What do we lose when machines take the decisions?

Bolander, Thomas

Published in:
Journal of Management & Governance

Link to article, DOI:
10.1007/s10997-019-09493-x

Publication date:
2019

Document Version
Peer reviewed version

What do we lose when machines take the decisions?

Thomas Bolander

2019

Abstract

This paper concerns the technical issues raised when humans are replaced by artificial intelligence (AI) in organisational decision making, or decision making in general. Such automation of human tasks and decision making can of course be beneficial through saving human resources, and through (ideally) leading to better solutions and decisions. However, to guarantee better decisions, the current AI techniques still have some way to go in most areas, and many of the techniques also suffer from weaknesses such as lack of transparency and explainability. The goal of the paper is not to argue against using any kind of AI in organisational decision making. AI techniques have a lot to offer, and can for instance assess a lot more possible decisions—and much faster—than any human can. The purpose is just to point to the weaknesses that AI techniques still have, and that one should be aware of when considering to implement AI to automate human decisions. Significant current AI research goes into reducing its limitations and weaknesses, but this is likely to become a fairly long-term effort. People and organisations might be tempted to fully automate certain crucial aspects of decision making without waiting for these limitations and weaknesses to be reduced—or, even worse, not even being aware of those weaknesses and what is lost in the automatisation process.

Keywords: Artificial intelligence (AI), connectionist AI, symbolic AI, explainability, trust, algorithmic bias, algorithmic decision making, human decision making.

The last few years have seen an explosive increase in industrial-scale applications of artificial intelligence (AI), and an even higher increase in the expectations of the problems and tasks that AI will be able to solve in the near future. There is no doubt that AI will have a profound impact on many aspects of our lives, jobs, and society as a whole. However, it is much less clear exactly what the impact will be. Many human cognitive tasks can seemingly be automated by AI, but we risk a loss in predictability and explainability when doing so.

The title of the paper, "What do we lose when machines take the decisions?", might seem a bit loaded. Why should we expect to lose anything at all? Why is it about loss and not gain? First of all, let me underline that when I talk about loss or gain, it is exclusively about the quality of the decision making, not about e.g. financial aspects. Automating human tasks through AI and robotics can potentially lead to significant financial gains. I will not try to assess these

potential financial gains, but simply point to how and where automated decision making might be different from human decision making—in particular where it might still not be on par with the human decision making it is sought to replace.

The reason we potentially might lose something by automation through AI is that—despite the goal of AI being to simulate aspects of human cognition—human and machine intelligence are still fundamentally different. We can not yet communicate with AI systems the way we communicate with fellow humans, and AI systems cannot explain their own reasoning and behaviour the way humans can. Human and machine intelligence have a rather different set of strengths of weaknesses. Some tasks that are easy for humans to solve have turned out to be exceedingly difficult for machines, and vice versa. To understand what we potentially lose by automating decision making, it is necessary to first understand these differences between human and machine intelligence. The paper will first give a brief overview of what AI is and the main paradigms it consists of, and then turn to address specifically the potential weaknesses of AI decision making compared to human decision making.

1 What is artificial intelligence (AI)?

More than 60 years ago John McCarthy, the father of artificial intelligence (AI), defined the field as "the science and engineering of making intelligent machines, especially intelligent computer programs." The complication of this definition is that we do not know exactly what intelligence is, and hence even less what it means for a computer program to be intelligent. In the 1950s and 1960s, AI was expected to develop very rapidly into computers and robots with human-level cognitive capabilities. This however did not happen, at least not yet.

Lacking a precise definition of AI, we can still give an approximate characterisation. It is almost always about building machines—computers or robots—that can perform tasks that otherwise only humans have been able to, e.g. play chess, drive a car, do medical diagnosis, or engage in a dialogue. Furthermore, *when* such machines are built, AI researchers are almost always directly inspired by how humans solve the same task. It can be in terms of the machine directly trying to mimic some of the neurological processes of the human brain (see *connectionist AI* in Section 4 below) or it can be via a more abstract model of human problem solving, e.g. an approximate model of the reasoning steps involved in a human deciding the next move in a game of chess (see *symbolic AI* in Section 4 below).

AI today is a wide range of different techniques for simulating different aspects of human cognition. Computers can play chess, drive cars, recognise skin diseases and engage in dialogues, but all these applications are based on different techniques within AI—and require individual programming tailored for the specific application at hand. This makes current AI very different from human beings solving similar tasks. Human beings can learn to master all of these different tasks during their lifetime without having to be preprogrammed specifically to solve them.

2 Characteristics of current AI

AI systems tend to be tailored to specific types of applications, and often new types of applications requires new methods to be developed (or existing methods to be combined in a novel way). Developing a robust driverless car is not just about taking an existing AI system from the shelf, plug it into the car and then let itself figure out how to drive. Given that AI systems need to be tailored to specific applications, the complexity of building such systems depend crucially on how well-defined and clearly delimited the problem to be solved is. As a rather robust rule of thumb, the more well-defined and clearly delimited a problem is, the easier it is to make AI that can solve it. With this in mind, it is not too surprising that already in 1997 it was possible to build a computer program, IBM Deep Blue, that was able to become world chess champion. Chess is extremely well-defined and clearly delimited: there is only a very few and strict rules to obey, and there is a very precisely formulated goal to achieve. For a human being, chess has an overwhelming combinatorial complexity in terms of possible move sequences, but modern computers are not easily overwhelmed by the need to consider an enormous number of options. Deep Blue could compute 200 million chess moves per second.

