Hidalgo César A, Orghian Diana, Albo-Canals Jordi, De Almeida Filipa & Martin Natalia (2021) How Humans Judge Machines. Cambridge, MA: MIT Press. 256 pages. ISBN: 9780262045520

Manh-Tung Ho

tung.homanh@phenikaa-uni.edu.vn

How humans judge machines by Hildago et al. (2021) is a very readable and informative book on the topic of human-machine relations. Central to the book's contribution is the study of more than 5,900 subjects, who were asked to judge the morality of scenarios where humans and artificial intelligence (AI) make consequential decisions. These scenarios are not far-fetched; for example, the respondents were asked to review how morally wrong and how much intention was involved when AI or humans made decisions about security checking in an airport or screening a job applicant. The book provides a systematic comparison of the differences between people judging humans versus judging machines, with the results presented in a series of pleasing and intuitive visualizations, bringing to light the complexity of our judgment towards AI, which depends not only on moral dimensions but also on contexts. For example, Hildago et al. show people judge machines the harshest when it comes to situations involving physical harm such as failure of diagnosis or a car crash. Meanwhile, people judge humans more harshly when the situations are perceived as not fair.

Here, the moral dimensions are derived from the theory of moral foundations by Jonathan Haidt (2007), which proposes there are five dimensions of morality: Harm, Fairness, Loyalty, Authority, and Purity. Hildago et al. argue that this method could "quantitatively unpack" the ethics of how humans relate to AI in the same way it has allowed psychologists to unpack variations in moral preferences. For more than 80 scenarios, the authors asked each respondent to pick four words that best describe each from a list of carefully selected twenty words. To illustrate, if the respondents picked indecent and harmful, then the scenario involves the purity and harm dimensions. The authors also introduce us to the moral space, a mathematical construct that quantifies the perceived morality of a situation as a function of a person's perception of how an agent (a human or a machine) involved in the situation has performed in each of the dimensions above. The data show most of the demographic variations in the data can be accounted for, implying the high applicability of Haidt's theory of the five moral foundations.

Going on this journey from one experiment to another, Hildago et al. show us many deep-seated intuitions we harbor about AI. The most crucial difference between our judgments towards AI versus towards humans is that we tend to not ascribe intention to AI, thus we judge them more by the outcomes, while the morality of a situation involving a human decision-maker is judged more by the intention. A poignant example is that in the event of a natural disaster, machines will be judged harshly if they try to save humans and fail, while people in the same scenario will still be judged positively. Such observation is greatly relevant since we are increasingly in the presence of AI systems whose performance is not of 100% success or accuracy rate but is nonetheless better than their human counterparts. For example, data from the 65,000 miles of self-driving cars by Waymo demonstrated how the current generation of autonomous vehicles can entirely avoid collision modes that are often caused by human

This work is licensed under a Creative Commons Attribution 4.0 International License

drivers such as road departure or fixed objects collision (Schwall et al., 2020). Such technologies can save many more lives and prevent many more deaths, and yet given human psychology, they would still be perceived as not trustworthy as humans.

The authors caution us that the book is strictly positive, meaning it merely describes how humans judge machines, not how we should judge machines. Yet, the aforementioned observations clearly imply that, for humans to create an AI-powered world that maximizes the benefits for people, we should relax our very human tendency to use intention as a heuristic to judge the morality of a situation. Toward the end of the book, the authors explore such a dilemma via the concept of machine responsibility, where legal concepts of liability, negligence, and recklessness are useful. In sum, the authors surmise that all liability must fall on humans. Thus, as a society, we need to think deeply about how to allocate responsibility to different humans: the engineers, the users, the sellers, etc., so as mitigate the unintended consequences that will occur upon the creation of new laws and regulations on AI use.

