Meaningful human control over opaque machines

Robbins, S.A.

Publication date
2020

Document Version
Final published version

Published in
18 International Conference on the Ethical and Social Impacts of ICT

Citation (APA)
Robbins, S. A. (2020). Meaningful human control over opaque machines. In J. Pelegrin-Borondo, M. Arias-Oliva, K. Murata, & M. L. Palma (Eds.), *18 International Conference on the Ethical and Social Impacts of ICT: Paradigm Shifts in ICT Ethics: Societal Challenges in the Smart Society* (pp. 354-356)

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

MEANINGFUL HUMAN CONTROL OVER OPAQUE MACHINES

Scott Robbins

Technical University of Delft (Netherlands)

s.a.robbins@tudelft.nl

EXTENDED ABSTRACT

In an increasingly autonomous world, it is becoming clear that one thing we cannot delegate to machines is moral accountability. Machines cannot be held morally accountable for their actions (Bryson, 2010; Johnson, 2006; van Wynsberghe & Robbins, 2018). This becomes problematic when a machine makes a decision that has a significant impact on human beings. Examples of such machines which have caused such impact are widespread and include machines evaluating loan applications, machines evaluating criminals for sentencing, autonomous weapon systems, driverless cars, digital assistants, etc. The question that governments, NGOs, academics, and the general public are asking themselves is: how do we keep meaningful human control (MHC) over these machines?

The literature thus far details what features the machine or the context must have in order for MHC to be realized. Should humans be in the loop or on the loop? Should we force machines to be explainable? Lastly, should we endow machines with moral reasoning capabilities? (Ekelhof, 2019; Floridi et al., 2018; Robbins, 2019b, 2019a; Santoni de Sio & van den Hoven, 2018; Wendall Wallach & Allen, 2010; Wendell Wallach, 2007). Rather than look to the machine itself or what part humans have to play in the context, I argue here that we should shine the spotlight on the decisions that machines are being delegated. Meaningful human control, then, will be about controlling what decisions get made by machines.

This proposal, of course, simply kicks the can down the road and forces us to ask how to carve up the decision space in such a way that we can ensure meaningful human control. I propose here that machines currently make three types of decisions: descriptive, thick evaluative, and thin evaluative (Väyrynen, 2019; Williams, 2012). For example, an image classification algorithm could classify the image (or items within the image) in these three types. Descriptively, the algorithm could decide that the image is of a 'man' and a 'black bag' and that the image was taken 'inside'. The algorithm could also classify the man as 'dangerous' and 'suspicious'. These would be thick evaluative decisions. Finally, the algorithm could also classify the man as 'bad' which is a thin evaluative description.

I argue that keeping meaningful human control over machines (especially AI which relies on opaque methods) means restricting machines to descriptive decisions. It must always be a human being deciding how to employ evaluative terms as these terms not only refer to specific states of affairs but also say something about how the world ought to be. Machines which are able to make decisions based on opaque considerations should not be telling humans how the world ought to be. This is a breakdown of human control in the most severe way. Not only would we be losing control over specific decisions in specific contexts, but we would be losing control over what descriptive content grounds evaluative classifications.

Restricting machines to making decisions about the descriptive would allow humans to keep control over what is meaningful: value. This can best be seen when looking at thick evaluative decisions like classifying a person as 'suspicious'. 'Suspicious' is a 'thick' evaluative term because it includes both descriptive elements and evaluative elements. If I called someone suspicious it might include the

description 'solitary person with ski mask on loitering and biting their fingernails'. It would also include the evaluative element 'bad'. It is important that if I describe someone as suspicious that there is good reason to do so as it has both short term and long term consequences which could result in harm. For example, if a white neighbourhood consistently calls the police because there is a 'suspicious' person around then there is a good chance that both: that person will be forced to have an interaction with the police AND it will be signalled that the look and behaviour of that person is unwanted in that neighbourhood. If the only reason that the person was labelled suspicious is because that person was black, then unjustified harm has been done. Those people wielding such labels should be held accountable for their unreasonable use.

Allowing machines to make such thick evaluative decisions means delegating to machines the reasons that lead to a negative or positive evaluation. Examples of this happening in a harmful way are plentiful (e.g. Denying women jobs (Dastin, 2018)). For true meaningful human control, humans should rely on the aid of machines to make descriptive decisions that can, if needed, be verified. Instead of labelling a person as 'suspicious', a machine should label a person as 'loitering' and 'solitary' and then allow a human being to reach the evaluative conclusion that the person is suspicious. This leaves a human being in meaningful control over what is important thereby keeping clear human accountability for important decisions.

The further upshot of this proposal for meaningful human control is that it is in line with the idea that humans and machines should work together rather than machines replacing humans (see e.g. Rosenfeld, Agmon, Maksimov, & Kraus, 2017). The most important part of any collaboration is understanding what your, and your partner's, strengths are.

KEYWORDS: AI Ethics; Meaningful Human Control; Artificial Intelligence; Hybrid Intelligence.

REFERENCES

Bryson, J. (2010). Robots Should Be Slaves. In Y. Wilks (Ed.), Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues (pp. 63–74). Amsterdam: John Benjamins Publishing.

Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Ekelhof, M. (2019, March 19). Moving Beyond Semantics on Autonomous Weapons: Meaningful Human Control in Operation. https://doi.org/10.1111/1758-5899.12665

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. Ethics and Information Technology, 8(4), 195–204. https://doi.org/10.1007/s10676-006-9111-5

Robbins, S. (2019a). A Misdirected Principle with a Catch: Explicability for AI. Minds and Machines. https://doi.org/10.1007/s11023-019-09509-3

Robbins, S. (2019b). AI and the path to envelopment: Knowledge as a first step towards the responsible regulation and use of AI-powered machines. AI & SOCIETY. https://doi.org/10.1007/s00146-019-00891-1

Rosenfeld, A., Agmon, N., Maksimov, O., & Kraus, S. (2017). Intelligent agent supporting human–multi-robot team collaboration. Artificial Intelligence, 252, 211–231. https://doi.org/10.1016/j.artint.2017.08.005

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful Human Control over Autonomous Systems: A Philosophical Account. Frontiers in Robotics and AI, 5. https://doi.org/10.3389/frobt.2018.00015

van Wynsberghe, A., & Robbins, S. (2018). Critiquing the Reasons for Making Artificial Moral Agents. Science and Engineering Ethics, 1–17. https://doi.org/10.1007/s11948-018-0030-8

Väyrynen, P. (2019). Thick Ethical Concepts. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Summer 2019). Retrieved from https://plato.stanford.edu/archives/sum2019/entrieshick-ethical-concepts/

Wallach, Wendall, & Allen, C. (2010). Moral Machines: Teaching Robots Right from Wrong (1 edition). New York: Oxford University Press.

Wallach, Wendell. (2007). Implementing moral decision making faculties in computers and robots. AI & SOCIETY, 22(4), 463–475. https://doi.org/10.1007/s00146-007-0093-6

Williams, B. (2012). Ethics and the Limits of Philosophy (1 edition). London New York: Routledge.