Modern computers are however much more easily overwhelmed by problems in which the rules or the goal—or both—are less clearly formulated and delimited. One such example is driverless cars. As in chess, there are also rules to obey in traffic, but there are many more rules than in chess, and they are much less formally specifiable. It is even more complicated for computers to successfully small talk with a human for a few minutes over a cup of coffee. *Chatbots* are computer systems for engaging in dialogue with humans, and the dialogue can either be in writing or through a voice interface. Building a chatbot that can engage successfully in small talk with humans is exceptionally difficult, as the rules of such dialogues are even much less clear than the rules in traffic. This is also why a lot of the chatbot technology that was in the early 2010s implemented on web pages to help customers, e.g. by IKEA, Scandinavian Airlines and Deutsche Post, has now been taken out of use. There seems to be a new wave of chatbots arising here on the edge to the 2020s, and they are probably better, but it does not change the fact that general natural language dialogue is everything but well-defined and clearly delimited, and hence extremely hard to bring to human level on a computer.

It is interesting to compare the difficulty for computers on the three task mentioned above, playing chess, driving cars, and chatting, with the difficulty for humans. For most humans, the relative difficulty of the three tasks is opposite the one for computers: It is easier for most humans to engage in small talk for a few minutes than it is to drive a car safely through downtown Rome on a Friday afternoon, which again is easier than becoming world chess champion. Figure 1 illustrates this. The fact that the two axes of difficulty are opposite one another illustrates that intelligence is not just *one* thing, and that different types of intelligence cannot always be compared along the same axis.

Many people seem to have the view that computers are becoming more and more intelligent, and that it is just a matter of time before they become more intelligent than us humans. But that view assumes that we can directly compare human and machine intelligence on a single axis of intelligence. The figure illustrates that it might not be as simple as that. Humans and machines

Figure 1: The axis of difficulty of a task for machines is opposite the axis of difficulty for humans.

currently have very different strengths and weaknesses, and there is no simple way of comparing their "level of intelligence". Computers will forever be better than humans at board games with high combinatorial complexity, but no matter how explosive the development of AI is going to be, it is conceivable that we humans will forever be better natural language users (not the least since natural language was invented by us, and developed in a way that is to a large extend dependent on our culture, brains and bodies [9]).

3 Human-machine dualism

The difference between human intelligence and current level machine intelligence is so large that it is probably more relevant to talk about a duality. We humans have a very flexible intelligence, are good at abstract thinking and conceptualising the world. Often we are good at solving problems that are not very clearly delimited and well-structured (but where the solutions do not have to be either). Conversely, machines are primarily good at clearly delimited and well-structured problems, but can then also provide solutions that are very precise and well-structured. They are much less competent at abstract thinking and conceptualising the world, though a lot of research is invested in developing AI systems that have these competences as well.

The human-machine duality can be illustrated by the case of the IBM Watson system that was originally developed for playing the game of Jeopardy [4]. Jeopardy is about answering questions concerning trivia knowledge, which is not a particularly well-defined and clearly delimited problem, but still much more well-defined than general dialogues. In 2011, IBM Watson became (unofficial) world-champion in Jeopardy. It didn't do so by being better at understanding the questions, but by compensating somewhere else: The system had 200 million pages of text in memory and could read through most of that text in a single second. This is of course far beyond what any human can do. The point here is that humans are actually much better at understanding questions and finding answers from relatively small amounts of data (the small amounts we can keep in our brains), but computers can in some cases compensate by having access to enormous amounts of data and being able to process that data with

exceptional speed. And in some cases, as with Watson, they can compensate so well that they actually outperform humans on certain tasks.

If the answer to a question is well hidden in a piece of text, Watson is likely not to find it. However, it compensates by having access to an enormous text library that it can browse through in seconds, and then it will probably find another source, where it is more trivial to extract the answer. When humans look for answers in texts, we can only read extremely slow compared to computers, but we have a much deeper understanding of what we read, and are very good at finding the deeper meanings and the well hidden answers.

One of the important conclusions of the Watson example is that even when we succeed in constructing a computer program that achieve above human level on a certain task, it doesn't at all imply that it solves it in the same way as humans, and therefore we cannot really use it to conclude anything about the relative intelligence between humans and machines. In contrast to Watson, humans can also answer questions that nobody has asked before, e.g. whether a crocodile can run a steeplechase [11]. In order to answer such a question, we need our human ability to create mental models of the content of the question, that is, to picture the poor crocodile with its extremely short legs trying to jump a high barrier. We use our rich existing models of the world to answer questions, whereas Watson simply tries to look the question up in its enormous library, and if the answer is not there because nobody considered that question before, it has no chance of answering.

