Horizon Scanning Series

The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing

Trust

This input paper was prepared by Reeva Lederman

Suggested Citation

Lederman, R (2018). Trust. Input paper for the Horizon Scanning Project "The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing" on behalf of the Australian Council of Learned Academies, www.acola.org.

The views and opinions expressed in this report are those of the author and do not necessarily reflect the opinions of ACOLA.

Trust in AI

Notes by Reeva Lederman

The issue of trust in Artificial intelligence systems raises a number of definitional problems: Do we mean trust in the effectiveness of the technology and that the algorithms behind the system will produce the desired output; trust in the values underlying the system; trust in the way data in the system is protected and secured; trust that the system has been developed for the good of all stakeholders? These questions of trust take users far beyond the simple matter of whether they believe the technology works.

Trust is the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party (Mayer et al., 1995) . When we discuss trust in technology we have similar expectations: that we can give ourselves over to the technology and it will perform reliably in a predetermined way.

The problem of trust in technology or in automation is not a new problem (Lee and See, 2004), However, the complexity of AI has made a deep understanding of the technology more difficult for users and consequently raises additional issues of trust. However, the potential benefits of AI for health and well-being mean that issues of trust need to be explored and dealt with to ensure they do not create any unfounded barriers to use.

Artificial intelligence systems offer tremendous potential benefits in a diverse range of application areas from transportation, finance, security, legal practice, medicine and the military. Most of the systems under consideration in these fields involve what we call "weak AI" in that it assists in the performance of specific tasks that involve probabilistic reasoning, visual or contextual perception and can deal with complexity in ways that far outpace the human mind. AI systems are not yet able to deal with ethical judgements or the effective management of social situations or mimic all facets of human intelligence. Nonetheless, they still provide significant opportunities to increase our ability to make effective use of available data.

Examples of AI systems currently in use or under development include household systems which use available data to anticipate human needs. For example, AiCure (Hengstler et al., 2016) which reminds patients to take medication and confirms their compliance, or household robots that can fetch and deliver . Further health systems include applications that can, for example, potentially replace the work of radiologists by performing diagnoses (https://www.forbes.com/sites/paulhsieh/2017/04/30/ai-in-medicine-rise-of-the-machines/#25eebc2dabb0) or applications that form relationships with mental health patients that simulate some of the features of the human psychologist (D'Alfonso et al., 2017).

In transportation, car manufacturers are well on the way to developing autonomous and semi-autonomous vehicles. BMW already has a semi -autonomous vehicle on the market, Daimler has a fully autonomous truck planned for 2025, the city of Nurnburg in Germany has operated a fully autonomous train since 2008.

In the military the potential for AI is huge. Scenarios where lone mission commanders direct unmanned military vessels controlled by AI have the potential to significantly reduce loss of life in combat. In the US, the Defense Advanced Research Project Agency (DARPA) are working on ways to use AI to extract military information from visual media captured in the field and turn available photos and videos into useable sources of intelligence.

The areas described above, health, transportation and military services, are central to our safety and wellbeing, as are many of the other areas in which AI is applied. Consequently, trust has become an important issue in the acceptance and adoption of such systems as human beings are protective over these areas of their lives and are reluctant to cede control to automatous devices.

While trust has traditionally been a concept used to describe human to human interactions, many previous studies have shown that it is valid to use the concept of trust to describe the way in which the relationship between humans and computers or automation is mediated (Zuboff, 1984). Trust in what were previously human led processes (where trust was previously not guaranteed) needs to somehow be extended to a new environment where the same processes are now automated. Trust is also difficult to achieve where complex algorithms are being implemented and a full understanding of the technology (Lee and See,2003) is difficult to achieve. Lack of trust in automating technologies, which can include AI, is shown to lead to misuse or disuse which can compromise safety or profitability (Lee and See, 2004).

Research suggests that trust in AI depends on several factors. Firstly the technology needs to have proven reliability. " A technology based on the delegation of control will not be trusted if it is flawed" (Hengstler et al., 2016). In AI applications, useability, reliability and consistent operation all engender trust (Siau and Wang, 2017). Previous work has shown that users of automation consider three factors as important in trust. 1. Performance (what the technology does) including specifically operational safety and maintenance of data security (Hengstler et al., 2016, Lee and See, 2004).. 2. Process, including useability and whether or not it can be trialled (Lee and See, 2004). 3. Purpose, or why the technology was developed and whether it benefits the consumer (Hengstler, 2016) and is visible (such as the Automated train (Rogers, 2003) and finally , 4. the design. Designs that humanise technologies are more trustworthy so very robotic designs need to make some of the other qualities more obvious and also make the users feel as though they have a significant level of control (Hengstler, 2016). .

Users are also found to experience greater feelings of trust if the innovating firm is known to them (Hengstler, 2016). Consequently, for firms like BMW or Daimler developing automaoted cars, positive brand identification is important but firms also need to build relationships with consumers through information provision and the involvement of users in project development. This issue highlights the difference in two concepts of trust in AI – whether the trust is in the technology or the technology provider (Siau and Wang, 2017). Both of these forms of trust are important to whether or not users are willing to interact with AI.

A further important factor in trust is the notion of explainability, where the actions of the AI are easily understood by humans. AI is being used by systems to arrive at important decisions in the lives of individuals, such as admission to education or provision of finance. Increasingly consumers are calling for the right to an explanation in decisions made by AI, but legal frameworks are yet to respond adequately (Edwards and Veale, 2017).

Previous work suggests that people seek explanations of AI when cases are contrastive (ie they wonder why one thing happened and not another); people use their cognitive biases to selective explanations for how AI performs; people are not always swayed by the most likely explanation for how AI has behaved unless they understand the cause of the most likely explanation; explanations for

AI are social and are influenced by a person’s beliefs (Miller, 2017). Trust in AI can be seen to be dependent on how much developers respond to these problems of explainability.

D'ALFONSO, S., SANTESTEBAN-ECHARRI, O., RICE, S., WADLEY, G., LEDERMAN, R., MILES, C., GLEESON, J. & ALVAREZ-JIMENEZ, M. 2017. Artificial-Intelligence-Assited Online Social Therapy for Youth Mental Health. *Frontiers in Psychology,* 8**,** 13 pages.

EDWARDS, L. & VEALE, M. 2017. Slave to the Algorithm. Why a "right to an explanation" is probably not the remedy you were looking for. *Duke Law and Technology Review,* 16**,** 18-84.

HENGSTLER, M., ENKEL, E. & DUELLI, S. 2016. Applied Artificial Intelligence and Trust - The case of autonomous vehicles and medicla assistance devices. *Technological Forecasting and Social Change***,** 15.

LEE, J. D. & SEE, K. A. 2004. Trust in Automation: Designing for Approparite Reliance *Human Factors,* 46**,** 50-80.

MAYER, R. C., DAVIS, J. H. & SCHOORMAN, F. D. 1995. An integrative model of Organisational Trust. *Acad. Manag. Review,* 20**,** 25.

MILLER, T. 2017. Explanation in Artificial Intelligence: Insights from the Social Sciences. *eprint arXiv:1706.07269*.

SIAU, K. & WANG, W. 2017. Building Trust in Artificial Intelligence , Machine Learning, and Robotics. *The Cutter Edge,* 31.

ZUBOFF, S. 1984. *In the Age of The Smart Machine*, BasicBooks.