Proposal AOM Specialised Conference

Big Data and Managing in a Digital Economy

Track 1: The Digital Organization and Digital Organizing

How to build responsible AI? Lessons for governance from a conversation with Tay

Mark van Rijmenam

University of Technology Sydney, Australia

Mark.vanRijmenam@student.uts.edu.au (corresponding author)

Jochen Schweitzer

University of Technology Sydney, Australia

Jochen.Schweitzer@uts.edu.au

Abstract

Artificial Intelligence (AI) is intelligence displayed by a machine observing its surrounding and acting to maximise its ultimate goal. The recent developments of AI have progressed to such an extent that many scholars now stress the importance of creating AI that is safe and responsible. This is important as organizations continue to develop algorithms that turn out harmful for those involved and affected by AI. We review the example of Microsoft's chatbot *Tay*, which posted inflammatory, offensive and racist tweets in 2016. The example demonstrates how AI is not without risks. To help avoid risks, we use a qualitative approach to study organisations that developed and implemented a chatbot. Our objective is to further understand both technical and organisational governance practices that ensure responsible AI.

Keywords: Artificial Intelligence, Governance, Ethics, Risks

Introduction

On March 23, 2016, Microsoft Corporation launched Tay, an AI chatbot on Twitter, with the objective of having conversations with Twitter users, learn from these interactions and improve itself. However, after 16 hours and 96.000 tweets, Tay was taken offline since it had started to post inflammatory, offensive and racist tweets, resulting in a public relations disaster for Microsoft (Hern 2016, Vincent 2016, Murgia 2016). The example of Tay showed that AI released in the real world, may behave different to when used in isolated controlled environments (Garcia 2016). The case of Tay also showed that biased data can have a huge (negative) impact on anticipated outcomes (O'Neil 2016, Spielkamp 2017). What became clear is that while AI is rapidly impacting all aspects of society and starting to make up for its promise (Yudkowsky 2008), it does not come without risks (Luca, Kleinberg, and Mullainathan 2016).

AI is intelligence presented by machines that perceive its environment and take action accordingly to maximise its ultimate goal (Bostrom 2014, Russell and Norvig 1995). In recent years, there has been a continuous stream of breakthroughs of AI, made possible through increased "computational capabilities, algorithm design and communication technology" (Alfonseca et al. 2016, 1), resulting in algorithmic businesses that rely on complex algorithms and Artificial Intelligence to automate business processes and improve decision-making (Prentice 2016). It is seen as a natural evolution of any business transformation to digital, whereby new value can be created thanks to AI (Prentice 2016). As such, algorithms are likely to take over jobs and result in significant loss of jobs across the globe (Ford 2015, Berriman 2017, Furman 2016). One place where this is quite visible is the area of call centres, where conversational artificial intelligence, also known as chatbots, are taking over much of the work previously performed by call centre agents (Accenture 2017). However, the challenges and potential negative implications of artificial intelligence have also become quite visible, with the debacle of the chatbot Tay as a warning. In fact, there are increasingly concerns on the risks that artificial intelligence may bring humanity in the coming decades (Müller and Bostrom 2016).

There are three forms of Artificial Intelligence; Narrow AI, Artificial General Intelligence (AGI) and Super Artificial Intelligence (SAI) (Bostrom 2014). Narrow AI refers to artificial intelligence that is more intelligent than humans on specific tasks in relatively narrow domains (Baum, Goertzel, and Goertzel 2011), e.g. a trading algorithm or Siri on your iPhone. Artificial General Intelligence refers to AI systems having autonomous self-control and self-understanding and the ability to learn new things to solve a wide variety of problems in different contexts (Goertzel and Pennachin 2007), e.g. Siri capable of driving your car, doing your accountancy and making you a coffee. The final phase of intelligence is Super Artificial Intelligence, which means intelligence that far exceeds the intelligence of any man however clever (Bostrom 2014, Good 1966), in continuation of the Siri metaphor, it means Siri ruling the world. SAI will result in new forms of intelligence unfamiliar to mankind today (Armstrong, Sandberg, and Bostrom 2012, Bostrom 2014), and offers different risks than any other known existential risk humans faced before and hence requires a fundamentally different approach. However, until now little research has been done to understand what this approach should be and how organisations can contribute to safe AI.