4 Symbolic AI

Since the 1960s, AI research has essentially been divided into two competing paradigms, the *symbolic paradigm* and the *connectionist paradigm* [6]. These two paradigms have completely opposite approaches to simulating aspects of human cognition. The symbolic paradigm follows a top-down approach by trying to directly simulate the highest levels of human cognition, our linguistic (symbolic), conscious reasoning. In this paradigm, one tries to build AI systems that have an explicitly represented language to reason about the world, and for instance use this to do logical inference or plan a sequence of actions to achieve a certain goal. AI within the symbolic paradigm is behind systems such as chess computers like IBM Deep Blue, the 200.000 robots working in the warehouses of Amazon around the globe, and intelligent personal assistants such as Apple Siri, Google Assistant and Amazon Alexa. The advantage of symbolic AI is that the systems constructed can be made *robust*, *predictable* and *explainable*. If Siri always misunderstands specific types of questions, Apple can inspect the code and understand why, and hence improve it. The drawback of symbolic AI is, though, that systems within this paradigm tend to have strictly delimited abilities, and they normally do not learn from experience.

5 Connectionist AI

Arguably, one of the main landmarks of human intelligence is our flexible intelligence and our ability to learn. We are not born with the ability to play chess or drive a car, but learn it during our lifetime. If we want artificial intelligence

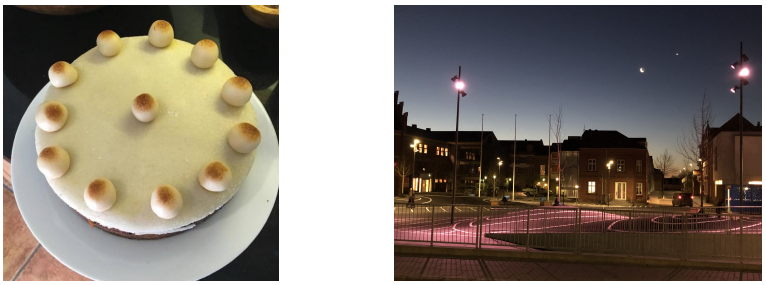

Figure 2: Two pictures that were blocked by the image recognition system of Instagram.

systems to share these abilities with humans, we have to consider the area of *machine learning* within AI. Machine learning is a very broad term that covers any AI algorithm that does not have a static behaviour, but can learn from experience. It could e.g. be an algorithm with the ability to learn to distinguish objects in the physical world, with the ability to learn better strategies in chess, or with the ability to learn the rules of new games. Some of the techniques of machine learning belong to the symbolic paradigm, but the currently most prominent ones belong to the *connectionist paradigm*. The connectionist paradigm is essentially constituted by AI techniques based on *(artificial) neural networks* (*ANN*s), including *deep neural networks* (*deep learning*).

In artificial neural networks, one tries to simulate the atomic processes of the human brain: the functioning of the individual neurons and neuron connections. Hence opposite the top-down approach of symbolic AI, connectionist AI has a bottom-up approach to simulating aspects of human cognition. The connectionist approach is behind image recognition software, e.g. for recognising skin diseases or as used by Instagram to decide whether a picture contains unacceptable nudity. The advantage of the connectionist approach is that it is possible to construct systems that have a more flexible intelligence and can learn from experience. The neural network employed at Instagram has not been programmed with a model of what unacceptable nudity is, but has simply been trained on a very large set of pictures that has in advance been manually labelled as either "acceptable" or "unacceptable". Eventually, the system itself has learned to recognise the patterns of unacceptable content. It would never be possible to do the same with symbolic AI: There is no way to give a sufficiently precise linguistic or symbolic definition of what "unacceptable nudity" is.

The drawback of the connectionist approach is, however, that it can never be 100% predictable, error-free or explainable. The systems build according to this approach are based on statistical learning from experience. When you do statistical learning from experience, your ability to correctly categorise new objects, for instance recognise certain types of pictures, gradually improves but can never become 100% precise. An example of this is illustrated by the two pictures in Figure 2. They don't seem to bear many similarities, and they seem to be fairly acceptable pictures from everyday situations. However, they were both blocked by the image recognition system of Instagram (in 2015 and 2019, respectively), and both were claimed by the system to contain unacceptable

nudity. There is currently no way of knowing exactly *why* these pictures were labelled as unacceptable, as the neural networks train an implicit model with millions of neuron weights, and it is the combination of all these neuron weights that decide the classification the network makes. There is also no simple way of telling the neural network that easter simnel cakes like the one on the left of Figure 2 are not examples of human nudity. The only way to try avoiding such misclassifications in the future is to provide the algorithm with more labelled pictures that it can train on—and then hope for the best. This is clearly very different than the situation in symbolic AI, where models are explicit rather than implicit, and can hence easily be inspected and modified.

