, 2001), which states that any form of dependency - termed coupling - is detrimental to how well a system will function, and should therefore be avoided or reduced. Another example is TRIZ (Altshuller, 1984), a knowledge-based inventive problem solving framework, which aims at supporting invention through identification and removal of so-called contradictions, which are incompatibilities between design objectives.

Interestingly, Ahmed *et al*. (2003), found that strategies to identifying or avoiding design trade-offs is largely based on tacit knowledge in industrial practice. This surely creates additional challenges in original design, given that experience is less useful when designers are met with design tasks and issues they have not faced before. When designing a multi-functional product, how does a designer then ensure that the decisions made during the embodiment phase, do not result in trade-off situations at a late stage of development, where the only recourse is constrained optimisation or acceptance? If the product could be embodied in such a way, that the aforementioned situation is avoided, then the development lead time for new complex products, and the importance of experience would be reduced. As discussed, heuristics are not necessarily sufficient in helping designers allocate functionality across a system, in a way that secures robustness and high performance. Meanwhile, the applicability of methods related to dependency can be limited in mechanical design given that:

1. Dependency modelling methods such as DSM, are qualitative and knowledge intensive in nature. Requiring a substantial insight into the workings of a design, DSM would also result in a substantial workload if one were to aim at identifying all the dependencies in a system, across

all levels of abstraction. While these methods are certainly valuable in the decomposition of complex systems, and the identification of dependencies, they do not necessarily aid the designer in then determining the impact of the dependency, and what to do about it.

2. Design optimisation methods and quantitative trade-off studies are mostly applied after a concept and embodiment has been defined, generally only adjusting the parameter design within the limits of the embodiment itself. Driving design and redesign activities with these is as such relatively time consuming, as it implies an iterative, knowledge intensive approach.
3. According to the first axiom of AD, the ideal design would be fully differentiated, meaning that each functional requirement is met by a unique set of design parameters. This implies little to no integration, and would therefore in practice often require a larger number of components to fulfil. This in turn increases the information content, in conflict with the second axiom. Similarly, Frey *et al*. (2007) found that increased coupling is not necessarily detrimental, when studying the relationship between part count and performance. What AD as such fails to capture, is that couplings can in fact have a positive effect on performance and robustness.
4. Being aimed at invention rather than mechanical design specifically, TRIZ is of a general nature, spanning any type of contradiction in any product. It is also fairly limited in its uptake in practice, often attributed to its' perceived enigmatic nature and complexity (Ilevbare *et al*., 2013). Frey *et al.* (2007) also did not find the *law of ideality* to be consistent with their observations from practice.

Given some of these challenges, Göhler and Howard (2015) put forward a metric, the contradiction index, introducing the notion that not all couplings need to have a negative impact. Their work combines AD, the notion of contradictions with system complexity considerations from DSM. The metric was meant as an indicator of robustness to be applied at an early stage of development- Göhler *et al*. (2016) later found a significant correlation between this index and system robustness. This begs the question; how can designers avoid distributing functionality in a way that results trade-offs?

3 TRADE-OFFS IN ROBUST MECHANICAL DESIGN

Achieving a design where all objectives are independent is unrealistic in practice. Design objectives will by interdependent be it due to design integration caused by e.g. manufacturing constraints, or simply due to inherent dependencies between physical phenomena. With this in mind, it seems that the understanding of how trade-offs between functional objectives arise in mechanical design, and how these can be managed, is essential to the embodiment of robust multi-functional products. In the following, a theoretical perspective is given on the reason behind the occurrence of robustness and performance reducing trade-offs, and what options are available to designers in such situations.

3.1 Occurrence and implications of trade-offs in mechanical design

As introduced by Göhler & Howard (2015), functional requirements from AD can be split in three categories; *min-is-best, nominal-is-best, and max-is-best*. This notion is also consistent with optimal design, where objective functions are aimed at minimisation, maximisation, while meeting a set of constraints. Conversely, when looking at the relationship between two coupled functional requirements, they concluded that there are three types of coupling, positive, negative, and nominal. Negative couplings always cause a trade-off between the two requirements, which will often result in a narrow design space, which in turn reduces both the achievable performance and the allowable variation, if both targets are to be met, as illustrated in figure 1. Negative couplings in other words reduce the feasible domain of a design, with the bound" stemming from the coupling rather than a constraint.

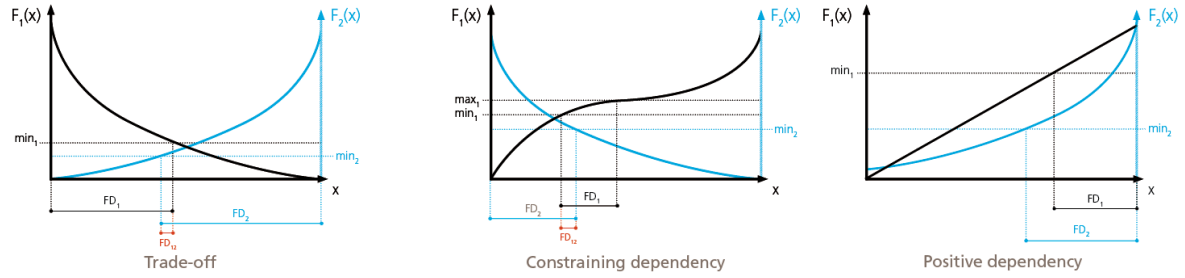


Figure 1 - The types of dependency between two objective functions F1 and F2, with a shared design parameter, x. Adapted from Göhler & Howard (2015)

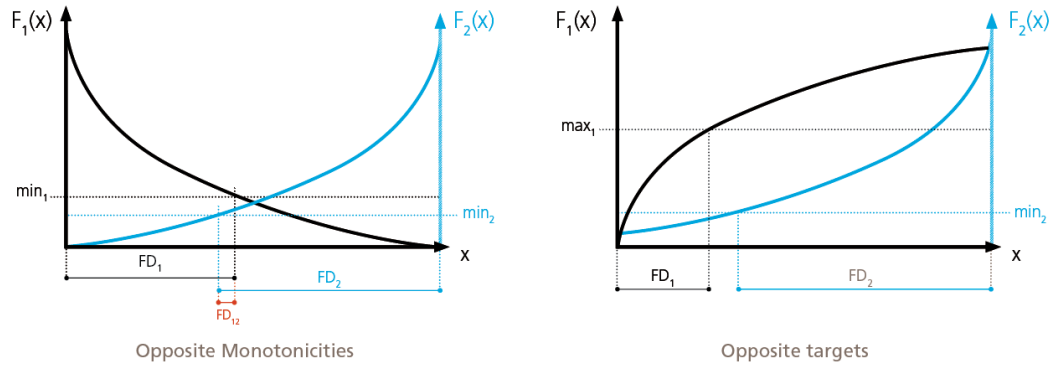


Figure 2 - The two types of functional trade-offs, Left: A narrow feasibility domain (FD12) requiring the design parameter x to have tight tolerance in order to stay inside a "sweet-spot". Right: An unfeasible design where no value of x exists where both objectives are met

Building on these notions, design trade-offs between two objective-functions must only be able arise in two situations. For two mutually dependant objectives F_1 & F_2 shown in fig 2, a trade-off occurs if:

1. F1 & F2 are of the same objective type (e.g. min-is-best vs. min is best), but are oppositely monotonic, either globally or locally within the feasibility domains of each objective.
2. F1 & F2 are of different objective type (e.g. max vs. min) but have the same monotonicities

What is more; these situations can occur in decoupled designs, if the order of the influence of the dependant parameter is sufficiently high, or if the independent parameters cannot be adjusted without violating other constraints. A wide range of analysis tools could be used to identify design trade-offs at a relatively early stage of design, e.g. through the use of monotonicity tables (Papalambros & Wilde, 1979) or using the DSM-based contradiction-index approach (Göhler & Howard, 2015).

The implication for mechanical design, is that systems that constrain themselves due to conflicting objective functions, must have a smaller feasible domain, than designs that are merely bound by geometry constraints. In other words, better optima must be achievable in a design without a trade-off, than in a design with one. Designs with positive dependencies also allow optima that can be achieved with lessened need for to tight parameter control (as the feasibility domain is wider), meaning that the design is more robust to variation. This is not to imply that multi-objective performance trade-offs are the same as the oft discussed and omnipresent performance vs. quality trade-offs, but rather that designs with performance trade-offs are more prone to being sensitive to variation, and will as such more commonly be at risk of variation-driven failure and loss of quality, i.e. that they are optimal but not robust. In summation, integrated mechanical systems can hence be made more optimal and robust if these negative dependencies are avoided, or their number at least reduced to the bare minimum.