The dangers of AI

Discussions of whether AI is as a catastrophic risk to mankind has recently also reached the academic world, where the “discourse about the existential risk related to AI” (Alfonseca et al. 2016, 1) has attracted many scholars to investigate the subject and conclude the importance of developing safe AI (Russell, Dewey, and Tegmark 2015, Anderson 2015). Even if the chances for it to come true might be extremely small, these risks should be taken serious because of what is at stake (Barrett and Baum 2016) and there have been already plenty examples of AI gone wrong (O'Neil 2016) and inflict damage (Luca, Kleinberg, and Mullainathan 2016). In fact, Moor (2005) argues that ethical problems will increase as technologies improve and develop into mutually enabling technologies, as is the case with artificial intelligence that uses reinforcement learning to improve itself (Bostrom 2014). In such cases, governance becomes increasingly important to ensure that high-quality, non-biased data (O'Neil 2016) is used and algorithms are developed in line with existing, as well as future, norms, values and ethics (Anderson, Anderson, and Armen 2004, Moor 2005, Reynolds 2011, Anderson and Anderson 2011). As the Tay chatbot example shows, artificial intelligence and machine learning rely on input data that trains the algorithm. As a consequence, it is not immune from the ‘garbage in, garbage out’ rule (Musib et al. 2017). Therefore, there is a moral obligation to ensure that artificial intelligence remains safe and friendly, now and in the future, which is why Mayer-Schönberger and Cukier (2013) argue for the need of ‘algorithmists’ who review algorithms as existing accountants review the financials of an organisation to ensure that AI performs correctly and is not affected due to low-quality, biased, input data.

The importance of governance

Although data governance has grown in importance because of artificial intelligence (Alhassan, Sammon, and Daly 2016, Malik 2013), many organisations don't see data as an asset (Tallon 2013) and hence lack a clear understanding of the importance of high-quality, unbiased, input data (O'Neil 2016). As a result, companies face a variety of problems that could harm the business (Cleven and Wortmann 2010), as we saw with the debacle of Tay. Wende (2007) argues, therefore, that organisations need the right governance processes that ensure high-quality data, resulting in a *single version of the truth,* to cope with the strategic and operational challenges of their environment (Watson, Fuller, and Ariyachandra 2004, Khatri and Brown 2010) and to prevent biases when dealing with artificial intelligence (O'Neil 2016). Other scholars, on the other hand, argue for the need of IT ethics, which refers to the ethical behaviour of information systems and their employees (Banerjee, Cronan, and Jones 1998, Tavani 2003, Reynolds 2011). With the rise of intelligent machines, this is gradually moving towards machine ethics, focused on how algorithms and machines can behave ethically now and in the future (Satell 2016, Anderson and Anderson 2011).

Human-Machine networks

The development of AI results in an increasing convergence of the human and the computer, resulting in social, technological, political and ethical implications where artificial intelligence and humans are becoming increasingly interwoven in mutually dependent networks (Fleischmann 2009). This is especially visible in areas where chatbots are implemented, as humans directly interact with artificially intelligence (Zamora 2017). Understanding the success or failure of Artificial Intelligence and subsequent interactions with different actors can be achieved by Actor-Network Theory (Tatnall 2005). As Tatnall (2005) argues, ANT helps to explain how agents with specific (artificial) characteristics interact with each other and allows an analysis of both artificial and non-artificial agents in the same context, avoiding the need to think in human/non-human barriers and ignoring hierarchical distribution of actors (Latour 2005). Such a flat ontology, where actors of different size and type are considered equally capable of creating interactions with each other, is especially relevant in considering the impact of (Super) Artificial Intelligence on organisations and humans across time and space. Since ANT assumes infinite pliability and freedom of actors, it helps to understand how actors make a certain network successful

(Latour 2005, Tsvetkova et al. 2015, Walsham 1997, Murdoch 1998)

Research Question

In the past decades, IT ethics and governance have played an increasingly important role and many researchers have focused on these topics (Reynolds 2011, Tavani 2003, Sarsfield 2009, Bruwer and Rudman 2015). Nonetheless, little research has been done on how organisational structures and processes could ensure that Artificial Intelligence will not have flaws or do no harm to actors involved, especially when AI evolves to Artificial General Intelligence (AGI) or even Super Artificial Intelligence (SAI). Therefore, we seek to develop a framework that enables organisations to understand what is required to ensure safe, or responsible, AI, i.e. AI that does what we want it to do and does not harm any actors involved. We aim to answer the research question *How can organisations ensure responsible AI and prevent AI from harming those actors involved?* We propose to answer this question via a qualitative study among Australian organisations that have developed and implemented artificial intelligence in the form of a chatbot. Our research will focus on understanding what technical as well as organisational governance practices organisations have taken to ensure responsible chatbots.