6 Symbolic versus connectionist AI

An essential difference between the paradigms is that symbolic AI is based on creating *explicit (symbolic) models*, whereas the connectionist approach is based on learning *implicit models*. This difference roughly corresponds to the difference between trying to predict a ballistic trajectory using the laws of mechanics and aerodynamics (explicit model) versus simply trying to learn from observing a high number of such trajectories (implicit model). When humans throw snowballs, there is no doubt that we use some kind of learned implicit model to predict where the snowball will land, which is consistent with the connectionist approach. However, when we play a game of chess or plan a dinner party, there is equally no doubt that we also use explicit symbolic (linguistic) models to reason about our possible action sequences, which is consistent with the symbolic approach. So it seems that human problem solving combines implicit and explicit models, and that certain aspects of problem solving are closest to the connectionist approach, whereas others are closest to the symbolic approach. For this reason, the last few years have seen a high increase in attempts at constructing AI systems that combine the symbolic and the connectionist approaches, with some of the notable examples being Google DeepMind building a system that taught itself to play old Atari arcade games [12], and building another system achieving world-class level in the game of Go [17]. Connectionist AI is mainly about simulating aspects of our perception system, whereas symbolic AI is mainly about simulating aspects of our higher cognition. Fully autonomous AI systems like driverless cars or household robots of course need both.

It seems to be hard in AI to get what we could otherwise reasonably have expected. We would like AI systems to be robust, predictable and explainable, which would push us towards symbolic AI. However, we would also like AI systems to be flexible and learn from experience, which would push us towards connectionist AI instead. There seems to be a fundamental and unavoidable trade-off involved: The more intelligent, flexible and easily trained we want a system to be, the less we have control over the system and the less we can guarantee it to behave in the intended way. It implies that not all demands for computer software and robots can be met by simply turning up the level of intelligence and flexibility of those systems. For many types of system, for instance database systems and e-voting systems, we still want to be able to prove that they have and will always maintain the intended behavioural properties.

7 Classifying the weaknesses of AI relative to human decision making

Many of the examples provided above illustrate that AI still has weaknesses compared to human intelligence. This doesn't imply that we should generally just try to minimise the use of AI. In Denmark, the company Corti developed an algorithm that listens to calls to the emergency central and uses pattern recognition to try to determine whether it is a case of cardiac arrest. If the algorithm assesses it to be so, it shows a warning on the screen of the medical professional taking the call. The algorithm has quite a number of false positives (warning when there is no cardiac arrest) and false negatives (not warning when there is a cardiac arrest), however the net effect is a higher number of detected cardiac arrests, and the price paid is only to send out a few more ambulances than one would otherwise have done [3]. Here, the combination of an AI system and a human expert provides better problem solving than the human alone.

Detecting cardiac arrest only based on the voice and breath of a human is clearly a perception task, and we would not even expect humans to be able to explain exactly what patterns made them suspect that one case is cardiac arrest and another is not. So adding AI doesn't seem to give any loss, neither in quality of decisions, nor in explainability and transparency of the decisions. The only immediate threat to such use of AI is if humans start having blind faith in the algorithmic predictions and stop using their own cognitive ressources to detect cardiac arrests and critically evaluate the algorithmic warnings. The two fatal Tesla crashes in autopilot mode that will be discussed further below happened exactly due to the drivers having achieved so much faith in the robustness of the autopilot that they stopped having their hands on the steering wheel and eyes on the road, even though the instructions for the autopilot specifically asked them to.

A rather demanding, but still reasonable, principle for the deployment of AI is to say that it can and should only be employed for partial or full automatisation when we can guarantee that it is better than what it replaces on all parameters. That is:

1. When a task is (partly) automatised by AI, we should require at least the same quality of problem solving as before automatisation—on all parameters.

2. AI should always enhance transparency, fairness and explainability, not diminish them.

For instance, you can't necessarily say that an autopilot or driverless car is better than a human driver even if it is involved in fewer crashes. If these crashes are completely incomprehensible from a human perspective and we have no way of explaining why they happened, we have a significant loss of transparency and explainability.

To get a clearer view of where AI might have weaknesses compared to human agents, let us look at the different stages of algorithmic problem solving. An algorithm takes an input that is then processed to provide an output, see Figure 3. Humans also process inputs through our senses and turn them into outputs: actions. If what a human does is different from what a machine does in

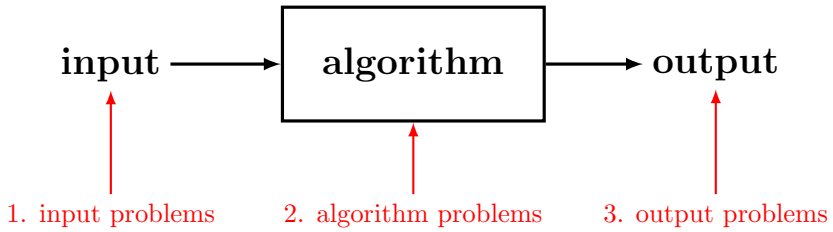

Figure 3: Illustration of an algorithm and the potential sources of shortcomings.

a particular situation, it must be due to differences in the input, in the processing of data—the "algorithm"—or in the output. We will consider these different sources of disparity in sequence below.

8 Input problems

Input problems is to a large extend about data. An algorithm is fed with data, often both during the learning phase and when later used for decision making. The performance of the algorithm can never be better than the data it is fed with ("garbage in, garbage out" as they say in machine learning). We can identify at least three possible problems related to input data:

1. Missing data points.
2. Missing data dimensions.
3. Errors or bias in data.

Missing data points Algorithms like neural networks build implicit models representing patterns in the data they are trained on. If we want a neural network to be able to distinguish pictures of cats and dogs, it is not enough to feed it with a couple of pictures of dogs and a couple of pictures of cats, as it will not be enough to learn the general patterns distinguishing the two types of animal. However, if we have too little training data (too few pictures) to distinguish cats and dogs, we will soon realise this when we start testing or employing the neural network. So this type of lack of data is not the most problematic.

