Chapter 5
Between Natural and Artificial Intelligence

Digital Sustainability in High-Risk Industries

Stian Antonsen

Abstract Algorithms have always been a key topic in safety science, whether they are governing technology through computer programming or human actors through organisational procedures. However, when the term “algorithm” is not limited to the static pre-programming of expert knowledge algorithms with the ability to change themselves, a new branch of uncertainties appears. With the concepts of epistemic uncertainty and epistemic accidents as a backdrop, I discuss safety-related challenges with the use of artificial intelligence (AI) in high-risk industries. The aim is to highlight uncertainties inherent in AI, paradoxes for safety management and risk governance, as well as the human contribution to safety in future.

Keywords Artificial intelligence · High-risk industries · Uncertainty

5.1 Introduction

Big data, algorithms and artificial intelligence (AI) are the current buzzwords in debates around technological development and how it may affect our lives, including the organisations in which we work. In this, chapter I reflects on the relationship between natural and artificial intelligence and the human contribution to safety in future.

The chapter consists of five parts. I first examine the changing division of labour between humans and technology, and the way existing safety science knowledge can be of use for future challenges. I then move on to the topic of epistemic uncertainty and epistemic accidents and concepts that I have borrowed from [5]. With Downer’s concepts as sensitising devices, I turn to the way algorithms and artificial intelligence incorporate different sources of uncertainty. This includes reflections around the very concept of intelligence and the human contribution to safety in future. I conclude with the delineation of three paradoxes that need resolving before artificial intelligence can be used in high-risk industries.

S. Antonsen (✉)
NTNU Samfunnsforskning, Trondheim, Norway
e-mail: stian.antonsen@samforsk.no

H. Laroche et al. (eds.), *Managing Future Challenges for Safety*,
SpringerBriefs in Safety Management,
https://doi.org/10.1007/978-3-031-07805-7_5

For readers looking for weaknesses, I can be of assistance by highlighting a limitation that should be borne in mind when reading the text. Being a sociologist means that if I were to inspect the programming of complex algorithms, I would have a hard time understanding what I am looking at—I simply don't know the language it is built in. However, this is part of the well-known challenge of *explainability*—the ability of algorithmic systems and programming languages to facilitate not only communication of computations to physical computers, but to human beings, allowing them to understand, scrutinise and criticise as we do with any other text [3]. Nevertheless, the sociologist's perspective means that the discussion is restricted to the *logic* of algorithms, the uncertainties associated with algorithms and the role of humans in an algorithmic world.

5.2 The Changing Nature of Work

Algorithms have always been a key topic for the design of safe and reliable sociotechnical systems. Broadly speaking, an algorithm is a finite set of rules aiming to govern actions on the form of "if this happens, in this context, then conduct these actions, in this sequence". On the technology side, automated safety systems (e.g. emergency shutdown systems) are indispensable for controlling hazardous energy. They consist of computer-programmed rules for actions taken by technology, e.g. "if this set of criteria is satisfied, then do the following actions to shut down this list of systems". On the organisational side, standard operation procedures follow the same logic, albeit with different challenges regarding compliance to the predefined instructions: "If an operator opens a valve in a pressurised system, then open gradually at 5, 20 and 50% before fully opening the valve". Hence, algorithms are already everywhere but as high-risk systems are digitalised and more real-time data becomes available, the use of complex algorithms will be a key topic for safety management in the years to come.

Increased use of algorithms involves a change in the division of labour between humans and technology. There is a myriad of classifications in this domain [10, 13], but they all have to do with the allocation of functions and decision-making authority between humans and technology, usually along an axis between fully manual and fully automated operations. While the taxonomies are simplifications, rarely distinguishing between different modes of operation (e.g. normal operation vs. unforeseen situations), they provide a backdrop for distinguishing between different uses of algorithms.

Such classifications are well-known within cognitive ergonomics, giving rise to concerns regarding human-in-the-loop issues, e.g. challenges to situation awareness, deskilling or automation failures [6]. The crash of AF447 in 2009 illustrates the human-in-the-loop paradox: inconsistent speed readings made the autopilot disconnect, immediately changing the aircraft from a highly automated system to a highly manual system, at an altitude where pilots rarely fly manually [9]. The assumption behind this sudden transfer of tasks and decision-making from technology to humans is that the human pilots will be ready to take over in a split second. Ready in terms of manual flying skills, training and situation awareness. It is such assumptions and expectations inscribed in technology that I will be discussing in this chapter.