3.2 Trade-off management strategies

If a trade-off situation affects a design, what can one then do? Looking at perspectives from prior research and industrial practice, a designer can generally speaking either accept the existence and influence of a trade-off, attempt to tweak/optimise the design, or change the design itself. Common for all three is that it can be difficult to predict the outcome of deciding on one approach rather than another. *Concept A or B? Optimise or redesign*? Accept compromise on *objective 1* or *objective 2*? In the authors' experience, these are all decisions designers can struggle with, and the right approach depends on a wide range of factors. Yet, it would seem that the decision on what approach to apply is not necessarily made explicitly industrial practice, but rather done through tacit and explorative means. As such the management of one or several trade-off scenarios in a design will in some cases be handled with simultaneous or parallel activities - e.g. with a team of design engineers implementing design changes while also performing parametric updates and investigating the impact of compromise. In the following, six generic strategies (c.f. figure 3) available to designers for the management of trade-offs between functional objectives are discussed, along with their benefits and limitations, and the methodological support available in applying these strategies. They are not necessarily independent, and one could as such argue that combinations of otherwise these otherwise different strategies exist.

A trade-off situation between two max-is-best objectives with opposite monotonicities is used as an illustrative example, but the strategies are equally applicable to opposite targets and to trade-offs

involving other types of objective functions. Real-world examples are used to clarify the nature of these strategies. While the examples are somewhat simplistic, they still serve to illustrate the value in- and the need for methods that support designers in selecting the best trade-off management strategy for a given context, and also point to the potential pitfalls involved.

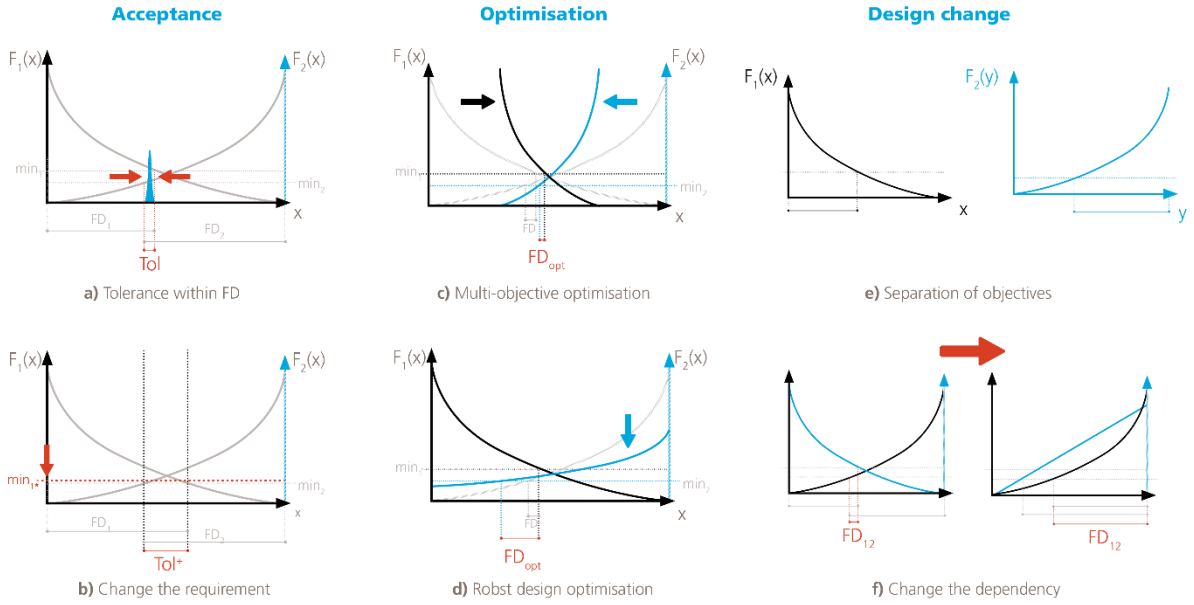


Figure 3 - Strategies to manage trade-off situations. a) Process control, b) Compromise, c) Improve performance d) Improve robustness e) Uncouple f) Change the dependency type

3.2.1 Trade-off acceptance

In accepting a trade-off, the designer does not alter the design in any way, rather relying on that the design can stay within specification in any conceivable state. In order to do so, it is necessary to ensure that the system always stays within the feasibility domain where both targets are met; in other words, the design parameter, x, cannot have a variance larger than the feasibility domain. This implies that the trade-off is managed through **tight tolerances/process control** (figure 3.a), the consequence being a potential increase in cost and a low robustness to degradation throughout the lifecycle. Entire books exist on the subject, with fields such as manufacturing engineering and process optimisation aimed at predicting and improving achievable tolerances and process quality. If it is impossible or costly to control the parameter sufficiently, an alternative would be to accept a **compromise by changing one or more of the requirements** (fig 3b). In doing so, the feasibility domain and therefore the allowable Parameter variance is widened. This implies a reduction in performance of the compromised objectives, and therefore potentially quality loss to the user. Methodologies such as compromise decision support problem (Mistree *et al.*, 1993) aim to support decision making in these situations, to help find the compromise that is most effective in increasing the overall performance.

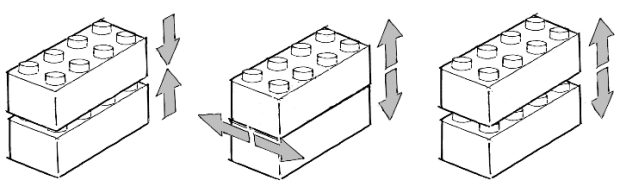


Figure 4 - LEGO should be easy to assembly and disassemble, yet stable once assembled

LEGO bricks are a good example where both approaches have been applied in design and production, to create a unique product offering. On one hand, LEGOs have to be easy to assemble and take apart, yet at the same time be highly stable once assembled. As such regular LEGO bricks are highly contradicting in that core objectives, assembly-, holding-, and disassembly force, have opposite objectives in relation to four controlling design parameters; pin diameter, thickness, interference fit, and material stiffness. Applying TRIZ would reveal that it is a physical contradiction, and solving it

involves separation in time, space, condition, or scale (Altshuller, 1984). These would require a solution that is vastly different from LEGOS, e.g. introducing a third joining component, some form of bi-stability, or a two-directional disengagement movement (e.g. pull and twist). Given that there is little added value to the user in achieving a lower assembly force and a higher stability, LEGO instead found a sweet-spot between all requirements, and managed it through tight process control and interface standardisation. This allows very simple use - single axis movement of two parts - yet in a more multi-functional product this would not necessarily be possible. While it is difficult to actually assert whether the inventor of LEGO consciously decided upon acceptance of the trade-off, it is still clear that the trade-off exists between two essential functional objectives exists. Given that the design still functions, variation in manufacture is surely being tightly controlled to ensure consistent quality.

3.2.2 Trade-off optimisation

Assuming that independent parameters exist that allow the objective functions to be adjusted individually, optimisation can be applied to reduce severity of the trade-off - either through systematic and formal optimisation techniques or through an iterative experience-driven approach where the designers tweaks and refines parameter values in the design to lessen the impact of the trade-off. Be it through formal or tacit means, applying **Multi-objective optimisation** (fig 3.c) and design optimisation techniques would ultimately change the gradients of the two objective functions locally within the limits created by the bounds and the trade-off itself. The result is an improved optimum of one or both of the objectives but a narrower feasibility domain, corresponding to a reduction in robustness. Alternatively, **Robust design optimisation** (fig 3.d) could widen the feasibility domain, but reduce the achievable optimum of one of the objectives. Both approaches are thoroughly described in a wide range of sources, yet in both cases, optimisation only changes the parametric design; the conflict still exits, and will limit the achievable optimum. The question when considering whether to optimise or redesign a system under trade-off, is whether the achievable optimum is sufficient, which is difficult to answer without performing the optimisation itself.