It becomes more problematic when we are potentially not even aware that we are missing data, for instance if we are missing training data for certain types of objects or patterns that are required for our AI system to safely make the required decisions. A relevant example of this problem is seen in car autopilots. In June 2016, a Tesla in autopilot mode made a fatal crash into the side of a semi-trailer. The Tesla drove at high speed and didn't even slow down when it was approaching the semi-trailer [10]. The problem was that the autopilot didn't even know that it was approaching a semi-trailer, even though it had seen that there was a huge rectangular object in front of it. Semi-trailers are big rectangular surfaces with text on, and so are highway signs. So from the perspective of a pattern recognising neural network they are very much alike,

Figure 4: What is it?

and the autopilot mistook the semi-trailer for a highway sign that it would have been possible to drive under. Indeed even humans might in some cases only be able to tell the difference between highway signs and semi-trailers from observing the context the pattern appears in. Consider the picture in Figure 4. With this picture I was able to fool a deep neural network to think that it is seeing a semi-trailer, but the sign itself also looks very much like a semi-trailer seen from the side except it is located high in the air and have some unexpected text on it. Not the pattern itself, but the context it appears in, makes it clear what we are seeing in this case. The importance of context will be addressed further in Section 9 below.

The Tesla autopilot has only been designed to be used on highways, and hence probably not trained specifically to recognise semi-trailer trucks turning out in front of it—and not tested on such scenarios either. The Tesla driver was not using the autopilot as intended, but the example is still interesting, since we can never guarantee the behaviour of an algorithm when it is being faced with input of a type that it hasn't been trained on. If we use algorithms for detecting whether a picture of skin change is cancer or whether a call to the emergency central is a case of cardiac arrest, then we know in advance exactly what we want to be able to categorise. This is however not true for autopilots and driverless cars. There can be many unexpected situations in traffic, and even human drivers can experience situations in traffic that doesn't look like situations we have been in before, situations that we haven't been "trained on". However, in such cases most humans are still are able to assess the situation and make some kind of sensible decision, because we have very rich general models of the physics of the world (how things move) and of other agents (inferring their intentions).

Algorithms so far only have rather limited models of rather limited parts of the world, and as soon as we step outside the border of what the algorithms have been trained on, they are likely to fail miserably. Even though neural networks are learning algorithms, it doesn't mean that an autopilot with a perception system based on a neural network automatically learns from its mistakes. In March 2019, another Tesla in autopilot mode was involved in exactly the same

type of fatal crash, driving into a semi-trailer at high speed.

The main point to be aware of here is that the less well-defined and predictable the environment of an AI system is, the less we can be guaranteed its sensible behaviour. Unless an environment is very well-defined and predictable (like, say, chess or pictures of skin changes), even humans can rarely completely define all the relevant situations that can occur in the environment. Hence even the AI programmers can not make sure that the system will act sensibly in all situations.

Missing data dimensions I have provided examples concerning missing data points, that is, relevant situations or patterns that the algorithm has not been trained on. Another type of lack of data is when the data points are there, but are lacking relevant dimensions. A simple example could be to attempt to fully automate the diagnosing of skin diseases by training an algorithm on pictures of skin changes. A human dermatologist might however also need to touch the skin and ask questions about how it feels in order to correctly diagnose the condition. In that case, the algorithm only trained on pictures has a much worse chance of correctly diagnosing the condition, since there are important dimensions that the algorithm simply doesn't have access to. Another example could be driverless cars using camera sensors with significantly lower resolution, higher latency and much more noise than the human eye has (imagine to attempt to drive a car remotely over a poor Skype connection). It might seem that it ought to be obvious whenever the data we feed into our algorithms are lacking dimensions compared to humans solving the same problems. In the examples just given, it indeed seems straightforward. However, we humans are not always aware of the relevant dimensions.

One of the impressive recent breakthroughs in natural-language processing are so-called *language models*. These are neural networks trained on massive amounts of text in order to be able to recognise patterns in text and hence also be able to generate new text. I can for instance feed the beginning of this section of my paper into one of these language models (the public version of GPT-2 [14]), and ask it to continue the text:

> As a result, it is not surprising that algorithms fail in many situations where humans are likely to perform the right decision. If we use such algorithms, it is not only possible for them to make mistakes, but also for them to be able to make mistakes that are completely unexpected to be made by humans. It is also possible that these mistakes will be missed by the user, and thus be used to make the wrong decision and make the crash. Therefore it is important to be aware of this potential problem and to design our algorithms as carefully as possible, to avoid this issue in the first place [11, 12].
>
> The second kind of problem with input data is related to the fact that if data does not exist, and we want to know an object is there, we need to look at the images again and search for objects we have not seen before. This is not the easiest task, especially if we are using the object for example as a target for a laser, but if we are not, then it will be far worse. Let me first explain how it can be more difficult to perform this task. In one of the famous movie "Star Wars" scenes, the Rebel Alliance uses the laser to shoot down a rebel

ships in a space battleship. The Rebel, in the movie, at first shoot the Rebel and then uses the laser to shoot down a ship...

