Understanding risk in the context of AITs, commentary

1 Introduction

Risk can generally be defined as the likelihood of something adverse happening. How risk is presented and how it is determined can vary massively depending upon the application. With the creation of AITs, there is a need to consider both the risks associated with their uses and their ability to objectify risk, as well as how these risks change in different use cases.

2 Chen JH, Asch SM. Machine learning and prediction in medicine—beyond the peak of inflated expectations. The New England journal of medicine. 2017 Jun 29;376(26):2507.

Chen (2017) considers how while it has been suggested that "big data" will transform healthcare, with AITs automating large parts of it, it may not be the silver bullet people want especially when it comes to predicting risks. Chen (2017) suggests that while the use of risk is nothing new when making decisions, the use of AITs can increases the accuracy of the predicted risk which is then used to inform decisions as to what treatments are undertaken. Chen (2017) also suggests that the data used by AITs is not necessarily perfect, with both too little and too much data creating problems for the risk prediction, as the aim is to use data to predict the future instead of reproducing what has already happened; with Clinical data having an effective half-life of 4 months. Furthermore Chen (2017) identifies that even if the right amounts of data is collected this may not cover sufficient sources of information to make accurate predictions, especially as when trying to make predictions further in to the future, small differences in the environment have the potential to cause massive changes. While AITs may be able to accurately assess the levels of risk, they are not able to give the source or ways in which the risk can be reduced, for instance with post-surgery complications.

- Chen (2017) describes how AITs are used to establish the risk of something occurring, however the potential risks of using AITs in an environment such as medicine, where decisions could have potentially massive impacts, are not considered.

- Chen (2017) makes the point that the creation of an AIT needs to be able to take account of nuances in a person's life, however not only is it hard to predict the longer future, but also the more data that is included the more arbitrary some of these features can be, and they may not have any effect on the health of the individual (such as seen with predictive policing using unrelated features to predict crime).

- Although Chen (2017) considers how AITs in medicine can assess risk limitations, and how these predictions can improve medicine, he makes no mention of the risks associated with using such systems.

3 Green B, Chen Y. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. Proceedings of the ACM on Human-Computer Interaction. 2021 Oct 18;5(CSCW2):1-33.

Green (2021) considers how decision making is affected by the use of risk assessments that aid the quantification of risk by attempting to make it more subjective, by bringing attention to specific factors. To understand how risk assessments affect decisions, the distinction between predictions and decisions is made; while risk assessments may improve peoples' predictions, they do not necessarily improve their decision making ability, as seen in the two scenarios considered: with the use of risk assessments for pre-trial detention increasing biases and as well as making people more risk averse when approving government loans for home improvements. This led Green (2021) to conclude that while risk assessments may aid in the quantification of risk, their use can lead to adverse outcomes counteracting any potential benefits which may derive from a person making the decision, with no differences identified between the behaviors of lay people and experts in the fields considered. Green(2021) finally suggests that to adopt algorithms in this decision making process, a baseline is needed to show that they actually improve decision making, especially as relying on human oversight rarely functions as desired.

- Green(2021) identifies a limitation of his work as the use of Mechanical Turk workers not being a real world environment where other factors may also effect the decision making meaning there may be a real world difference between lay-people and experts.

- Green (2021) picks up on how the use of risk assessments attempts to make decisions more objective, however with the use of risk assessments, both questions as to their validity and their objectivity should be answered, especially if they amplify biases, and how to mitigate their negative effects.

- While Green (2021) makes the interesting point that risk assessments make us more risk averse, he makes no mention of why this happens, whether this is a bad thing, or ways in which to prevent this, especially as risk is often only one of many factors which can influence a decision

4 Ananny M. Toward an ethics of algorithms: Convening, observation, probability, and timeliness. Science, Technology, Human Values. 2016 Jan;41(1):93-117.

While not explicitly about risk, Ananny (2015) indirectly considers risk through the lens of ethical considerations arguing that ethics should be asking *When, How* and *For whom* AITs work. Ananny (2015) makes the point that AITs embody their designer's own ethical standards; this by extension, means that they take on the designer's stance towards risk and what their accepted levels of risk are. These accepted levels/ethical stances change over time due to their experiences as well as standards that are created. Ananny (2015) goes on to consider how when talking about the ethics of an algorithm, it may be better to talk about an algorithm in a specific scenario and use cases; giving the example of the Google news algorithm, which learns from the user over time causing the algorithm change and become specific to they user. This leads onto a question as to who is responsible when the outcomes are ethically questionable, as questionable outcomes can arise even if the code is transparent, designed with good intentions and in a well-regulated area of use (a prime example of this would be Microsoft's chatbot *"Tay"* which became radicalised in less than 24 hours[1]).

- While Ananny (2015) recognizes that the situation in which an AIT affects the ethical considerations which have to be made, he doesn't provide any way to guide these ethical considerations

- Although Ananny (2015) may have considered the ethics of AITs, and that they take on the ethical stance of their designers and that ethical standards may converge, an important point which he fails to make (especially as the EU is trying to regulate AITs based on their risk[2]) is that if we are attempting to regulate AITs based on ethics or risks, it is down to the designers/regulators to hold to their interpretation of the standard

5 Questions to consider:

- As AITs are unlikely to be able to give 100% accurate predictions of risk, or tell us the cause of the risk or how to mitigate it, should we be using them in safety critical situations? When things go wrong where should the blame lie/who should be held accountable?
- If more accurate quantification of risk through the use of risk assessments can lead to adverse outcomes, why might we want to accurately quantify risk? Is this even possible?
- When trying to control the design of AITs how much can the risks they pose be used to create credible regulation? When considering the risks that AITs pose, are all parts of the implementation process equally valid?

References

1. Suárez-Gonzalo S, Mas Manchón L, and Guerrero Solé F. Tay is you: The attribution of responsibility in the algorithmic culture. Observatorio (OBS*). 2019; 13 (2):[14] p. 2019
2. European Comission. Regulation of the European Parliament and of the Council - Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and amending Certain Union Legislative Acts. 2021; COM(2021) 206 final