One sometimes gets the feeling that classifications of the division of labour between humans and technology are treated as maturity scales that systems are inevitably moving and even *should* be moving from left to right on the scale. It seems to have become common knowledge that the information processing capabilities of “artificial intelligence” (AI) far surpass the capacity of human intelligence and that decisions and actions should, therefore, be transferred from humans to technology to reduce the occurrence of human error. While this assertion may be valid when referring to algorithm-based tools used in stable contexts, the presence of major accident potential and a high level of variability makes this more problematic. Scales describing the division of labour between humans and technology should, therefore, be treated as a lens for the design of tasks and responsibilities within sociotechnical systems, considering where it may be wise to use some form of automation and where human capabilities outshine those of technology. This is by no means a new question for safety science, meaning that many of the lessons from the past are relevant for the challenges of the future.

However, when the term “algorithm” not only refers to the static pre-programming of expert knowledge but includes the ability for systems to change themselves and communicate autonomously with other technological actors, a new branch of challenges and uncertainties appears. This is where algorithms become far more than rules automating simple actions and recurring decisions, relieving human beings of routine tasks. They are no longer only replacing human actions; they are touted as alternatives to human *intelligence*. This raises a different set of questions—questions related to uncertainty. Uncertainty here refers to the way technology is always based on incomplete knowledge and assumptions. As these assumptions can never be fully tested, they will be sources of inevitable surprises as a system operates over time [5].

5.3 Uncertainty and Epistemic Accidents

The works of Downer [4, 5] are important when it comes to safety-related challenges in algorithms and AI. Drawing on perspectives from Science and Technology Studies, Downer sheds light on the way assumptions about reality are built into technology, models, tests and verification. This is a particular form of uncertainty, arising from assumptions about the world in which a piece of technology is going to operate, assumptions which can never be fully representative of the world. This forms the basis of a particular form of accidents, which he calls epistemic accidents.

Epistemic accidents are the results of specific events revealing holes in the knowledge underlying the tests and models devised to represent real-life operational contexts. [5, p. 83]

The point is that technology cannot be seen in isolation from the knowledge on which their design is based—assumptions about everything from material fatigue, the needed strength of physical barriers or the way a piece of software should work under different operating conditions. These assumptions cannot be tested in an experimental environment that is representative of all possible contextual variability in the reality in which the systems are put to work. This means that the success of ultra-safe systems like aviation is not only due to technology being tested in simulators and laboratories, but that part of the success comes from learning the hard way—through a history of disasters [5]. This raises some tricky questions about "residual risk" which are hard to speak about, both politically and ethically, but which are important issues to raise around the use of algorithms and AI in high-risk systems.

5.4 Assumptions and Uncertainty in Artificial Intelligence

Keeping Downer's concepts in mind when moving on to algorithms involved in AI, there is an obvious need to say something about the term AI. As is the case with the general concept of intelligence, the definition and understanding of artificial intelligence is contested [7]. For this discussion, a broad definition will suffice. I see AI as referring to any kind of computing technology that aims to mimic or otherwise resemble human intelligence. According to Boucher [2], existing AI technology can be divided into two waves. The first wave consists of "good old-fashioned AI" based on precise rules that are the encoding of human knowledge in contexts where there is little variability and where it is possible to specify right and wrong solutions by means of strict "if–then-else" rules.[1] While the first wave is human-driven, the second wave is data-driven in the sense that it consists of various forms of machine learning (ML). In ML, the learning does not consist only of humans refining the rules (algorithms) to improve performance but includes an ability to improve the fit of the rules through identified patterns in large quantities of data. This is where the terms "artificial neural networks" and "deep learning" (referring to artificial neural networks with at least two hidden layers) come in.

As the aim of the chapter is to discuss the logic of AI for use in high-risk domains, I will not go into the details of the techniques in question. Instead, I will delineate key steps in developing and modifying AI to show that the algorithms not only incorporate epistemic uncertainty into the systems stemming from the assumptions of its creator—it may also produce brand new uncertainties based on their own assumptions. My argument is that there is a form of epistemic uncertainty built into models and algorithms that gain a form of objectivity because they are seemingly untouched by human fallibility of judgement, while in fact they are not.

[1] This includes both systems where each variable has an absolute value true (1) or false (0), and systems based on fuzzy logic allowing any value between 0 and 1.

