ARTIFICIAL INTELLIGENCE VALUE ALIGNMENT PRINCIPLES: THE STATE OF ART REVIEW FROM INFORMATION SYSTEMS RESEARCH

SHENGNAN HAN[1] & SHAHROKH NIKOU[2]

[1] Stockholm University, Department of Computer and Systems Sciences; e-mail: shengnan@dsv.su.se
[2] Åbo Akademi University & Stockholm University; e-mail: shahrokh.nikou@abo.fi

Abstract Information and communication technologies (ICTs) must be designed and used for humane ends. The rapid adoption of Artificial Intelligence (AI) has raised the critical question of whether we can ensure AI's alignment with human values to guide its design and use. We perform a selective literature review with the specific search terms of the papers published in the top information systems (basket of 8 journals and 5 AI journals in IS) from 2000-2020 to answer this question. The findings indicate that IS research has contributed insufficiently to a deeper understanding of human values and AI value alignment principles. Moreover, the mainstream IS research on AI is mostly dominated from its technical and managerial aspects. Thus, the future research agendas are proposed accordingly. The paper provides some food for thoughts in studying human values and AI alignment within the context of IS research.

Keywords: artificial intelligence, alignment principles, human values, information systems research, literature review

University of Maribor Press

DOI https://doi.org/10.18690/978-961-286-485-9.47
ISBN 978-961-286-485-9

1 Introduction

Berente et al. (2019) define AI as machines performing cognitive functions that we typically associate with humans, including perceiving, reasoning, learning, and interacting with others. They emphasize that "AI is not confined to one or a few applications, but rather is a pervasive economic, societal, and organizational phenomenon" (p. 1). To achieve what Walsham (2012) has argued that we must direct ICT at humane ends, AI should be aimed at making this a better world by using its highly optimized mechanistic functions and super intelligence to serve human needs, satisfy human desires and to maximize the realization of human values (e.g., Yudkowsky, 2011). This is also proposed as the AI value alignment principles (Russell, 2019) or as Sutrop (2020) put forward designing AI that conforms to human values is called 'value alignment'. One fundamental and critical question is raised and intensively debated: *how can we ensure AI alignment with human values through AI operations from design to use?* Yamposkiy (2017) argued that, because of the unresolved disagreements in the disciplines of philosophy and axiology regarding the nature and content of human values, the question of how to align these values in AI development and use, is also moot. The IS community has not yet paid sufficient attention to this AI phenomenon and has contributed insufficiently to a deeper understanding of human values in general (Carman and Rosman, 2020, Lyytinen et al., 2020). In this paper, we first analyze the AI phenomenon as it is discussed in the top IS research outlets (basket of 8 journals and 5 AI journals in IS). It should be noted that the AI value alignment and its connection to ethical concerns is not included in the search because, while significant, this topic is not the paper's focus. Upon the results from the literature review, we propose the future research agendas.

2 AI Value Alignment Principles

Russell (2019) has proposed the three AI value alignment principles for creating a safe and beneficial AI. (1) A principle of altruism: the AI's only objective is to maximize the realization of human values. Here, human values are defined as what "we" would "prefer our life to be like". (2) A law of humility: AI as the digital agents is initially not certain of what human values are. But AI agents, in support of advanced machine learning capabilities, may learn those values and preferences by observing "our" behaviors. (3) To achieve the value alignments between AI and

humans, we, in this process, must learn to be better persons, or, perhaps, simpler. The aim should be ensuring that AI agents can learn the essential value-goods such as safety, healthcare, food and shelter, and meaningful work from "us". AI agents must be explicably programmed to make such values primary where and when needed. We acknowledge that advanced technical solutions are not sufficient for fulfilling the AI value alignment principles (e.g. Christian, 2020). Therefore, the multiple aspects of human values should be fully explored in AI design and use.

3 Research Method

We followed Lowry et al. (2004) recommendation to select the journals and articles. We searched Web of Science, INFORM, EBSCO, and Google Scholar. We separately also searched the basket of 8 IS journals and the top 5 AI journals in IS. As inclusion criteria, an article had to be original study and published in that top IS journal between 2000-2020 and written in English. Moreover, to be included in the review, articles had to match exact the search terms used during the publication search. As we used specific search terms (such as artificial intelligence AND human value*, or AI AND human value*), the initial database search retrieved 327 articles. In the next step, we excluded all duplicated and articles that did not adhere to our search criteria (n = 302 in total). The most frequent reason for excluding an article was that, although drawing to some extend on AI, the article did not primarily use AI in the context of human value or focused mainly on ethical issues. This final dataset is composed of 25 AI articles[1], which were downloaded in full text and reviewed by authors. After reviewing the 25 articles, it has become clear that none of the articles, although appeared to be relevant, discussed or approached "*AI and human value"* like our current approach. Nonetheless, the review results are briefly presented in the next section.