The text is, at first, impressive in the sense of being grammatically correct and initially matching the style and overall content of my paper. However, it soon starts to diverge significantly and, as a whole, it is hard to grasp the overall message of the text. The problem here is that the model is missing several crucial things in order to create texts with a clear and meaningful message. First of all, the training input for the language model has been text only, not giving the algorithm any chance of getting to understand the connection between words and their meanings in the physical world. It corresponds to trying to teach a non-Japanese speaking person to become proficient in Japanese by only being exposed to transcripts of dialogues in Japanese, and not being explained what any of the words mean or have any idea about the original context of those dialogues. This could potentially be sufficient to learn to form grammatically correct sentences, to say hello and goodbye, and it might even fool Japanese people for a short while to think that the speaker understands what is being said. But it would not be enough for the person to figure out how to convey a specific message in Japanese. Every word uttered in a dialogue would simply be based on a statistical estimate of the likelihood of seeing that specific word in that specific position in a dialogue, based on the training data. In order to convey a specific message, an algorithm of course additionally needs to be able to reason in terms of intended messages, and have the ability to take the perspective of the receiver in order to figure out how to best formulate it. Language models as GPT-2 does not have any explicit representation of intentions or other agents.

Another example of missing data dimensions comes from a small research collaboration project in 2017 between Nordisk Film A/S (the leading cinema chain in Denmark and Norway) and the IT University of Copenhagen, Denmark. The goal was to train a neural network to predict movie success in cinemas on a scale from 1 to 9. Nordisk Film was already making such predictions using human experts in order to decide where to show which movies and for how long. The input to the neural network was essentially all the textual information that could be found about each movie, including genre, budget, country, prequels, rating, cast, length, Google Trends posts, Twitter posts, Wikipedia entries, prerelease, competition, year, reviews, source material, amount of marketing, preorders. Even with all this data, the neural network was not even close to match the precision of the human experts. But this is to be expected. After all, the number of tickets sold is also about humans being able to watch the movie and subjectively assess its quality and target audience. The movie itself was of course not part of the input to the algorithm, and even if it had been, the algorithm would not be able to make sense of it. A similar project attempted to predict the next winner of the Eurovision Song Contest by feeding an algorithm with online data about each song. Again, it is hard to judge who will win a song contest without hearing the song, and without having any idea about human taste. For an algorithm to learn human cultural and aesthetic norms is insanely difficult if not impossible.

Errors or bias in data As mentioned earlier, a machine learning algorithm is never better than the data it is trained on. When the company LEO Innovation Lab recently tried to build an algorithm to recognise psoriasis from

pictures of skin changes and wanted to train it on pictures correctly labelled by human experts, they found that even the human experts had significant internal disagreement on the correct classification (even sometimes disagreeing with themselves when seeing the same picture several times) [7]. If an algorithm is trained on erroneously labelled data, it will of course learn the errors and forever replicate the errors in its decision making.

Training data can also be biased, hence leading to biased algorithmic decisions. A famous example is the system for assessing the risk of criminals committing new crime during parole that turned out to highly overrate the risk of black people and highly underestimate the risk of white people [1]. The most likely reason for this is that the training data already had a racial bias based on the historical data it was trained on.

9 Algorithm problems

Algorithm problems concern the limitations or weaknesses that AI algorithms might have in processing their input. We can identify at least the following three potential types of problems:

1. Oversimplification: Ignoring context or reducing dimensions.
2. Missing dimensions in the model.
3. Identifying correlations with causal relations.

Oversimplification: Ignoring context or reducing dimensions Cases of context insensitivity in classification algorithms has been illustrated in Figures 2 and 4. For an algorithm to believe it most plausible that the image in Figure 4 is a semi-trailer, it most be very ignorant of the fact that semi-trailers rarely are attached to gantries high above the ground. In fact the algorithm I asked to classify the picture also suggested that it could be a road sign, it just assigned much lower probability to that option. When an algorithm only tries to recognise the geometrical pattern of the main object in the picture and ignores the context, it can easily be mistaken, as also humans can be. So why doesn't it take the context into account? The issue is that the more sensitive to context we want an algorithm to be, the harder it naturally is for it to learn any generalisable patterns. Often the ability to learn general patterns is about *ignoring* the context, e.g. being able to recognise the same cat independent of which background we see it on.

The tendency of the current generation of neural networks to be somewhat insensitive to context can be problematic in the employment of such methods e.g. in organisational decision making. Fair and valid assessments often require detailed data, preserving contextual information. The former Danish government in March 2019 released a National Danish AI Strategy [15]. It included a so-called signature project concerning targeted employment efforts. It stated: "With the help of artificial intelligence, it will potentially be possible to reduce the period of unemployment. By analysing patterns in historical data on successful efforts, the caseworker will have a better opportunity to target employment efforts to the individual citizen". A potential issue with such a project is that in order to be able to find any generalisable patterns at all, it could be

necessary to severely reduce the dimensions of the data or ignore significant parts of the context of each data point. Then one might exactly end up generalising beyond what is reasonable, and take decisions on patterns identified in isolation from their context.