Let's take an example from supervised machine learning. Say we want to create an algorithm able to separate criminals from non-criminals based on images, a project carried out in 2016 [1].

The first thing we need is an existing data set consisting of images of *known* criminals and non-criminals. Since we already know who are criminals and non-criminals, we can label each picture accordingly, thereby providing the system with enough cues to allow a learning algorithm to be run. The system is sent looking for patterns in the pictures, singling out recurring differences in the labelled data. The learning algorithm has now created a programme where it knows what to look for—which markers in the images that explain the most variance in the data. Now comes the crucial step: exposing the programme to new data which it has never seen before. Its creator hopes it will make correct predictions when sorting the test data into the two labels (criminals vs. non-criminals). As the process goes on and the software gets feedback on its performance, the programme will change to improve its fit, based on how the learning algorithm is set up.

Where are the possible sources of uncertainty? Starting with the selection of training data and labelling—where are the images found, and how do you know how to label them? To know someone is a criminal, you will need police information. Which images do you select of them? And how do you know that someone that is *not convicted* of anything is not a criminal? Add gender, race, clothing style, etc. to the selection criteria, and things get complicated both in terms of data selection, fairness and ethics. In the case in question, the algorithm basically learned to separate people smiling from the ones not smiling, since they used photographs of criminals taken by the police and ordinary pictures of happy people on the non-criminal side [1].

Moving to the learning algorithm—what is the algorithm going to pay attention to in the pixels of the images of criminals and non-criminals? In supervised machine learning, it will need guidance—colour, contrast, background, angles between nose and lips, cheek bones, etc. In unsupervised learning, it will create these patterns itself. In any case, it will be a selection, and this will be a process worth considering in terms of understanding what it is doing and why.

Moving on to the test data, this is where things really get messy. This is when the programme is faced with data it has never seen before and tries to classify new observations based on what it has learned from the training data. In the example of recognising criminals, it will be given more data, with more variation and encounter far more problems. To predict whether a person is in fact a criminal, the system will need a model ranking the observations according to several scales and a line dividing the criminals from the non-criminals. This is, in fact, a form of generalisation, drawing lessons from experience to something it has never seen before. There are numerous examples showing that this is very challenging, even in presumably simple image recognition like the reading of handwriting on envelopes in the postal services [1]. A final point is the lack of transparency as to what the software actually looks like once it has run for a while and changed its criteria for classification—who is able to verify what the programme is doing, and which assumptions it has created for its own classifications?

The example is about image recognition and not safety–critical decision-making. What does this tell us about algorithms in safety–critical planning, management and operations? First of all, it provides a general warning that algorithms are never neutral. Algorithms will reflect the shortcomings in knowledge of their creators, and no matter how brilliant the test data, there is no way of escaping the uncertain assumptions about the world underlying their working. Whether these assumptions stem from the algorithm's creator or from its ability to change itself, they will still be uncertain.

But, there are even simpler questions to ask concerning data—both the training and test data. The most obvious question is whether there is *enough* data. In terms of monitoring the technical condition of equipment, there might be enough data to do this. This might replace manual human work by providing predictive recommendations for maintenance and replacement, as long as there is no need for professional and contextual judgement. But, these are routine tasks that are repeated over and over. If we are talking about the more complex tasks that human operators do, that have a more situational component of judgement, it is much harder to see where one would find the data able to match the choices made with the situational characteristics that make them meaningful.

Another question is how these data become available for analysis in the first place. A disproportionate share of attention and resources within computer science is devoted to analysis methods, treating the input data as pre-existing objects [11]. Data is rarely "discovered" as objective facts and analysed as such—it is both selected and prepared before it is available for analysis (ibid.). This is a more serious challenge than the well-known potential for "garbage in, garbage out". It points, again, to the invisible production of uncertainties, only that it provides both data and algorithms with a form of misplaced objectivity that only becomes visible after some kind of failure.

In safety–critical contexts, we cannot afford to overlook the fact that data will be biased, labelling contains bias, and that learning algorithms can create their own set of bias. Therefore, there is no reason to believe that AI removes human fallibility. It replaces one form of human fallibility with another.