In the paper, we will discuss the theoretical background related to various forms of Artificial Intelligence, the role governance can play within organisations to ensure responsible AI and we discuss the theoretical lens of Actor Network Theory that we will apply to answer our research question. This will lead us to our proposed conceptual framework, which we aim to validate in the qualitative study. Finally, we will end with a discussion on the results, our limitations and an agenda for further research in this continuously developing research field.

References

Accenture. 2017. At Your Service - Embracing the Disruptive Power of Chatbots.

Alfonseca, Manuel, Manuel Cebrian, Antonio Fernandez Anta, Lorenzo Coviello, Andres Abeliuk, and Iyad Rahwan. 2016. "Superintelligence cannot be contained: Lessons from Computability Theory." *arXiv preprint arXiv:1607.00913*.

Alhassan, Ibrahim, David Sammon, and Mary Daly. 2016. "Data governance activities: an analysis of the literature." *Journal of Decision Systems* 25 (sup1):64-75.

Anderson, Jon. 2015. "The Doomsday Invention." *The New Yorker*.

Anderson, Michael, and Susan Leigh Anderson. 2011. *Machine ethics*: Cambridge University Press.

Anderson, Michael, Susan Leigh Anderson, and Chris Armen. 2004. "Towards machine ethics." Proceedings of the AOTP'04-The AAAI-04 Workshop on Agent Organizations: Theory and Practice.

Armstrong, Stuart, Anders Sandberg, and Nick Bostrom. 2012. "Thinking inside the box: Controlling and using an oracle ai." *Minds and Machines* 22 (4):299-324.

Banerjee, Debasish, Timothy Paul Cronan, and Thomas W Jones. 1998. "Modeling IT ethics: A study in situational ethics." *Mis Quarterly*:31-60.

Barrett, Anthony M., and Seth D. Baum. 2016. "A model of pathways to artificial superintelligence catastrophe for risk and decision analysis." *Journal of Experimental & Theoretical Artificial Intelligence*:1-18. doi: 10.1080/0952813X.2016.1186228.

Baum, Seth D, Ben Goertzel, and Ted G Goertzel. 2011. "How long until human-level AI? Results from an expert assessment." *Technological Forecasting and Social Change* 78 (1):185-195.

Berriman, Richard; Hawksworth, John. 2017. Will robots steal our jobs? The potential impact of automation on the UK and other major economies. PWC.

Bostrom, Nick. 2014. *Superintelligence: Paths, dangers, strategies*: OUP Oxford.

Bruwer, Rikus, and Riaan Rudman. 2015. "Web 3.0: Governance, Risks and Safeguards." *Journal of Applied Business Research* 31 (3):1037-n/a. doi: 10.1016/j.websem.2007.09.005. URL http: //linkinghub.elsevier.com/retrieve/pii/S1570826807000376.

Cleven, Anne, and Felix Wortmann. 2010. "Uncovering four strategies to approach master data management." System Sciences (HICSS), 2010 43rd Hawaii International Conference on.

Fleischmann, Kenneth R. 2009. "Sociotechnical interaction and cyborg–cyborg interaction: Transforming the scale and convergence of HCI." *The Information Society* 25 (4):227-235.

Ford, Martin. 2015. *The Rise of the Robots: technology and the threat of mass unemployment*: Oneworld Publications.

Furman, Jason; Holdren, John; Muñoz, Cecilia; Smith, Megan; Zients, Jeffrey. 2016. Artificial Intelligence, Automation and the Economy.

Garcia, Megan. 2016. "Racist in the Machine The Disturbing Implications of Algorithmic Bias." *World Policy Journal* 33 (4):111-117.

Goertzel, Ben, and Cassio Pennachin. 2007. *Artificial general intelligence*. Vol. 2: Springer.

Good, Irving John. 1966. "Speculations concerning the first ultraintelligent machine." *Advances in computers* 6:31-88.

Hern, Alex. 2016. "Microsoft scrambles to limit PR damage over abusive AI bot Tay." The Guardian, Last Modified 2016-03-24, accessed June 19. http://www.theguardian.com/technology/2016/mar/24/microsoft-scrambles-limit-pr-damage-over-abusive-ai-bot-tay.

Khatri, Vijay, and Carol V Brown. 2010. "Designing data governance." *Communications of the ACM* 53 (1):148-152.