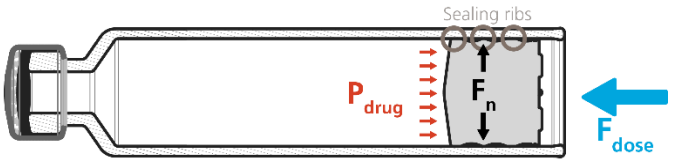


Figure 5 - Drug cartridges for medical injection devices are highly optimised designs, with a trade-off between avoiding leakage and achieving a low dosing force

Drug cartridges for medical devices (fig 5) are a simple example of a design where a trade-off has been mitigated through robustness- and multi-objective optimization. A dose is delivered by pushing a viscoelastic plunger forward, expelling the drug through a needle - the less resistance there is to this movement the faster the drug can be expelled. The plunger also acts as a hydraulic seal, preventing the drug from leaking out during dosing, with the plunger being pretensioned against the cartridge. Amongst the many design objectives in cartridges, are therefore a max-is-best drug sealing objective and a min-is-best dose force objective. A potential conflict arises in that the pretension of the plunger results in a normal force causing friction between the plunger and cartridge walls, which negatively influences the dosing force. This is handled through the minimisation of the coefficient of friction - an independent parameter that only influences the dose force - using a lubricant. Furthermore, the interface between plunger and cartridge walls has been reduced to three ribs to reduce the variation in sealing pressure that would otherwise occur in a continuous interface influenced by shape variation. Finally, the shape of the plunger is optimised towards achieving as homogenous a load distribution as possible while dosing, permitting an equal distribution of sealing pressure. Some of these features have been achieved through formal optimization, while others have been achieved through tweaking and experimentation (i.e. DoE), but it is some form of quasi-optimization nonetheless.

3.2.3 Trade-off mitigation through redesign

If a trade-off is to be avoided entirely, the design will inevitably need to be changed. This is the basic rationale behind theoretical frameworks such as Axiomatic Design and TRIZ. As discussed in section 2, the most common approach to trade-off avoidance or removal is to strive for designs that are

independent, i.e. uncoupled. In this context, the obvious approach to managing a trade-off in a design would be to change the design in a way that allows the **Separation of design objectives** (fig 3.e). By designing towards each objective being independent, there is no direct risk of trade-offs. However, this approach is not without limitations; sometimes objectives are inherently interdependent and a useful uncoupled solution therefore difficult to create. Furthermore, independence can require more components and sub-systems, to ensure that no elements of the design are shared between objectives. From a mechanical design perspective, this could - but does not necessarily - imply: 1) lowered mechanical efficiency and reliability, due to more contributors to losses and more features that can fail and 2) increased cost driven by extra components. Robustness wise, more parameters means longer tolerance chains, more load paths, and potentially more variation. A degree of integration (and ergo dependency) can as such be beneficial. With this in mind, aiming to **Change the dependency** (fig 3.f) between two objectives by design, to a positive dependency instead - thereby removing the trade-off - would allow simultaneous of optimisation of both objectives, and result in no reduction in the size of the feasibility domain of the system. g

An oft cited **example of separation** is the difference between the Newcomen steam engine and the Watt steam engine (Suh, 2001). The Newcomen engine (figure 6) is challenged in that the efficiency of the expansion and condensation cycles cannot be improved simultaneously, as both cycles occur inside the cylinder. Efficient and fast condensation relies amongst others on a high heat transfer through the outer wall of the cylinder, while the efficiency of expansion relies on no heat transfer at all; ideally the cylinder would always run warm. As such, the cylinder would ideally be infinitely thermally conductive in one state and infinitely insulated in another; a physical contradiction when viewed from a TRIZ perspective. In other words, this is a case of excessive integration, as two functionalities with opposite targets for the same parameter, with the Watt engine separating the condensator and cylinder, thereby separating the two in space, vastly improving the optimisability of the system.

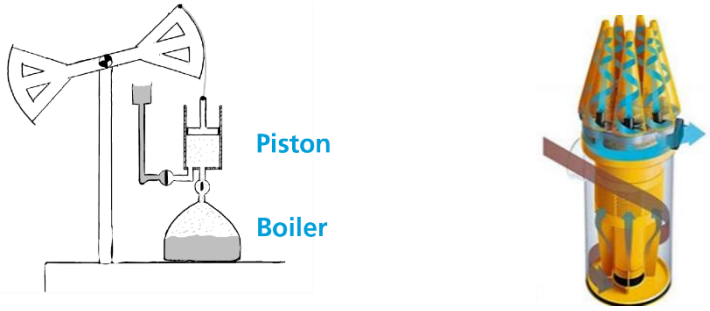


Figure 6 - Left) The Newcomen steam Engine (adapted from Suh (2001)), Centre) A bag-based vacuum cleaner (Right) A cyclonic separator from a Dyson Vacuum.

Changes in dependency can be difficult to exemplify, given that the design changes involved can be quite significant, not to mention the often large amount of design parameters and objectives involved. **A simple example of positive dependency** however, is the difference between vacuum cleaners with bags and bag-less vacuums. Regular vacuums created an airflow using an electric motor and fan, which then sucks air through a hose, with dust and debris being filtered out by an intermittent bag and secondary filter. The filtration in the bag is critical to avoid debris causing damage to the motor, but also to ensure that the dust is captured and not blown out again. Looking at two objectives, suction pressure and filtration quality which are of the type max-is-best, a trade-off reveals itself. The more efficient the bag is at filtration, the more resistance it creates, hindering the flow. In other words, the better the filtration, the more powerful a motor is required to generate a given suction pressure at the end of the hose. This also means that the suction pressure is reduced, the more the bag is filled. Bag-less vacuums meanwhile, commonly rely on cyclonic separation - a process that incurs less loss to the suction path. In fact, the filtration quality increases with the suction pressure, making the vacuums with cyclonic separation more optimisable and robust by design, with regards to these two objectives.

4 DISCUSSION - IMPLICATIONS OF TRADE-OFFS IN DESIGNS

As shown and discussed in the prior section, trade-offs can often be handled through process control in manufacture, requirement change, and optimisation. Such approaches are widely research and already supported by numerous well established methodologies and frameworks, (e.g. Papalambros & Wilde

(2017), Marler &Arora (2004), Mistree *et al*. (1994)). Yet, these are only truly applicable when the achievable feasibility domain is sufficiently wide and the achievable performance is sufficient. As such, trade-offs between functional objectives cause substantial issues relating to:

- **Tight tolerances or out-of-spec performance -** organisations that manufacture products that only function within too narrow sweet-spots between design objectives, will either have to rely on tight tolerances, or live with scrap, an increased functional dispersion, and the added reliability issues given the of two-sided variation driven failure that follows these dependencies.
- **Reduced performance** - trade-off situations limit the feasible domains of a design due to dependencies rather than constraints, reducing the achievable optimum.
- **Increased lead time** - With an increase in functional integration in a system, the risk of trade-offs grows proportionally. Designing highly integrated products is in the authors' experience often a matter of managing a "*system-of-sweets-spots"*, where each design change with one objective has a negative cascading effect across the system. The result is a highly iterative design process where time is spent on tweaking the design to continuously find new sweet-spots.

In other words, improperly managed trade-offs can result in less robust products that are less optimisable and more challenging to design. The decision-making involved in whether to accept, optimise or change a design, is in other words a cardinal process in designing high-performance, robust products. Yet it does not seem that there is any approach to support this process - i.e. helping the designer identify and classify the functional trade-offs in a design, and subsequently decide how to manage it (i.e. select one of the outlined approaches). Instead, the only way for the designer to identify the right strategy, is to actually apply all of them in parallel, and then compare the outcome.

Given that this is a time-consuming approach, experience shows that the *accept-optimise-redesign* decision is made based on tacit knowledge, and sometimes based on comparative analysis of the consequences of some of the options (e.g. assessing the tolerance required to accept the design, vs. the impact of changing the requirement). Based on the authors' experience from practice in numerous industries, this leads development teams to unconsciously applying all three approaches simultaneously in an unstructured manner, with substantial coordination complexity to follow. It is for instance not uncommon for parts of a team to start investigating the influence of requirement or specification change, while some colleagues investigate potential redesigns, and others attempt to tweak parameter values and evaluate the design through simulation or experiments. This is further complicated in multi-disciplinary applications such as electro-mechanics and mechatronics, where these trade-off scenarios become more multi-dimensional. While the strategies discussed could also apply in these cases, they are perhaps not exhaustive, with other options existing, given that redesign in e.g. the software or control domains do not necessarily imply radical changes to the system itself.