4 IS Research on "AI and Human Values"

The IS community has provided limited exploration of human values and of the possible AI alignment with those human values. Most of the reviewed articles limit their contributions towards AI technical problems (e.g., Li et al., 2009; Wong et al., 2020), and very few have implicitly discussed AI's impacts on humans, organizations,

[1] Due to page limitation, we cannot include a full list of all authors' information in the reference list.

and society in general. For example, Ransbotham et al. (2016, p. 1) argued that while IT provides many advantages to humans, their organizations, and to society in general, also have the potential to create new vulnerabilities such as online harassment, incivility, a merely algorithmic ethics, and bias towards minorities. In another study, Aleksander (2017) argued that, as robots and other machines operate in an algorithmic way and not in a truly cognitive and conscious human way, AI can present serious threats to humanity if the algorithms are not aligned with broader sets of values than those of pragmatic efficiency. Elkins et al. (2013) demonstrated that using artificial technology and integrating AI into advanced expert systems inadvertently imposes threats even to human experts and inhibits users from adopting the technology. Nonetheless, Aleksander (2004) argued that as AI technology develops more and more, it has greater potential for overcoming some of the unforeseen difficulties as humans pursue some very ambitious projects. Glezer (2003, p. 65) argued that using AI for automation of tasks is problematic, for the software agents often interfere with the human ability to specify the amount of control they would like to have over the agent's behavior. Nicolescu et al. (2018) investigated the emerging meanings of "value" associated with the Internet of Things (IoT) and argued that the multiple meanings of "value" are invariably articulated at the juncture of three domains: social, economic, and technical. Huysman (2020) asserted that we should create societal awareness about the rise of low quality of work due to AI rather than focusing merely on the effect of AI on job losses. Grønsund and Aanestad (2020, p. 14) argued that while research on algorithmic and intelligent technologies has generated insights about their potential to replace human work; however, the emergent configurations by which humans and algorithmic interplay emerge has not been investigated. In summary, the current AI research in IS field is restricted to technical developments and design issues. AI design and its alignment with human values are not yet being fully considered.

5 Discussion and Future Research Agendas

The study results clearly demonstrate that IS research has not yet sufficiently contributed to a deeper understanding of human values and AI value alignment and how to achieve the AI value alignment principles. Thus, we propose the following three research foci. First, we need to understand what are human values from different philosophical and ontological schools of thoughts within IS. Ågerfalk (2020) argues that IS research can contribute significantly to advance AI

development and use. We need to focus on the three key components, contextualization, communication, and practice to complete the inquires of AI phenomenon. AI phenomenon is much more complex with great uncertainties. We can explore this complexity from various school of thoughts with the aim of producing more comprehensive understanding of what are the human values and add IS perspectives to this multidisciplinary theme. Second, we need to understand what are the critical human values within the contexts of AI design and use. Human values are deeply rooted in cultural and social traditions. Gabriel (2020) points out that human preferences that are always embedded in a range of human values, may not be sufficient, though necessary, to give instructions to an AI agent for achieving desired outcomes. Our "immediate" preferences may differ largely than what we prefer in the longer time. AI systems are kind of IS artefacts (Chatterjee et al., 2017; Lee et al., 2015). The study of the critical human values can be conducted in the context of AI as an information artefact, AI as a technology artefact and AI as a social artefact. As well we need to study the interactions among the three artefacts components to reach the conclusion, i.e., what are the critical human values should be aligned with AI design. Third, we need to prioritize the critical human values in AI design and use for different user groups in various cultural, social, and personal contexts. To keep a positive reciprocal relationship of human/AI, we need to become a better person and use AI in a positive and ethical way. This is also what the AI alignment principles have proposed. Since the AI may learn from us (a law of humitity), we need to behave well and generate more positive values that benefit AI design. Thus, future studies can investigate the effects of the prioritized human values on the users' behaviors towards AI systems within a specific context.