Earlier I argued that missing dimensions in the input data could be a problem for machine learning algorithms. Now I suggested that, conversely, too many dimensions can also be a problem. The problem with too many dimensions is however not a problem with the input itself, but a problem with the algorithm that has to learn from it. Just as algorithms can have a hard time learning generalisable patterns if they have to be sensitive to the context of each data point, they can struggle to learn generalisable patterns if the input has too many dimensions.

The more dimensions the input data has, the more complex the algorithm handling it has to be. For machine learning algorithms like neural networks this means a significant increase in trainable parameters, and the more trainable parameters, the more training data is needed. For algorithms predicting human behaviour, the training data will of course always be limited by the number of humans we have existing data on, and for very high-dimensional data this means it will be impossible for the algorithm to learn patterns in the input. The solution to these problems is then often to reduce the dimensions of the data, that is, include fewer attributes of each person. Fewer dimensions of data means smaller models, and also models that have an easier time generalising the patterns observed (which is the crucial aspect of algorithmic learning). However, this might obstruct fairness in algorithms used for assessing humans like the considered examples of predicting recidivism and predicting the effect of employment efforts.

Missing dimensions in the model Missing dimensions in the model concern algorithms that are blind to certain aspects of the world they are trying to model—dimensions that might be required for the decisions they are supposed to make. To be able to model the mental states of other agents, their intentions and beliefs, some kind of Theory of Mind is necessary [13, 2]. Not everything humans do depend on the ability of creating mental models of other agents, but many thing do. A general practitioner has to be able to take the perspective of the patient and explain his diagnosis in a way that the patient understands, providing the right amount and level of information. This is impossible to do without any representation of the mental model of the patient, without knowing what the patient already knows and how the patient understand things. Even a driverless car probably needs some basic abilities in social perspective-taking in order to decipher social situations in traffic, including the possible intentions of other road users.

A lot of research effort goes into providing AI systems with Theory of Mind abilities, but so far the success is limited. One could have hoped that Theory of Mind and other advanced cognitive abilities would emerge from artificial neural networks when they became sufficiently large, but this is not what has happened. So if we want to have algorithms that can think logically, form plans, reason about other agents, etc., we need to specifically program them to have these abilities. Most of the AI systems equipped with such abilities are still based on symbolic AI, even though there is an increased focus on trying to achieve the

same with connectionist AI, or with combining the two paradigms.

If a neural network is excellent at performing perception tasks, but doesn't have anything that remotely resembles a Theory of Mind, we should of course be cautious as to what we use such an algorithm for. We should not naively believe that neural networks can be used for decision making in situations where we have already established that some Theory of Mind reasoning is required to make optimal decisions. For instance, consider the use of AI in performance management. In performance management it is important that the management understand how the employees perceive the world, and it is important to align the overall organisational goals with the employee goals. Performance management decisions can then of course not be taken by agents that do not have the ability to understand how the employees perceive the world and what goals they might have, i.e., agents with no Theory of Mind.

Identifying correlations with causal relations Another potential issue with machine learning algorithms is that they are trained to find statistical correlations. Such correlations might or might not be caused by causal relationships. If there is no causal relationship, we often would not like the algorithm to make a connection. One example is the company Lenddo doing credit scoring based on the users' smart phone usage including messaging, browsing activity, activity on social media, battery level, etc. [8]. Their trained algorithm found a strong correlation between average battery level and creditworthiness. The natural interpretation is that people who most of the time have a low battery level are in general more risk seeking, and hence also less creditworthy. However, one can discuss whether it is fair to reject a loan to a person just because that person is bad at charging her phone. Even if low battery level might statistically speaking be correlated with low chance of repayment, there certainly isn't a strict causal relationship.

It is quite easy to find spurious correlations that do not signify causal relationships. An entire web page, `tylervigen.com`, is devoted to point out such spurious correlations and hence bring light to the issue. For instance, according to the web page, there is a correlation of 0.99 between the divorce rate in Maine and the per capita consumption of margarine in the US in the period 2000-2009. We wouldn't like an algorithm to make predictions concerning the margarine consumption in the US based on the divorce rate in Maine, but if an algorithm had access to those figures, it might very well make the connection. The issue is that the pattern of correlation is indeed present in the data. Humans might frown upon such correlations and dismiss them for not representing causal relationships, but how do we know? We only know because we already have a very rich model of the world where such causal relationships seem unlikely. Machine learning algorithms don't have such rich models of the world, and for them any strong statistical correlation could as well be a causal relationship.

10 Output problems

One of the main output problems of AI algorithms is that they are often just algorithms for making a specific classification or taking a specific decision but without *any* means to explain the reasoning behind it. If we would like an AI system to decide whether to grant parole to an inmate, or decide what

employment effort to offer a citizen, it is problematic if we only get a decision but no explanation—even if the decision is almost always correct. The citizen offered a particular employment effort might reasonably ask for an explanation, and might also reasonably request to engage in a dialogue concerning the choice. If the algorithms themselves are expected to become able to explain their reasoning and engage in dialogues with humans, they most likely will need a high level of linguistic and social intelligence, including Theory of Mind. In some cases, it might be enough that a human expert can explain the algorithmic decision, possibly using some additional algorithmic tools. This leads me to introduce two different types of explainability, *explainability of* vs *explainability by*.