In lieu of the above, and the obvious benefits in designing functionally integrated products, why not attempt to design systems in a way that reduces the risk of functional trade-offs impacting robustness and performance? As discussed by Matthiassen (1997), a more comprehensive approach to integration is required, but no research has so far pointed to when to integrate and when to separate, beyond to do what is beneficial to the system. It would seem obvious that integration should be performed with trade-off management in mind, with a focus on embodying systems in a way that avoids severe trade-offs by design. In this context, the following conclusions can be drawn based on the prior sections:

- **Optimality and robustness by design** - A mechanical design will be optimisable and often robust, when all the objectives can be improved on without detriment to others. As such, the theoretically ideal integrated mechanical system only has positive dependencies.
- **Allocation of functionality** - Functionality integration and part reduction should, when possible, only be performed when the types and monotonicities of the objective functions are either the same, or when both type and monotonicity are simultaneously opposite.

How is this achieved? Systematic methodological support within this domain is scarce (c.f. sec. 2), which is surprising given the amount of methodological support within optimisation and trade-off acceptance. While TRIZ and axiomatic design both address the notion of avoiding detrimental dependency, they do necessarily not support the designer in deciding between acceptance, optimisation, or redesign, both in principle prescribing that all detrimental dependency should be solved by design. Furthermore, they primarily focus on approaches to removing the underlying the dependency (figure 3.e), rather than changing it (figure 3.f). There is in other words a potential in providing methodological support for designers aimed at managing trade-offs, specifically on how to decide between *acceptance, optimisation,* and *redesign*, and how actually perform said redesign.

5 CONCLUSION

Trade-offs between functional objectives can have a significant impact on the performance of a mechanical system and the complexity and lead time of designing it. In the context of ensuring robustness and optimisability, six different approaches to managing trade-offs have been identified, falling within one of three categories; *accept*, *optimise*, or *redesign*. Yet there is no approach to assist designers in identifying the approach that best suits a given design, nor is the aspect of how to perform redesign to remove a trade-off well described. This can drive designers to select the wrong approach at the cost of robustness and/or optimisability. What this points to, is a vast potential in further research within methods for the identification, classification and management of trade-offs between objectives, and embodiment design methods that support optimal functional integration in mechanical systems.

REFERENCES

Ahmed, S., Wallace, K. and Blessing, L. (2003), "Understanding the differences between how novice and experienced designers approach design tasks", *Research in Engineering design*, Vol. 14 No. 1-1, pp. 1.

Altshuller, G.S. (1984), *Creativity as an exact science: The theory of the solution of inventive problems*, Taylor & Francis Group.

Andreasen, M.M. and Howard, T.J. (2011), "Is Engineering Design Disappearing from Design Research?", Chapter 2, *Future of Design Methodology*, Springer-Verlang, London.

Andersson, P. (1997), "On Robust Design in the Conceptual Design Phase: A Qualitative Approach", *Journal of Engineering Design*, Vol. 8, pp. 75–89.

Arthur, W.B. (1993), "Why do things become more complex?" *Scientific American*, May 1993, p. 144.

Boothroyd, G., Dewhurst, P. and Knight, W. (2002), *Product design for manufacture & assembly*, Taylor & Francis.

Ebro, M. and Howard, T.J. (2016), "Robust Design principles for reducing variation in functional performance", *Journal of Engineering Design*, Vol. 26 No. 1-3, pp. 75–92.

Eppinger, S.D. and Browning, T.R. (2012), *Design structure matrix methods and applications*, MIT press.

French, M. (1971), *Conceptual Design for Engineers*, Springer, Berlin.

Frey, D., Palladino, J., Sullivan, J. and Atherton, M. (2007), "Part Count and Design of Robust Systems", *Systems Engineering*, Vol. 10 No. 3, pp. 2007.

Göhler, S.M. and Howard, T.J. (2015), "The Contradiction Index (CI): A New Metric Combining System Complexity and Robustness for Early Design Stages", *Proceedings of the ASME IDETC/CIE 2015*.

Göhler, S.M., Frey, D. and Howard, T.J. (2016), "A model based approach to associate complexity and robustness in engineering systems", *Research in Engineering Design*, pp. 1–12.

Ilevbare, I.M., Probert, D. and Phaal, R. (2013), "A review of TRIZ, and its benefits and challenges in practice", *Technovation*, Vol. 33, pp. 30–37.

Marler, R.T. and Arora, J.S. (2004), "Survey of multi-objective optimization methods for engineering", *Structural & Multidisciplinary Optimisation*, Vol. 26, pp. 369–395.

Matthiassen, B. (1997), *Design for Robustness and Reliability - Improving Quality Consciousness in Engineering Design*, Technical University of Denmark 1997.

Mistree, F., Hughes, O. and Bras, B. (1993), "The Compromise Decision Support Problem and the Adaptive Linear Programming Algorithm", *Structural Optimization: Status and Promise*, AIAA, Washington, DC.

Olesen, J. (1992), *Concurrent Development in Manufacturing – based upon dispositional mechanisms, PhD-Thesis*, Technical University of Denmark.

Otto, K. and Antonsson, E. (1991), "Trade-off strategies in engineering design", *Research in Engineering Design*, Vol. 3 No. 2, pp. 87–104.

Pahl, G. and Beitz, W. (1996), *Engineering Design: A systematic approach*, Springer, Berlin.

Papalambros, P. and Wilde, D. (2017), *Principles of Optimal Design - Modelling and Computation*, 3rd edition, Cambridge University Press, Cambridge.

Ross, A., Hastings, D. and Warmkessel, J. (2004), "Multi-attribute Tradespace Exploration as Front End for Effective Space System Design", *Journal of Spacecraft and Rockets*, Vol. 41 No. 1.

Suh, N.P. (2001), *Axiomatic Design - Advances and Applications*, Oxford University Press.

Wagner, T.C. (1993), *A general decomposition methodology for optimal system design, Doctoral dissertation, Dept. of Mechanical Engineering*, University of Michigan.

Vianello, G. and Ahmed-Kristensen, S. (2012), "A comparative study of changes across the lifecycle of complex products in a variant and a customised industry", *Journal of Engineering Design*, Vol. 23 No. 2.

ACKNOWLEDGMENTS

The authors would like to thank the Danish Innovations Fund and the Novo Nordisk STAR-programme for funding this industrial research project (grant nr. 7038-00221B).

Appendix 5: Paper D (Supplementary)

Title: Limitations of Design Space-based Indicators for Early Robustness Assessment
Authors: Juul-Nyholm, H. B.; Sigurdarson, N. S.; Ebro, M.; Eifler, T.
Publcation: The Proceedings of the 23nd International Conference on Engineering Design (ICED 21), held in Gothenburg, Sweden.

LIMITATIONS OF DESIGN SPACE-BASED INDICATORS FOR EARLY ROBUSTNESS ASSESSMENT

Juul-Nyholm, Herle Bagh (1);
Sigurdarson, Nökkvi S. (1,2);
Ebro, Martin (2);
Eifler, Tobias (1)

1: Technical University of Denmark;
2: Novo Nordisk A/S

ABSTRACT

This paper seeks to address the gap between qualitative Robust Design principles and parameter optimization. The former often fails to consider the challenging amount of details in embodiment and configuration design, while the latter is the widely accepted main thrust in traditional Robust Design. The gap is addressed by exploring the value of five quantitative robustness indicators for Design Space Exploration based on variables, objectives and constraints: The set level indicators, Design Space Size and Pareto Set Dispersion, and the point level indicators, Neighbourhood Performance, Failure Rate and Distance to Failure. As a background for the discussion of the limitations of these indicators an industrial case is presented. The case is an incremental encoder and includes two configurations for comparison, five objectives, eight variables, and a range of constraints. The design spaces are sampled and they show conflicting objectives, dispersed spaces and variables dependencies. Based on this it is suggested that set level indicators are more suitable than point level indicators of early robustness evaluation, but the available indicators are limited in their considerations of design space discontinuity and conflicts.