One of the issues that could be expanded on is the problem of cross-cultural differences in building and judging AI systems. The authors conclude that different AI systems trained with datasets from different societies will be encoded with different biases and preferences. For example, since the data of the book come from people living in the United States, a more individualistic and libertarian society, it is expected that in the scenarios where the government deploys the AI will be viewed with more distrust. However, in a country where communitarian ethics are more dominant such as East Asian nations, we can expect different results (Vuong, 2022; Roberts et al., 2021; Mantello et al., 2021).

Nevertheless, the beauty of the moral space construct and the experimental design in How humans judge machines is that future studies can build upon such methods and further explore how different moral values interact with each other and determine the perceived morality of a situation that involved machines. In this sense, the book offers a novel, interdisciplinary set of methods and tools for quantitatively probing how our moral intuitions are shifting with each encounter with ever more impressive and prevalent AI systems. Critically, it supplements the lack of emphasis on moral dimensions in technological adoption among previous empirical studies dominated by the Technological Acceptance Model (Taderhoost, 2018). The Technological Acceptance Model and its variations postulate that acceptance of new technology is primarily a function of perceived utilities and ease of use. This intuition might not hold anymore since AI systems interact with us in more sophisticated, yet subtle ways and often produce surprising results. For example, an AI system that analyzes the emotions of workers in an Amazon factory might not be visible to the workers, yet the knowledge of its existence can have outsized effects on workers' well-being and productivity (Du, 2022). More importantly, these effects can have very different cultural underpinnings depending on the native understanding of values such as individual liberty, privacy, autonomy, security, or fairness (Ishibushi, 2021; Degli Esposti et al., 2017; Miyashita, 2021).

As shown in How humans judge machines, perceptions of how machines change and interact with our behaviors and psychology can be a great source of unease in society. Thus, by placing human values and moral psychology at the heart of studying human-AI interaction, Hidalgo et al. (2021) remind us of our coevolving and increasingly interdependent relationship with technologies.

References

Degli Esposti S, Pavone V and Santiago-Gómez E (2017) Aligning security and privacy: The case of Deep Packet Inspection. In: Friedewald MJ, Burgess P, Čas J, Bellanova R, and Peissl W (eds) S*urveillance, Privacy and Security*. London: Routledge, pp. 71-90.

Du L (2022) Amazon grapples with more labor strife, this time in Japan. *Bloomberg,* September 6. Available at: https://www.bloomberg.com/news/articles/2022-09-06/amazon-grapples-with-more-labor-strife-this-time-in-japan?leadSource=uverify%20wall (accessed 14.2.2023).

Haidt J (2007) The new synthesis in moral psychology. Science 316(5827): 998-1002.

Hidalgo CA, Orghian D, Canals JA, De Almeida F and Martin N (2021) *How humans judge machines*. Cambridge, MA: MIT Press.

Ishibushi K and Matsakis L (2021) Masafumi Ito, leader of the Amazon Japan Union, sues for wrongful termination. Available at: https://restofworld.org/2021/lawsuit-amazon-japan-union/ (accessed 3.10.2023).

Mantello P, Ho MT, Nguyen MH and Quan-Hoang V (2021) Bosses without a heart: socio-demographic and cross-cultural determinants of attitude toward Emotional AI in the workplace. *AI & Society 3*8: 97–119. https://doi.org/10.1007/s00146-021-01290-1

Miyashita H (2021) Human-centric data protection laws and policies: A lesson from Japan. *Computer Law & Security Review* 40: 105487. https://doi.org/https://doi.org/10.1016/j.clsr.2020.105487

Roberts H, Cowls J, Morley J, Taddeo M, Wang V and Floridi L (2021) The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. *AI & Society*:36(1): 59-77.

Schwall M, Daniel T, Victor T, Favaro F and Hohnhold H (2020) Waymo public road safety performance data. *arXiv preprint* arXiv:2011.00038.

Taherdoost H (2018) A review of technology acceptance and adoption models and theories. *Procedia Manufacturing* 22: 960-967. https://doi.org/https://doi.org/10.1016/j.promfg.2018.03.137

Vuong QH (2022) *Mindsponge theory*. Warsaw: De Gruyter.