5.5 The Human Contribution to Safety in Future

It follows from the discussion so far that the human contribution to safety will be a topic for safety science in future, although our study of it may require additions to our theoretical repertoire. One of the questions in the workshop from which this book arises was "what forms can the human contribution to safety take in future?" When reflecting on this question, I realised that I am not that worried that humans will be made obsolete by artificial intelligence anytime soon. While the automation of routine actions and decisions that has been ongoing since the industrial revolution is not likely to come to a halt, safety–critical sensemaking may be one of the last instances where humans could be replaced. And the reason for this is that our intelligence is *not* artificial. I will try to explain my argument by reflecting on what intelligence can be.

The concept of intelligence is contested and multifaceted. It has been used to refer to numerous behaviours linked to some form of performance, but where the performance cannot be separated from its specific contexts [12]. The point here is not to go into detail on the concept of intelligence, but it is worthwhile considering what we mean by the term when we discuss the relationship between human and artificial intelligence. This includes a consideration of the contexts where intelligence is regarded as key to successful performance, for which tasks which form of intelligence is beneficial and the difference between artificial and natural intelligence in this respect.

Specialised, narrow intelligence is the first form. This means mastering very specific tasks to perfection, but the capabilities are limited to that particular task. This is what Google Assistant does. It masters the translation from voice commands to tasks like an internet search or turning an electrical appliance on and off. This kind of AI is everywhere. It is impressive and cool, and it keeps improving. But, it can only apply intelligence to the specific problems for which it is programmed. All it can do is basically search the internet faster and more comprehensively than an individual can. Google Assistant and Siri are little more than natural language processing algorithms used together with predefined rules.

Physical intelligence is another category and is derived through physical and practiced learning. Sports, dancing or craftsmanship is common examples. When you think about it, these are complex skills because they link aspects of the mind with the motor skills of the body. Robots that are able to run on uneven terrain or ride a bicycle have been around for more than a decade, impressing the audience with their ability to perform (well, almost) human-like movements based on a triangulation of data across a number of sensors. But, my guess is that the development of these robots has taken years and millions of dollars, while your average 6-year-old child can learn it intuitively in a day or two. This gives the human contribution to safety a continued role in future—while our information processing capacity in narrow domains may be limited, our action repertoire in the physical world is not particularly limited, since we can improvise with whatever tools and information we have at hand.

The last category is general, common-sense intelligence. This is probably what we have in mind when we use the term intelligence—the ability to interpret and understand virtually any situation and learn how to act in that situation. This includes understanding cause and effect. AI can understand that two observations are correlated, but it is still more or less clueless when it comes to causality. A statistical model can establish a correlation between clouds and rain and make good predictions about the probability of rain, but it has no idea what a cloud is. Translated to the domain of safety, the model may be able to rank different decision and action alternatives according to the labelled severity of consequences, but it has no understanding about accidents or death. The same reasoning goes for a human's contextual understanding of a social situation and the ability to pick up subtle cues from its environment, even though it has never been in a similar situation before. This is one of the key human contributions to safety. Common-sense knowledge, along with professional judgement, is what make humans capable of sensemaking. Humans are able to make good guesses in rare situations, based on incomplete information. In the field of AI, common-sense intelligence has simply not been invented yet.

The point of these reflections is that when we say that AI has superior information processing capacity compared to humans, this is only partly right. It can apply some form of intelligence in the form of information processing on *narrowly defined areas, where there is sufficient data and the similar situations occur over and over again.* It can also beat humans at chess or Alpha Go, but this is based on the computer being faster and better at calculating and simulating a wide range of different game scenarios. And make no mistake, this is a fantastic skill. However, we are only comparing a small part of the intelligence repertoire and the one aspect where AI currently has an edge. It is somewhat of a paradox that comparisons between human and artificial intelligence seem to be based on the premises of what artificial intelligence is capable of, not what human intelligence is capable of.

As follows from the previous sections, the human contribution to safety is more than standardised information processing around narrowly defined tasks. In most other aspects of intelligence, artificial intelligence does not even come close. This is where the buzz around AI is ridiculously hyped and mystified. In its current stage of development, AI is software written to do specific tasks. It is not alive, it does not have a consciousness, and it is completely incapable of understanding, creativity and empathy [2]. This is important to stress, because both the promises and worries sometimes seem to be aimed at technology which has not yet been invented.

5.6 Implications—The Future of Risk

Summing up these considerations, there are several paradoxes involved in the use of algorithms and AI in safety–critical settings. These are paradoxes that will need some form of resolution before AI is put into use in safety–critical decision-making.