Keywords: Robust design, Computational design methods, Evaluation, Design Space Exploration, Configuration Design

Contact:
Juul-Nyholm, Herle Bagh
Danmarks Tekniske Universitet / Technical University of Denmark
Denmark
hbaju@mek.dtu.dk

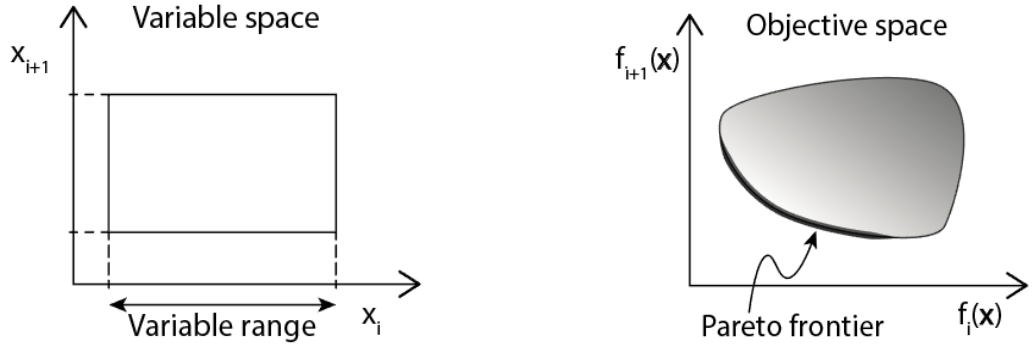
Cite this article: Juul-Nyholm, H. B., Sigurdarson, N. S., Ebro, M., Eifler, T. (2021) 'Limitations of Design Space-Based Indicators for Early Robustness Assessment', in *Proceedings of the International Conference on Engineering Design (ICED21),* Gothenburg, Sweden, 16-20 August 2021. DOI:10.1017/pds.2021.459

1 INTRODUCTION

Well-accepted in academia and practice, Robust Design (RD) provides an approach for ensuring the "insensitivity of products and processes against different sources of variation" without eliminating the sources of variation themselves (Taguchi, 2005). Seemingly offering a comprehensive development procedure, based on the three phases of the seminal quality engineering framework (1) *System Design*, (2) *Parameter Design*, and (3) *Tolerance Design*, RD consequently aims at developing products that show a consistently high quality and performance despite noise factors such as production variation in form of tolerances, unintended or variable load scenarios, ambient conditions of use, etc.
However, most authors agree that Taguchi's work on phase (2), i.e. the optimization of parameter settings by means of suitable experimentation strategies and the corresponding statistical analyses, is the main thrust in a traditional RD approach (e.g. Jugulum & Frey, 2007; Hasenkamp et al., 2009). Unfortunately, this implies a relatively narrow focus on one single, albeit important, design task in embodiment, which is the efficient optimization of parameter settings for a previously defined product configuration. As a consequence, traditional RD almost exclusively focuses on the time-intensive and often computational costly optimisation of parameter settings of a largely matured product solution, and hence completely ignores the possibility of improvements by design. This question is instead left to qualitative, early stage RD principles (Blanding, 1999; Suh, 2001), which oftentimes fail to address the challenges that come with the increasing level of details in embodiment and configuration design.
This paper seeks to address this decisive gap between early stage design principles and late stage optimisation approaches by exploring the value of quantitative Design Space-based indicators for an early robustness assessment in configuration design. The underlying reasoning, that these indicators could be applied to evaluate and compare configurations of a chosen concept, has a twofold basis. On the one hand, corresponding approaches allow for an integrated consideration of a product's functionality (objectives) and its structural characteristics (imposed constraints). Examples for the latter are geometric tolerance chains, which are largely relevant for the overall robustness, but often only considered towards the end of the development process. On the other hand, they also provide the possibility to extend conventional RD-thinking towards a rigorous considerations of resulting trade-offs, which have previously been identified as a largely relevant driver of product robustness (Göhler & Howard, 2015; Göhler, Frey & Howard, 2016b; Sigurdarson et al., 2019). As a consequence, the overall aim is to enable the engineering designer to avoid unnecessary iterations and to proceed to the embodiment task with confidence in the chosen configuration. In other words to improve robustness by design!

2 BACKGROUND AND METHODOLOGY

Design Space Exploration (DSE) is the iteration and exploration of the design or variable space (Fig. 1a) of feasible design points and the corresponding objective space (Fig. 1b) defined by objective functions describing the performance of a design configuration. The variable space is defined by the variable ranges and each design point is a set of variables and objective measures subject to design specific constraints. The design points fulfilling the constraints are feasible. In the objective space, the design point performance is evaluated based on optimality, i.e. either maximum or minimum. If two objectives cannot be optimal at the same time, as can be seen in Fig. 1b, the two objectives are conflicting. The conflict is represented by the Pareto frontier, or its approximation, on which none of the objectives for each point can be improved without deteriorating another.
While the variable space represents possible parameter settings, including potential variation, the correlation between variables and objectives represents sensitivity and the objective space represents the design output possibilities and the trade-offs inherent in the design. The latter, also referred to as contradictions, is of particular interest in the context of this paper. Based on the predictive value of trade-offs as robustness indicator for early product solutions (Göhler & Howard, 2015; Göhler, Frey & Howard, 2016b), DSE bears the potential to be used as early stage, quantitative RD assessment of configurations, and in this way to provide a valuable complement to the detailed sensitivity studies of traditional RD approaches.
The review of existing robustness indicators in the field of DSE started by identifying the relevant body of literature. Based on an initial literature review, which yielded a collection of 74 publications on indicators, nine publications containing the search terms *robust* and *Pareto* were identified. Subsequently,



(a) Variable space and variable range for the variables x_i and x_{i+1}.

(b) Objective space and Pareto frontier for minimization of the objectives f_i and f_{i+1}.

Figure 1. Visual definitions of the 2D variable and objective spaces.

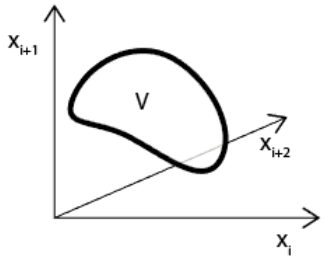
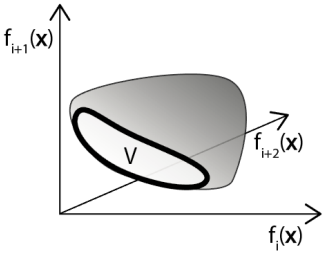
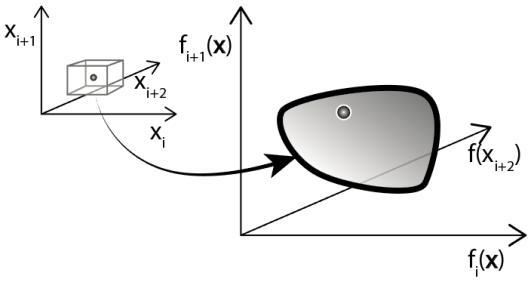
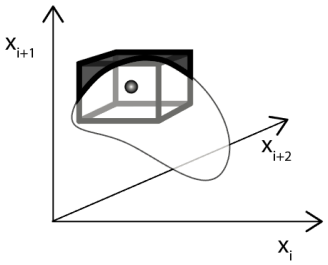
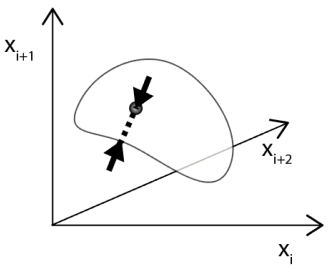
the indicators were grouped according to their information content, i.e. the necessary information for their calculation and the output provided, resulting in the classification presented in Sec. 3 below.
In order to discuss the challenges of the listed indicators, an exemplary case study from industry is presented. This anonymized case involves the design of an angular encoder. The case study is set up as a sampled design space exploration model to compare two encoder configurations. It was chosen as it combines multiple engineering disciplines and the design is subject to a range of demanding constraints. Furthermore, both constraints and objectives require evaluation of analytical expressions as well as simulations due to the incremental nature of the encoder scale. The chosen variables are also both continuous and discrete, highlighting the strengths of a design space exploration approach. All of this makes the case suitable for a discussion of robustness indicator limitations. Variables, objectives and constraints were identified through semi-structured interviews with design engineers in the development project, and the results are visualized by their design and objective spaces for easy comparison.
Based on the presented case study, the limitations of the identified indicators are discussed. The discussion is based on the extend to which the indicators address the challenges of the case, the required information and the computation effort required.