6 Conclusions

We briefly address current IS research on human values and AI in this paper, as well as some perspectives on the importance of achieving AI value alignment principles. We also suggest three research directions that IS researchers can pursue, but they are tentative and might be naïve in their current state of development. We believe that the paper provides some food for thought about the significance of studying human values in AI design and application in IS research.

References

Ågerfalk, P. J. (2020). Artificial intelligence as digital agency. European Journal of Information Systems, 29 (1), 1-8.

Aleksander, I. (2004). Advances in intelligent information technology: re-branding or progress towards conscious machines? Journal of Information Technology, 19 (1), 21-27.

Aleksander, I. (2017). Partners of humans: a realistic assessment of the role of robots in the foreseeable future. Journal of Information Technology, 32 (1), 1-9.

Berente, N., Gu, B., Recker, J., and Santhanam, R. (2019). Managing AI. Call for Papers, MIS Quarterly, 1-5.

Carman, M., and Rosman, B. (2020). Applying a principle of explicability to AI research in Africa: should we do it? Ethics and Information Technology. Available at: https://doi.org/10.1007/s10676-020-09534-2.

Chatterjee, S., Xiao, X., Elbanna, A., and Sarker, S. (2017). The Information Systems Artifact: A Conceptualization Based on General Systems Theory. Paper presented at the Hawaii International Conference on System Sciences.

Christian, B (2020). The alignment problem: machine learning and human values, W. W. Norton & Company, New York.

Elkins, A. C., Dunbar, N. E., Adame, B., and Nunamaker, J. F. (2013). Are users threatened by credibility assessment systems? Journal of Management Information Systems, 29 (4), 249-262.

Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. Minds and Machines, https://doi.org/10.1007/s11023-020-09539-2.

Glezer, C. (2003). A conceptual model of an interorganizational intelligent meeting-scheduler (IIMS). The Journal of Strategic Information Systems, 12 (1), 47-70.

Grønsund, T., & Aanestad, M. (2020). Augmenting the algorithm: Emerging human-in-the-loop work configurations. The Journal of Strategic Information Systems, 29(2), 101614.

Huysman, M. (2020). Information systems research on artificial intelligence and work: A commentary on "Robo-Apocalypse cancelled? Reframing the automation and future of work debate". Journal of Information Technology, 0268396220926511.

Lee, A., Thomas, M., and Baskerville, R. L. (2015). Going back to basics in design science: from the information technology artifact to the information systems artifact. Information Systems Journal, 25, 5-21.

Li, X., Chen, H., Zhang, Z., Li, J., and Nunamaker, J. F. (2009). Managing knowledge in light of its evolution process: An empirical study on citation network-based patent classification. Journal of Management Information Systems, 26 (1), 129-154.

Lowry, P., Romans, D., and Curtis, A. (2004). Global Journal Prestige and Supporting Disciplines: A Scientometric Study of Information Systems Journals. Journal of the Association for Information Systems 5(2), 29-77.

Lyytinen, K., Nickerson, J. V., and King, J. L. (2020). Metahuman systems= humans+ machines that learn. Journal of Information Technology. Available at: https://doi.org/10.1177/0268396220915917.

Nicolescu, R., Huth, M., Radanliev, P., and De Roure, D. (2018). Mapping the values of IoT. Journal of Information Technology, 33 (4), 345-360.

Ransbotham, S., Fichman, R. G., Gopal, R., and Gupta, A. (2016). Special section introduction—ubiquitous IT and digital vulnerabilities. Information Systems Research, 27 (4), 834-847.

Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Penguin.

Sutrop, M. (2020). Challenges of Aligning Artificial Intelligence with Human Values. Acta Baltica Historiae et Philosophiae Scientiarum, 8(2).

Walsham, G. (2012). Are we making a better world with ICTs? Reflections on a future agenda for the IS field. Journal of Information Technology, 27(2), 87-93.

Wong, N., Ray, P., Stephens, G., and Lewis, L. (2012). Artificial immune systems for the detection of credit card fraud: an architecture, prototype and preliminary results. Information Systems Journal, 22(1), 53-76.

Yampolskiy, R. V. (2017). AI Is the Future of Cybersecurity, for Better and for Worse. Harvard Business Review.

Yudkowsky, E. (2011). Complex Value Systems are Required to Realize Valuable Futures. The Singularity Institute, San Francisco, CA. http://intelligence.org/files/ComplexValues.pdf.