Latour, Bruno. 2005. *Reassembling the social: An introduction to actor-network-theory*: Oxford university press.

Luca, Michael, J. O. N. Kleinberg, and Sendhil Mullainathan. 2016. "Algorithms Need Managers, Too." *Harvard Business Review* 94 (1):96-101.

Malik, Piyush. 2013. "Governing big data: principles and practices." *IBM Journal of Research and Development* 57 (3/4):1: 1-1: 13.

Mayer-Schönberger, Viktor, and Kenneth Cukier. 2013. *Big data: A revolution that will transform how we live, work, and think*: Houghton Mifflin Harcourt.

Moor, James H. 2005. "Why we need better ethics for emerging technologies." *Ethics and information technology* 7 (3):111-119.

Müller, Vincent C, and Nick Bostrom. 2016. "Future progress in artificial intelligence: A survey of expert opinion." In *Fundamental issues of artificial intelligence*, 553-570. Springer.

Murdoch, Jonathan. 1998. "The spaces of actor-network theory." *Geoforum* 29 (4):357-374.

Murgia, Madhumita. 2016. "Microsoft's racist bot shows we must teach AI to play nice and police themselves." The Telegraph, accessed June 19. http://www.telegraph.co.uk/technology/2016/03/25/we-must-teach-ai-machines-to-play-nice-and-police-themselves/.

Musib, Mrinal, Feng Wang, Michael A Tarselli, Rachel Yoho, Kun-Hsing Yu, Rigoberto Medina Andrés, Noah F Greenwald, Xubin Pan, Chien-Hsiu Lee, and Jian Zhang. 2017. "Artificial intelligence in research." *Science* 357 (6346):28-30.

O'Neil, Cathy. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*: Crown Publishing Group (NY).

Prentice, Stephen. 2016. "Defining Algorithmic Business." *Gartner*.

Reynolds, George. 2011. *Ethics in information technology*: Cengage learning.

Russell, Stuart, Daniel Dewey, and Max Tegmark. 2015. "Research priorities for robust and beneficial artificial intelligence." *AI Magazine* 36 (4):105-114.

Russell, Stuart, and Peter Norvig. 1995. "Artificial intelligence: a modern approach."

Sarsfield, Steve. 2009. *The data governance imperative*: IT Governance Publishing.

Satell, Greg. 2016. "Teaching an Algorithm to Understand Right and Wrong." [Website]. Harvard Business Review, Last Modified November 16, 2016, accessed February 7. https://hbr.org/2016/11/teaching-an-algorithm-to-understand-right-and-wrong.

Spielkamp, Matthias. 2017. "Inspecting Algorithms for Bias." MIT Technology Review, accessed June 23. https://www.technologyreview.com/s/607955/inspecting-algorithms-for-bias/.

Tallon, Paul P. 2013. "Corporate governance of big data: Perspectives on value, risk, and cost." *Computer* 46 (6):32-38.

Tatnall, Arthur. 2005. "Actor-network theory in information systems research." In *Encyclopedia of Information Science and Technology, First Edition*, 42-46. IGI Global.

Tavani, Herman T. 2003. "Ethics and technology: Ethical issues in an age of information and communication technology."

Tsvetkova, Milena, Taha Yasseri, Eric T Meyer, J Brian Pickering, Vegard Engen, Paul Walland, Marika Lüders, Asbjørn Følstad, and George Bravos. 2015. "Understanding Human-Machine Networks: A Cross-Disciplinary Survey." *arXiv preprint arXiv:1511.05324*.

Vincent, James. 2016. "Twitter taught Microsoft's friendly AI chatbot to be a racist asshole in less than a day." The Verge, accessed June 19. https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.

Walsham, Geoff. 1997. "Actor-network theory and IS research: current status and future prospects." In *Information systems and qualitative research*, 466-480. Springer.

Watson, Hugh J, Celia Fuller, and Thilini Ariyachandra. 2004. "Data warehouse governance: best practices at Blue Cross and Blue Shield of North Carolina." *Decision Support Systems* 38 (3):435-450.

Wende, Kristin. 2007. "A Model for Data Governance-Organising Accountabilities for Data Quality Management."

Yudkowsky, Eliezer. 2008. "Artificial intelligence as a positive and negative factor in global risk." *Global catastrophic risks* 1:303.

Zamora, Jennifer. 2017. "Rise of the Chatbots: Finding A Place for Artificial Intelligence in India and US." Proceedings of the 22nd International Conference on Intelligent User Interfaces Companion.