3 REVIEW OF ROBUSTNESS INDICATORS FOR DESIGN SPACE EXPLORATION

In Tab. 1 five indicator types are listed and illustrated. In the following, the short descriptions of Tab. 1 is elaborated and the indicators' implications on robustness is introduced.
The size of the design space represents the ranges of feasible design variables as a function of the constraints. Opposed to the shortest distance to a constraint, the design space size indicates the room for maneuver in choosing design points with the achievement of the maximum allowable tolerance at the center of the design space. The design space size can be evaluated as the full hypervolume of the design space as indicated in the illustration or as a box of independent variable ranges.
The size of the Pareto set on the other hand, indicates how contradictory the objective functions are in the feasible design space. It is calculated as the hypervolume between the Pareto frontiers in the objective space represented by the black edged volume in the illustration. In design space exploration the Pareto set can only an approximated as this method does not allow an evaluation of whether the set fulfils mathematical optimality criteria.
Neighbourhood performance provides an extended view on the sensitivity of objective performance to perturbations in a single variable by evaluating the performance of a specified variable neighbourhood of a design point. The neighbourhood performance can be evaluated as the average performance of the neighbourhood, which would allow for inclusion of effects of non-linear objective functions, or the worst case performance of the neighbourhood.
Failure rate is a measure of how many parts will violate constraints based on the variable distribution. The failed outcomes are represented by the dark shaded part of the box in the illustration. Rather than focusing on the worst case scenario as some neighbourhood performance indicators, it relates the nominal design point statistically to large scale production outcomes.

The shortest distance from a point to failure indicates the maximum size of a design point tolerance, which has implications for the required production control. It can be measured either on specific variable axes or as a Euclidean distance as suggested by the arrows in the illustration.

Table 1. List of identified robustness indicators for design space exploration. References in right column: [1] (Göhler, Eifler & Howard, 2016a), [2] (Beyer et al., 2007), [3] (Harbrecht et al., 2019) [4] (Rötzer et al., 2020), [5] (Riquelme et al., 2015), [6] (Barrico et al., 2006), [7] (Yannou et al., 2007), [8] (Frank et al., 2018), [9] (Deb et al., 2006)

#	Indicator & Short Description	Illustration	Ref.
Indicators on Set Level			
I	**Design Space Size** The n dimensional volume of the design space, either in terms of range-based, independent boundaries or the total space size. It indicates the room for maneuver in choosing a design point and hence the global robustness of the design. It requires suggested variable ranges, objective functions and constraints.		[1][2] [3][4]
II	**Pareto Set Dispersion** The n dimensional volume of the objective space between the Pareto frontiers between all objective pairs. It indicates how conflicting the objective functions are for the current design configuration. It requires suggested variable ranges, objective functions and constraints.		[5]
Indicators on Point Level			
III	**Neighbourhood Performance** The objective performance of the chosen neighbourhood of the design point based on either average, variance or worst case. This indicates the sensitivity of the chosen set as it describes the correlation between the design and the objective space. It requires a design point, desirable tolerances as well as the variable set and the objective functions.		[1][2] [6][7] [8][9]
IV	**Failure Rate** The rate of outcomes failing due to violation of constraints for the chosen variable space and variation distribution type. This indicates the scrap rate of a mass produced product at the chosen, nominal design point. It requires the design point, constraints, objective functions and distribution characteristics.		[1][2]
V	**Distance to Failure** The distance from the chosen design point to the closest constraint, either in terms of variable units or as a Euclidean distance. This indicates the quality margin of the chosen design point and hence the local robustness of the design. It requires a design point, objective functions and relevant constraints.		[1]

4 ENCODER CONFIGURATION CASE

In the following, a configuration case study for an encoder will be introduced and used to assess the applicability of the indicators identified in Sec. 3.
The case study focuses on the physical configuration of the reader and scale of an incremental, conductive encoder that is retrofitted into an existing mechanical device to register and transmit angular displacements. The reader is produced from sheet metal through cutting, bending and punching and the scale pads are separate conductive pads on the surface of a print board. The displacements correspond to critical events and the aim of the retrofit is to improve the functionality of the mechanical device by offering a digital logging of these events. Furthermore, the mechanical device is mass produced and a robust design is of paramount importance.
The fact that the encoder is an add-on leads to a series of predetermined size and functionality constraints. The encoder cannot be serviced, calibrated or replaced, and the cost has to be kept at an absolute minimum, and needs to be produced in an enormous volume (above 10^7 units/year). On top of this, the displacement that needs to be measured, are incremental and have a predefined maximum range, which means that the outcomes are a known discrete set with its own unit.
Exemplified in Fig. 2, the encoder is of the conductive and incremental type as the registration of displacement is achieved as a current conducting reader is sliding over a scale consisting of separate conductive pads representing logically interpretable code digits. Hence, the rotational, mechanical input is converted to an electrical code, which is interpreted by the software as an angular displacement corresponding to a number of units belonging to the outcome set.
The configuration should maintain the original mechanical function of the device to the largest extent possible while ensuring a adequate measuring accuracy including electrical signal quality and logical code interpretability. These demands cannot be achieved without compromises. Yet, the required degree of compromise depends on the choice of the encoder configuration.

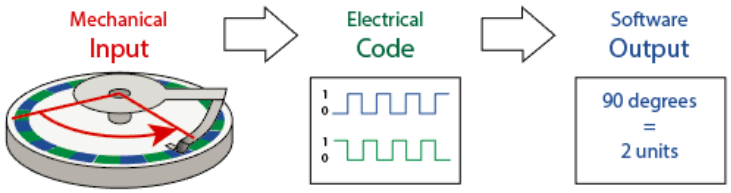
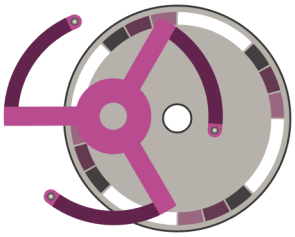
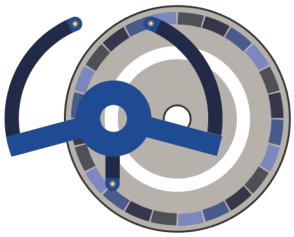


Figure 2. Encoder working principle.

The aim of this case study is to compare the robustness potential of two encoder configurations. These are shown in Fig. 3a-3b. Configuration 1 has one scale track, while Configuration 2 has two scale tracks. The white pads represent the ground connection which is essential for the conduction of current. The colored pads represent different code pads, like the green and blue pads in Fig. 2.



(a) Configuration 1 with three arms on one track and different code pads with grounding in between.



(b) Configuration 2 with two arms on the four code track and a separate track and arm for grounding.

Figure 3. Encoder configurations.

5 DESIGN SPACE EXPLORATION FOR ENCODER CONFIGURATION

In this section the model variables, objectives and constraints are described. They are based on a number of assumptions and therefore the results offer a means of comparison of the two configurations in

Fig. 3a-3b. The encoder configurations were explored in MATLAB by looping through the objective functions with two identical sets of random design point samples. Subsequently, the feasible samples were identified as well as the local Pareto optimal samples for each configuration.

5.1 Variables

The embodiment of the configurations were varied based on the eight design variables illustrated in Fig. 4. The dimensional variables subject to variation are sampled randomly and uniformly within a set of chosen limits. They represent both noise factors and control parameters as there is no need to distinguish between them when analyzing for dependencies (Göhler, 2017). All variables are continuous except the number of scale pads, *m*, which divides the configurations into subconfigurations. Design parameters like material properties, production capabilities etc. are included in the model as well.

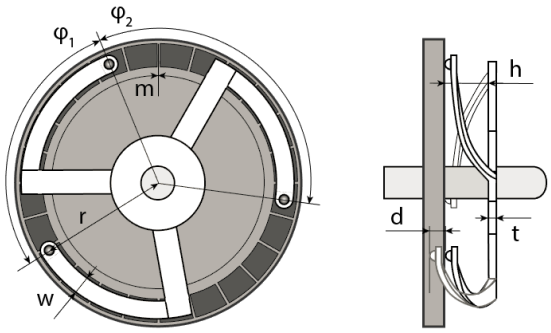


Figure 4. Configuration model variables.

5.2 Objectives

The objectives of the encoder design are to impact the original function of the mechanical device as little as possible, to ensure electronic registration of angular displacement as well as robustness against variation in the position of reader and scale.

Four of the five design objectives are analytical expressions subject to a multitude of assumptions but also computationally inexpensive. For modelling of encoder outputs however, simple simulations were developed. This was possible because the design space is explored by sampling.

The height of the encoder should be minimized to have as small an impact on the existing components of the mechanical device as possible. Due to the focus on the reader design, the total encoder height only varies with the thickness of the wiper plate, *t*, and the distance from the reader plate to the PCB surface, *h* (Eq. 1).

Minimization of the frictional torque of the reader due to the scale contact, seeks to ensure as small an effect on the displacement to be measured due to the addition of the encoder (Eq. 2).