First is what we can call an intelligence paradox. It consists of two branches where the first is associated with the access to and selection of training data. Intelligence requires experience. If there is "garbage in, garbage out" in training data, it might be artificial, but it is not intelligence. The other branch has to do with the kind of intelligence needed in different circumstances in high-risk industries and the human contribution in this respect. As long as general AI simply does not exist, the drive towards reducing the human contribution in high-risk settings should not only be a question of *how*, but a question of *why.* Answering this question should not only focus on the narrow capabilities of artificial intelligence but the general capabilities of human intelligence.

Second is a transparency paradox. When algorithms learn, they become *actors* in safety management. It is a basic principle of HROs and most advice on safety management that important decisions and actions must be checked and double-checked. Assigning a critical task or decision to technology does not revoke this need. If it is hard to qualify technological actors' interpretations and assumptions, it is a breach of basic principles of redundancy and accountability.

Third is what I call a verification paradox. It is a variant of the transparency paradox, but the third one has more to do with the long-term follow up of learning algorithms. I think, or at least hope, that regulators and supervisory authorities will never allow critical decision-making processes to change unsupervised. The governance of self-learning algorithms requires regulation, audit tools and competence, something which is not in place, and it is hard to see how this can be done, at least within a prescriptive regulatory regime. The matter of verification also touches upon the larger problematics of bias, ethics and fairness of AI systems [8], which is also likely to be of great importance for safety science.

High-risk industries are moving into a landscape where software becomes more safety–critical than before, making software engineers more critical than before. It is not a wild guess that the patterns of failure will change, from traditional operator (human) errors to more software-related (human) errors. This includes the possibility for algorithmic surprise. If tests and simulation are by definition incomplete (remember Downer's argument), surprises will occur. Trial and error will persist as a prerequisite for learning, possibly creating some very unpleasant situations for both companies and regulators. As algorithms and AI enter the field of human factors, we should rethink the way we conceptualise and study the relationship between the human and technological agent of safety. The need for human factors expertise is probably more relevant in the age of artificial intelligence than it has ever been before.

Acknowledgement The chapter is published as part of the project Theoretical Advances of Cyber Resilience (TECNOCRACI), funded by the Research Council of Norway through grant no. 303489.

References

1. C.T. Bergstrom, J.D. West, *Calling Bullshit. The Art of Skepticism in a Data-Driven World* (Penguin Random House, New York, 2020)
2. P. Boucher, *How Artificial Intelligence Works* (European Parliament, 2019). Retrieved from https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/634420/EPRS_BRI(2019)634420_EN.pdf
3. S. Dasgupta, *Computer Science: A Very Short Introduction* (Oxford University Press, Oxford, 2016)
4. J. Downer, "737-cabriolet": the limits of knowledge and the sociology of inevitable failure. Am. J. Sociol. **117**(3), 725–762 (2011)
5. J. Downer, On ignorance and apocalypse: a brief introduction to "epistemic accidents", in *Safety Science Research: Evolution, Challenges and New Directions*, ed. by J.-C. Le Coze (CRC Press, Boca Raton, 2020)
6. J. Hoc, From human-machine interaction to human-machine cooperation. Ergonomics **43**(7), 833–843 (2000). https://doi.org/10.1080/001401300409044
7. S. Legg, M. Hutter, A collection of definitions of intelligence. Front. Artif. Intell. Appl. **157**, 17–24 (2007)
8. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A Survey on Bias and Fairness in Machine Learning (2019). Retrieved from ArXivabs/1908.09635

9. N. Oliver, T. Calvard, K. Potočnik, Cognition, technology, and organizational limits: lessons from the Air France 447 disaster. Organ. Sci. **28**(4), 729–743 (2017)
10. R. Parasuraman, T.B. Sheridan, A model for types and levels of human interaction with automation. Syst. Man Cybern. **30**(3), 286–297 (2000)
11. E. Parmiggiani, T. Østerlie, P. Almklov, In the backrooms of data science. J. Assoc. Inf. Syst. **23**(1), 139–164 (2022)
12. H.D. Schlinger, The myth of intelligence. Psychol. Rec. **53**(1), 15–32 (2003)
13. T.B. Sheridan, W.L. Verplanck, *Human and Computer control of Undersea Teleoperators* (MIT Man–Machine Systems Laboratory Report, MIT, Cambridge, MA, 1978)