Published in Behavioral and Brain Sciences
https://doi.org/10.1017/S0140525X17000255

Human-like machines: Transparency and comprehensibility

Piotr M. Patrzyk, Daniela Link, Julian N. Marewski

ADDRESS
Department of Organizational Behavior
Faculty of Business and Economics
University of Lausanne
Quartier UNIL-Dorigny, Internef
CH-1015 Lausanne
Switzerland

Emails:
piotr.patrzyk@unil.ch (Piotr M. Patrzyk)
daniela.link@unil.ch (Daniela Link)
julian.marewski@unil.ch (Julian N. Marewski)

Abstract: AI algorithms seek inspiration from human cognitive systems in areas where humans outperform machines. But on what level should algorithms try to approximate human cognition? We argue that human-like machines should be designed to make decisions in transparent and comprehensible ways, which can be achieved by accurately mirroring human cognitive processes.

How to build human-like machines? We agree with the authors' assertion that "reverse engineering human intelligence can usefully inform AI and machine learning" (p. 6) and in this commentary we offer some suggestions concerning the direction of future developments. Specifically, we posit that human-like machines should not only be built to match humans in performance, but also to be able to make decisions that are both *transparent* and *comprehensible* to humans.

First, we argue that human-like machines need to decide and act in transparent ways, such that humans can readily understand how their decisions are made (see Arnold & Scheutz, 2016; Indurkhya & Misztal-Radecka, 2016; Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016). Behavior of artificial agents should be predictable and people interacting with them ought to be in a position that allows them to intuitively grasp how those machines decide and act the way they do (Malle & Scheutz, 2014). This poses a unique challenge for designing algorithms.

In current neural networks there is typically no intuitive explanation for *why* a network reached a particular decision given received inputs (Burrell, 2016). Such networks represent statistical pattern recognition approaches that lack the ability to capture agent-specific information. Lake et al. acknowledge this problem and call for structured cognitive representations, which are required for classifying social situations. Specifically, the authors' proposal of an "intuitive psychology" is grounded in the *naïve utility calculus* framework (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016). According to this argument, algorithms should attempt to build a causal understanding of observed situations by creating representations of agents who seek rewards and avoid costs in a rational way.

Putting aside extreme examples (e.g., killer robots and autonomous vehicles), let us look at the more casual AI task of scene understanding. Cost-benefit based inferences about situations such as the one depicted in the left-most picture in Figure 6 in Lake et al. (p. 39) will likely conclude that one agent has a desire to kill the other and that he or she values higher the state of the other being dead than alive. While we do not argue this is necessarily wrong, a human-like classification of such a scene would likely lead to the conclusion that the scene depicts either a legal execution or a murder. The returned alternative depends on the viewer's inferences about agent-specific characteristics. Making such inferences requires going beyond the attribution of simple goals – one needs to make assumptions about the roles and obligations of different agents. In the discussed example, although both a sheriff and a contract killer would have the same goal to end another person's life, the difference in their identity would change the human interpretation in a significant way.

We welcome the applicability of naïve utility calculus for inferring simple information concerning agent-specific variables, such as goals or competence level. At the same time, however, we would like to point out some caveats inherent to this approach. Humans interacting with the system will likely expect a justification of why it has picked one

interpretation rather than another and algorithm designers might want to take this into consideration.

This leads us to our second point. Models of cognition can come in at least two flavors: (1) *as-if models*, which only aspire to achieve human-like performance on a specific task (e.g., classifying images), and (2) *process models*, which seek to both achieve human-like performance and to accurately reproduce the cognitive operations humans actually perform (classifying images by combining pieces of information in a way humans do). We believe that the task of creating human-like machines ought to be grounded in existing process models of cognition. Indeed, investigating human information processing is helpful for ensuring that generated decisions are comprehensible (i.e., that they follow human reasoning patterns).

Why is it important that machine decision mechanisms, in addition to being transparent, actually mirror human cognitive processes in a comprehensible way? In the social world, people often judge agents not only according to the agents' final decisions, but also according to the process by which they have arrived at these (e.g., Hoffman, Yoeli, & Nowak, 2015). It has been argued that the process of human decision-making does not typically involve rational utility maximization (e.g., Hertwig & Herzog, 2009). This in turn influences how we expect other people to make decisions (Bennis, Medin, & Bartels, 2010). To the extent that one cares about the social applications of algorithms and their interactions with people, considerations about transparency and comprehensibility of decisions become critical.