Sensitivity to swash (tilting between reader and scale planes) is undesirable because the signal strength is a function of the contact force. Assuming a constant tilting angle, ϕ, to obtain a relative objective, the difference between the minimum and maximum contact force is a function of the contact radius, *r*. The contact forces are calculated assuming a constant tilting angle and assuming that the reader arms are straight and slender with rectangular cross sections. It is desired that the minimum and maximum contact forces are equal, which is why the objective function ratio is to be maximized (Eq. 3).

Minimization of constriction resistance is representing the function of the electrical code circuits connected to the scale pads. The lower the resistance the lower the required power. In the interface between the reader arm and the scale, the contact resistance is a sum of the constriction and the film resistance. The constriction resistance occurs because the current is constricted to travel through a reduced area across the rough interface, while film resistance is caused by a thin layer of dirt and oxides on the surfaces, which has a higher resistivity than the conductive bulk material. The film and bulk resistances of the conductive materials are not included in the model as it hard to quantify and will be apply to both configurations. The constriction resistance is a function of this radius and the resistivities of the reader and scale pad surfaces, which are the same (Eq. 4).

The maximization of repeatability in terms of the number of pads per unit depending on the starting point is important in order to avoid ambiguities in the code interpretation. The challenge of repeatability is

pronounced in the case of this particular encoder, because the encoder can neither be calibrated during production nor use. The main contributor to the repeatability challenge, apart from dynamic effects, is the random starting position of the wiper arms on the scale pattern. The repeatability objective is evaluated by simulating rotations of 1-30 units for a range of starting positions and counting the number of pads passed. The most frequent pad count per dose, i.e. what the controller would be programmed to interpret as one specific dose, is used as a reference when calculating the repeatability for each dose. Because the objective is to be maximized, the lowest calculated repeatability among the doses is chosen as the objective measure for one particular sample. The number of starting positions where chosen based on a convergence study (Eq. 5).

$$\min \quad t + h \tag{1}$$

$$\min \quad T_{\mu} \tag{2}$$

$$\max \quad \frac{F_{min}}{F_{max}} \tag{3}$$

$$\min \quad \frac{\rho}{2r_a} \tag{4}$$

$$\max \quad \text{Rep.} \tag{5}$$

5.3 Constraints

The encoder configurations are subject to the following constraints. The configuration embodiment should:

1. ensure constant grounding.
2. ensure a continuous code signal on a full rotation.
3. ensure any unit interpretation regardless of the starting point on the scale.
4. not exceed the maximum yield stress of the reader arm material.
5. not exceed the minimum requirement for the contact resistance.
6. not exceed the allowable diameter of the retrofit in the mechanical device.
7. be manufacturable in terms of reader and pad dimensions.
8. ensure rotation despite friction loss due to the reader arm contact.

5.4 Results

The DSE model was run with $2 \cdot 10^6$ design points distributed uniformly across a range that complies with the size requirements and was identified through interviews. The number of feasible solutions for Configuration 1 and 2 were 562 and 176 respectively. The results are presented in Fig. 5 as selected variable and objective spaces, respectively, including all eight variables and all five objectives. The visual interpretation of these will be discussed in Sec. 6.

6 DISCUSSION

6.1 Case Results

Four selected variable spaces for Configuration 1 and 2 can be seen in the upper part of Fig. 5. On the top left a dependency between the thickness, t, and the deflection, d, can be sensed for both configurations, though stronger for Configuration 1. The global optima are quite evenly distributed in d, but for t the optimal tendency is on the low end of the range. On the top right of the variable space plots, the number of scale pads per revolution can be seen in relation to the reader height. The discrete steps in m each represent subconfigurations of Configurations 1 and 2 and for Configuration 1 only four out the seven are feasible due to differences in division of three and four digit codes. Not only does Configuration 2 have one third of the number of feasible solutions compared to Configuration 1, its solution space is also more dispersed, especially in t and h, and might be discontinuous. The discontinuity can however not be assessed due to the combination of the sampling size and the inclusion of the discrete variable, m. On the bottom left of the variable space plots, both feasible and globally optimal design points are dispersed across the sampled design ranges for both configurations in w and r. This might imply discontinuity, but also shows little dependency to optimality as the global Pareto points are dispersed

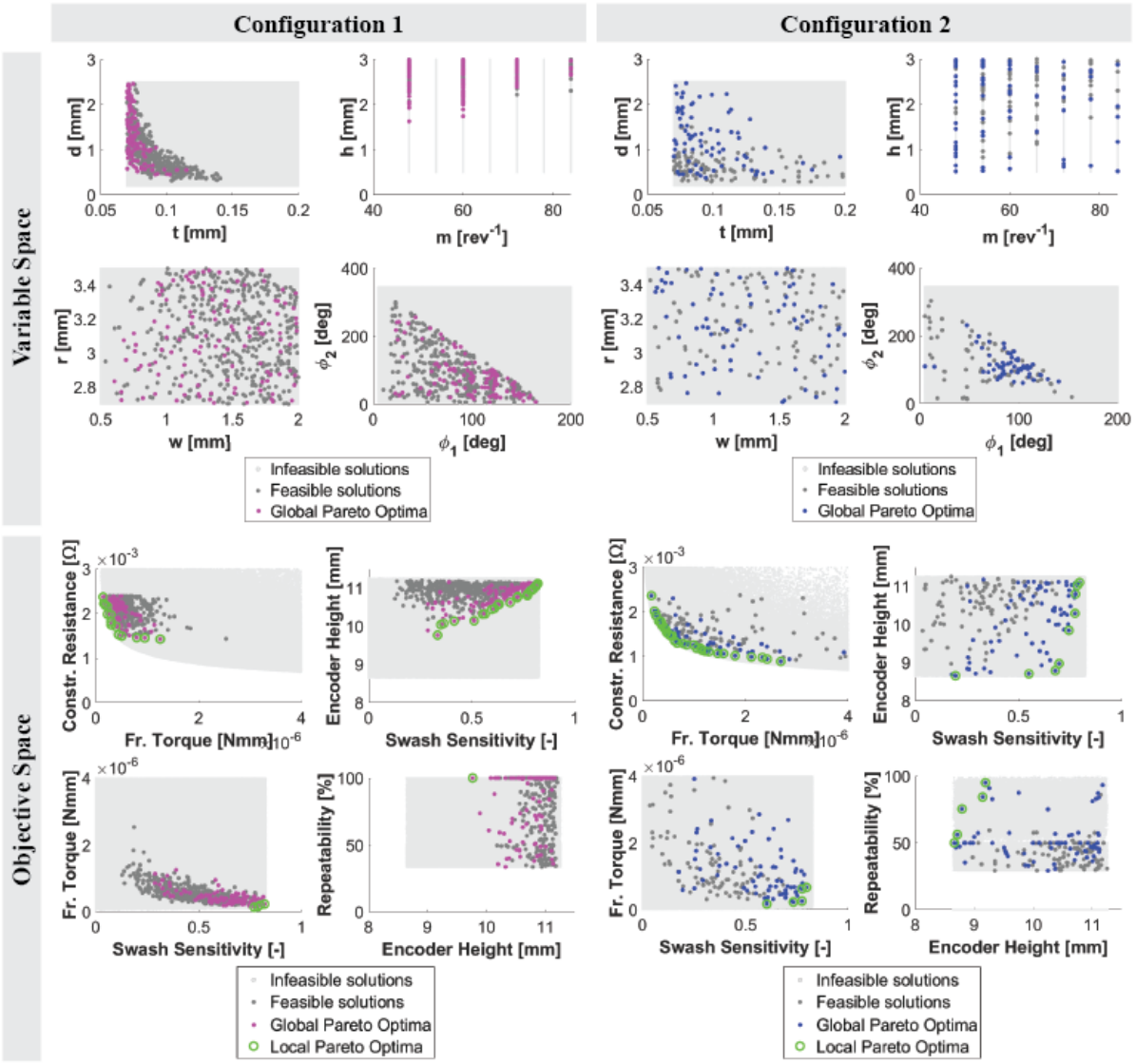


Figure 5. Variable and objective pairs for Configuration 1 and 2 including infeasible sets, feasible sets, global Pareto sets and local Pareto sets.

as well. On the bottom right of the variable space plots, the expected dependency induced by one of the manufacturability constraints (7.) creates a triangle between ϕ_1 and ϕ_2. For both configurations, clusters of feasible points can be seen, especially the three bands between 100 and 150 degrees for Configuration 1. This might be related to discontinuity due to either subconfigurations or constraints related to the discrete encoder scale.