Explainability *of* vs explainability *by* What does it require to deem a decision explainable? Explainability *of* means that we have some external means to explain what led an agent to a particular decision. Explainability *by* means that the agent itself can explain its decision, and we could possibly also engage in further discussion with the agent to get a deeper understanding of what lies behind the decision.

Imagine a hedgehog taking a risky decision to cross a busy street. This decision can at best be explainable *of*. Through our research in animal behaviour we might be able to develop sufficiently rich models that we can use to explain and interpret such behaviour. However, the decision can never be explainably *by*, since the hedgehog can not itself provide us with an explanation. If a human took the same risky decision, we would expect it to be explainable *by* that human. Humans can of course in general not give precise, full descriptions of all the reasoning leading to a particular decision, but at least we can in most cases point to central decisive factors that is usually sufficient for others to understand why a particular decision is made. This is partly because language is an integrated part of our decision making. When reflecting consciously about whether to do A or B, we use language to reason about the possible consequences of the two decisions.

In AI, we can also distinguish between explainability *of* and explainability *by*. We might achieve explainability *of* a neural network if we can add algorithms that can point to the decisive factors in the classification made by the network. This for instance takes place in tools like LIME that can point to the parts of a picture that were decisive in choosing a particular classification of it [16]. If applied to the picture in Fig 2 left, we might expect the marcipan balls to be highlighted as the reason to label the picture as inappropriate. This at least gives us a hint to why the wrong classification was made.

Explainability *of* can be helpful, but to achieve the same level of explainability—and hence trust—in AI that we can have in fellow humans, we probably need explainability *by*. The AI should itself be able to explain its decisions in a human-friendly way, and ideally we should also be able to have a dialogue with the system about the underlying reasoning (such dialogues could also, as in the world of humans, become central in becoming a better decision maker). To achieve such AI systems seems exceptionally hard unless we have an AI system that integrates language, planning and (logical) reasoning with learning and perception.

11 What are the solutions—if any?

The paper has pointed to numerous potential weaknesses related to the automation of human decision making. Based on these insights, I can make two recommendations. The first recommendation is to reflect carefully on what we use AI for, and whether the current level of the technology is sufficient for a full automation. If not, possibly a partial automation is possible involving some collaboration between algorithmic advice giving and human decision making. And for some types of tasks, automation is simply still not viable, e.g. in many cases where social or linguistic intelligence plays a crucial role.

The second recommendation is to invest significant research resources in addressing the most serious of the above mentioned weaknesses of AI decision making. This includes significant efforts in increasing the level of explainability of AI, and increasing the level of social and linguistic intelligence in AI. As mentioned already in the abstract, these are not easy problems, and are likely to require rather long-term efforts. Fortunately, we are not completely at loss at where to start attacking these problems. Symbolic AI scores high on explainability and transparency whereas connectionist AI scores high on ability to learn from experience. Symbolic AI is excellent for planning, logical reasoning and Theory of Mind, whereas connectionist AI is excellent for perception tasks. We need the best of both worlds and in particular we need to combine the strengths of both paradigms. If we succeed in a deep integration of symbolic and connectionist approaches, we might have a hope to get future AI systems that can both learn, reason about others, use language to explain themselves in human-comprehensible ways, and engage in dialogues with humans about their reasoning and decisions. That would make algorithmic decision making much more trustworthy and have a much larger general potential.

References

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, 23 May, 2016.

[2] S. Baron-Cohen. *Mindblindness: An essay on autism and theory of mind*. MIT press, 1997.

[3] Stig Nikolaj Blomberg, Fredrik Folke, Annette Kjær Ersbøll, Helle Collatz Christensen, Christian Torp-Pedersen, Michael R Sayre, Catherine R Counts, and Freddy K Lippert. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. *Resuscitation*, 138:322–329, 2019.

[4] David A Ferrucci. Introduction to "This is Watson". *IBM Journal of Research and Development*, 56(3.4):1–15, 2012.

[5] Ben Goertzel and Cassio Pennachin. *Artificial general intelligence*, volume 2. Springer, 2007.

[6] Achim G Hoffmann. *Paradigms of Artificial Intelligence: A methodological and computational analysis*. Springer, 1998.

[7] Martin Holm Jensen, Dan Richter Jørgensen, Raluca Jalaboi, Mads Eiler Hansen, and Martin Aastrup Olsen. Improving uncertainty estimation in convolutional neural networks using inter-rater agreement. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 540–548. Springer, 2019.

[8] Hope King. This startup uses battery life to determine credit scores. *CNN Business*, 24 August, 2016.

[9] George Lakoff. *Women, fire, and dangerous things*. University of Chicago press, 2008.

[10] Fred Lambert. Understanding the fatal Tesla accident on Autopilot and the NHTSA probe. *electrek*, July 1, 2016.

[11] Hector J Levesque. On our best behaviour. *Artificial Intelligence*, 212:27–35, 2014.

[12] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[13] D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, 1978.

[14] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 2019.

[15] Regeringen. National strategi for kunstig intelligens. Technical report, 2019.

[16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016.

[17] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.