While as-if models relying on cost-benefit analysis might be reasonably transparent and comprehensible, for instance, when problems are simple and do not involve moral considerations, this might not always be the case. Algorithm designers need to ensure that the underlying process will be acceptable to the human observer. What research can be drawn up to help build transparent and comprehensible mechanisms?

We argue that one source of inspiration might be the research on *fast-and-frugal heuristics* (Gigerenzer & Gaissmaier, 2011). Simple strategies such as fast-and-frugal trees (e.g., Hafenbrädl, Waeger, Marewski, & Gigerenzer, 2016) might be well-suited for providing justifications for decisions made in social situations. Heuristics are not only meant to capture *ecologically rational* human decision mechanisms (see Todd & Gigerenzer, 2007), but are also transparent and comprehensible. Indeed, these heuristics possess a clear structure composed of simple if-then rules specifying (1) how information is searched within the search space, (2) when information search is stopped, and (3) how the final decision is made based upon the information acquired (Gigerenzer & Gaissmaier, 2011).

These simple decision rules have been used to model and aid human decisions in numerous tasks with possible moral implications, for example, in medical diagnosis (Hafenbrädl et al., 2016) or classification of oncoming traffic at military checkpoints as hostile or friendly (Keller & Katsikopoulos, 2016). We propose that the same heuristic principles might be useful to engineer autonomous agents that behave in a human-like way.

References

Arnold, T., & Scheutz, M. (2016). Against the moral Turing test: Accountable design and the moral reasoning of autonomous systems. *Ethics and Information Technology*, *18*(2), 103-115. doi:10.1007/s10676-016-9389-x

Bennis, W. M., Medin, D. L., & Bartels, D. M. (2010). The costs and benefits of calculation and moral rules. *Perspectives on Psychological Science*, *5*(2), 187-202. doi:10.1177/1745691610362354

Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 1-12. doi:10.1177/2053951715622512

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451-482. doi:10.1146/annurev-psych-120709-145346

Hafenbrädl, S., Waeger, D., Marewski, J. N., & Gigerenzer, G. (2016). Applied decision making with fast-and-frugal heuristics. *Journal of Applied Research in Memory and Cognition, 5*(2), 215-231. doi:10.1016/j.jarmac.2016.04.011

Hertwig, R. & Herzog, S. M. (2009). Fast and frugal heuristics: Tools of social rationality. *Social Cognition, 27*(5), 661–698. doi:10.1521/soco.2009.27.5.661

Hoffman, M., Yoeli, E., & Nowak, M. A. (2015). Cooperate without looking: Why we care what people think and not just what they do. *Proceedings of the National Academy of Sciences, 112*(6), 1727-1732. doi:10.1073/pnas.1417904112

Indurkhya, B., & Misztal-Radecka, J. (2016). Incorporating human dimension in autonomous decision-making on moral and ethical issues. In B. Indurkhya & G. Stojanov (Eds.), *AAAI Spring Symposium: Ethical and Moral Considerations in Non-Human Agents*. Palo Alto, CA: American Association for Artificial Intelligence.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences, 20*(8), 589-604. doi:10.1016/j.tics.2016.05.011

Keller, N., & Katsikopoulos, K. V. (2016). On the role of psychological heuristics in operational research; and a demonstration in military stability operations. *European Journal of Operational Research, 249*(3), 1063-1073. doi:10.1016/j.ejor.2015.07.023

Malle, B. F., & Scheutz, M. (2014). Moral competence in social robots. In *2014 IEEE International Symposium on Ethics in Science, Technology and Engineering*, IEEE. doi:10.1109/ETHICS.2014.6893446

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 1-21. doi:10.1177/2053951716679679

Todd, P. M., & Gigerenzer, G. (2007). Environments that make us smart: Ecological rationality. *Current Directions in Psychological Science, 16*(3), 167-171. doi:10.1111/j.1467-8721.2007.00497.x