In the lower part of Fig. 5 four selected objective spaces for Configuration 1 and 2 can be seen. On the top left, the set of grey, infeasible points shows that the objective space is shaped by the relationship between the Frictional Torque and the Constriction Resistance. On the top right and bottom left of the objective space plots the infeasible spaces are square, indicating that the concept and configuration does not predispose a conflict between the objectives. On the top right, however, the Pareto front is placed very differently in the space for the two configurations due to a difference in the influence of constraints. In both top right and bottom left, the feasible Swash Sensitivity objective is dispersed along the range for both Configurations, while this is only the case for the Encoder Height and Frictional Torque for Configuration 2. This might indicate design space discontinuity or a high sensitivity to variables, but this cannot be derived from the plots. In the bottom right of the objective space plots the Repeatability objective is clearly different for the two configurations. The majority of the feasible design points for Configuration 2 are below 50% repeatability and the Pareto set indicates a conflict, but with the scattered feasible set the objective space size is hard to assess visually.

As mentioned in Sec. 2, this case is a challenging design task due to the number of constraints and objectives, and as shown in the variable space plots of Fig. 5 the constraints and objectives lead to both conflicts and limited design spaces for both configurations.

6.2 Limitations of Robustness Indicators

The five indicators listed in Tab. 1 each have their strengths and weaknesses and are applicable at different design stages due to differences in required in- and outputs, as can be seen in Tab. 2. Indicator I and II requires variable ranges as input, while Indicators III-V requires a design point. While III and IV requires tolerances as inputs, I and V provides information about feasible tolerances. On the output side, I and II consider global configuration characteristics, while III-V consider local information directly related to tolerances as well as sensitivity and scrap. Depending on the defined neighbourhood, III might however also reflect global characteristics, for example in cases of monotonicity. Based on this, the set level (I,II) indicators are applicable for coarse exploration of variable feasibility and conflicts, while the point level (III-V) indicators more often support the evaluation of a chosen design point by offering a detailed variation performance for more mature designs.

Table 2. Summary of limitations of robustness indicators for design space exploration.

Indicator **In-/Output**	**I** **Design Space Size**	**II** **Pareto Set Dispersion**	**III** **Neighbourhood Performance**	**IV** **Failure Rate**	**V** **Distance to Failure**
Variable ranges	Input	Input			
Objective Functions	Input	Input	Input	Input	Input
Constraints	Input	Input	Input	Input	Input
Design point			Input	Input	Input
Tolerances	(Output)		Input	Input	Output
Variation Distributions				Input	
Sensitivity			Output		
Scrap Rate				Output	
Optimality		Output	Output		
Conflict		Output			

The case and DSE model presented in Sec. 4-5 show discontinuity due to the constraints and the discrete variable (number of scale pads, m), which will result in issues both for set and point level indicators. The set level indicators would need to divide the spaces into continuous ones and the point level indicators would be unsuitable for the identification of alternative spaces. Furthermore, the case study shows two different kinds of objective conflicts; objective function and constraint induced. None of the indicators take this difference into account even though it provides important information about how to manipulate the design space and hence the obtainable robustness, e.g. through changing the configuration. The objective conflicts are also interesting on the topic of dispersion, i.e. the length of the Pareto frontiers, as it might indicate the sensitivity of the design depending on the dispersion of the corresponding variable space. This is not covered in the set level indicators either.

The presented robustness indicators differ from Taguchi's work on parameter design as they focus on robustness of multiple configurations in the early embodiment stage based on trade-offs and design flexibility as opposed to optimization of one configuration that has already been matured. This is supported by the correlation between trade-offs and robustness identified by Göhler, Frey & Howard (2016b) and its relevance is further highlighted by the multidimensional and discontinuous design spaces of the exemplary encoder configurations. Exploration of design spaces and conflicts through the reviewed indicators could increase the predictability of designs for further detailing and optimization.

The presented review and discussion of robustness indicators for DSE was based on theory and the challenges highlighted by the industrial case. It would however be interesting to implement and evaluate the indicators quantitatively and maybe even apply them to a set of diverse cases to assess their applicability based on degree of dispersion, discontinuity or conflict as well as design space size and objective types.

7 SUMMARY

The gap between early and late stage RD has been addressed through a review and classification of available indicators for DSE. Five indicators were identified: Two set level indicators, Design Space Size and Pareto Space Dispersion, and three point level indicators, Neighbourhood Performance, Failure Rate and Distance to Failure. Based on an industrial case with two encoder configurations highlighting conflicting objectives and demanding constraints as well as a table showing indicator in- and outputs, a discussion of the limitations of the five indicators were presented. For early evaluation of configuration robustness, set level indicators are more suitable than point level indicators, but the available indicators are limited in their considerations of design space discontinuity, conflicting objectives due to constraints or objective relationships as well as the robustness implications of the shape of Pareto frontiers.

ACKNOWLEDGMENTS

The authors would like to thank Novo Nordisk A/S for financial support to the DTU-NN Robust Design program.

REFERENCES

Barrico, C. & Antunes, C. H. (2006) *Robustness Analysis in Multi-Objective Optimization Using a Degree of Robustness Concept*. 2006 IEEE Congress on Evolutionary Computation, 1887-1892.

Beyer, H. G. & Sendhoff, B. (2007). *Robust optimization – a comprehensive survey*. Computer methods in applied mechanics and engineering, 196(33-34), 3190-3218.

Blanding, D. L. (1999). *Exact constraint: machine design using kinematic principles*. American Society of Mechanical Engineers.

Deb, K. & Gupta, H. (2006). *Introducing robustness in multi-objective optimization*. Evolutionary computation, 14(4), 463-494.

Frank, C. P., Marlier, R. A., Pinon-Fischer, O. J., & Mavris, D. N. (2018). *Evolutionary multi-objective multi-architecture design space exploration methodology*. Optimization and Engineering, 19(2), 359-381.

Göhler, S. M. & Howard, T. J. (2015). *The contradiction index (ci): a new metric combining system complexity and robustness for early design stages.* International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (Vol. 57175, p. V007T06A023). American Society of Mechanical Engineers.

Göhler, S. M., Eifler, T., & Howard, T. J. (2016) *Robustness metrics: Consolidating the multiple approaches to quantify robustness*, Journal of Mechanical Design, 138(11).

Göhler, S. M., Frey, D. D., & Howard, T. J. (2016). *A model-based approach to associate complexity and robustness in engineering systems.* Research in Engineering Design, 28(2), 223-234.

Harbrecht, H., Tröndle, D., & Zimmermann, M. (2019). *A sampling-based optimization algorithm for solution spaces with pair-wise-coupled design variables.* Structural and Multidisciplinary Optimization, 60(2), 501-512.

Hasenkamp, T., Arvidsson, M., & Gremyr, I. (2009) *A review of practices for robust design methodology*, Journal of Engineering Design, 20(6), 645-657.

Jugulum, R. & Frey, D.D. (2007) *Toward a taxonomy of concept designs for improved robustness*, Journal of Engineering Design, 18(2), 193-156.

Riquelme, N., von Lücken, C., & Baran, B. (2015). *Performance metrics in multi-objective optimization.* In 2015 Latin American Computing Conference (CLEI) (pp. 1-11). IEEE.

Rötzer, S., Thoma, D., & Zimmermann, M. (2020). *Cost Optimization of Product Families Using Solution Spaces.* In Proceedings of the Design Society: DESIGN Conference (Vol. 1, pp. 1087-1094). Cambridge University Press.

Sigurdarson, N., Eifler, T., & Ebro, M. (2019). *Functional Trade-offs in the Mechanical Design of Integrated Products - Impact on Robustness and Optimisability.* Proceedings of the 20th International Conference on Engineering Design.

Suh, N.P. (2001) *Axiomatic Design: Advances and Applications*. United States, OUP.

Taguchi, G., Chowdhury, S. & Wu, Y. (2005) *Taguchi's Quality Engineering Handbook*, Wiley & Sons.

Yannou, B., Troussier, N., Chateauneuf, A., Boudaoud, N. & Scaravetti, D. (2007) *Design exploration, robust design and reliable design: Three successive and complementary approaches*. International Conference on Engineering Design, ICED'07.

Technical
University of
Denmark

Nils Koppels Allé, Building 404
2800 Kgs. Lyngby
Tlf. 4525 1700

www.mek.dtu.dk