Dec 16th, 12:00 AM

Fighting the real AI Danger: How to Design Virtuous AI for Virtuous Decision-making

Adeline Frenzel
University of Augsburg, adeline.frenzel@uni-a.de

Shilpi Jain
FORE School of Management, shilpijain@fsm.ac.in

Shizhen (Jasper) Jia
Washington State University, shizhen.jia@wsu.edu

Maximilian Welck
University of Augsburg, mw.research@welck.de

Nishtha Langer
Rensselaer Polytechnic Institute, langen@rpi.edu

Fighting the real AI danger: How to design virtuous AI for virtuous decision-making

Paper-a-Thon

Adeline Frenzel
University of Augsburg
Universitaetsstrasse 16
86135 Augsburg
Germany
adeline.frenzel@uni-a.de

Shilpi Jain
FORE School of Management
New Delhi, India
shilpijain@fsm.ac.in

Shizhen (Jasper) Jia
Department of Management, Information Systems, & Entrepreneurship
Carson College of Business,
Washington State University,
Pullman, WA, USA
shizhen.jia@wsu.edu

Maximilian Welck
University of Augsburg
Universitaetsstrasse 16
86135 Augsburg
Germany
mw.research@welck.de

Nishtha Langer
Lally School of Management, Rensselaer Polytechnic Institute
Troy, New York, United States of America
langen@rpi.edu

Abstract

Artificial Intelligence, i.e., complex algorithms that learn to perform functions associated with human minds, such as perceiving, decision-making, and demonstrating creativity. Indeed, more often than not, AI is trained on biased datasets, that is, this data is disproportionately weighted in favor of or against certain individuals or groups of individuals. The roots for such biases are very diverse, sometimes they are technical in nature, but often they originate in the minds of people, making it difficult to identify these biases, e.g. if such a disproportionate weighting is perceived as 'normal' by many, while it is still devastating to few.

We argue that the use of virtue ethics in AI can help to mitigate the consequences of biases and to help identify such biases. In particular, we aim to assess in multiple online and field experiments how the virtue 'transparency' affects an individual's decision-making and perceptions regarding the AI.

Keywords: Artificial intelligence, bias, virtue ethics design, decision-making, individual factors, experiment

Introduction

Artificial Intelligence (AI), i.e., complex algorithms that learn to perform functions associated with human minds, such as perceiving, problem-solving, decision-making, and demonstrating creativity from large

datasets. Indeed, more often than not, from biased datasets, that is, this data is disproportionately weighted in favor of or against specific individuals or groups of individuals. The roots for such biases are very diverse, sometimes they are technical in nature, but often they originate in the minds of people. The latter makes it very difficult to identify biases, as such a disproportionate weighting could be perceived as 'normal' to many, yet, be devastating to few. For example, Facebook has been criticized that its technology to run job advertisements is biases against women. Employers used Facebook's targeting technology and AI algorithms to exclude women from receiving ads that advertised open positions, such as for truck drivers, and window installers (Scheiber 2018). In the same vein, Amazon disclosed that their hiring algorithms were also biased against women, whose resumes were downgraded (Gonzalez 2018). While Amazon promptly scrapped the application of this AI, it has risen concerns about clandestine, and inherent biases in AI.

Theoretical background

The issue of biases and ethics in AI technology has gained a lot of interest in the Information System (IS) field in recent years. Ajunwa (2019) argues that the pervasive adoption and usage of AI in our daily lives entails the emergence of ethical issues concerned with biases. Current IS research has investigated biases and discriminations in labor markets. For example, studies find that AI can cause the resistance of professionals in the context of AI-augmented decision-making (e.g., Jussupow et al. 2020; LaBrie and Steinke 2019), some studies point out individuals' adverse reactions to algorithm-based management decisions (Tambe et al. 2019), and some research finds that AI technology may offer unintended consequences for employees, projects, and organizations (Reis et al. 2020). This underpins the statement of Tarafdar et al. (2020) "Bias cannot be avoided in the use of big data algorithms and exists in the technical as well as the social aspects." (p. 2). This makes the statement of Manyika et al. that "AI can help humans with bias — but only if humans are working together to tackle bias in AI." (Manyika et al. 2019), even more important.

Our approach to improving the identification and mitigation of the consequences of biased AI draws from the virtue ethics theory, one primary mode of ethical analysis. Virtue ethics theory focuses on personal character, how to develop good qualities, i.e., "virtues", and the ability to use them (Goldsmith and Burton 2017). Virtue ethics is a big-picture theory, and it argues that individuals' behaviors are strongly associated with social contexts. We argue that virtuous AI needs to transparently inform users about the setup of potential roots of biases. Virtuous AI used to select applicants in hiring processes, for example, needs to inform about the diversity of the training data, e.g., 65% female vs. 35% non-female, and its certainty levels, e.g., this candidate is very likely not suited for the respective position. Such transparency is precious for users and can motivate AI-developers to (more) carefully consider and mitigate potential biases. Accordingly, with the help of virtue ethics, this research seeks to answer the following two research questions*:* 1) **How does exposure to ethically virtuous designed AI affect individuals' decision-making?** 2) **How is this relation moderated by individual differences (e.g., gender, ethnicity)?**

In this research, we are going to conduct experiments to create scenarios that manipulate bias cues (x2) and ambiguity cues (x2). We propose two experimental designs. First, in order to understand how virtue ethic cues affect decision making in general, we will conduct an (online) experiment utilizing vignettes. The proposed sample are both crowd workers, and students. Second, we aim to apply a field experiment to understand decision making in the context of AI and virtue ethic cues. To get a more in-depth understanding we aim to also get responses to certain open-ended questions, next to our quantitative post-experimental assessments.

We anticipate that our multi-study design research approach will help theory development by explaining how virtue ethics in AI affects individuals' decision making. Further, we anticipate that we can contribute to practitioners by elaborating how virtuous AI affects users' decision making, and how AI can be designed virtuously.

Proposed Research Model

In this study, we apply the virtue ethics approach as our theoretical foundation to investigate how bias cues and ambiguity cues influence individuals' decision-making. The main idea of virtue ethics is the pursuit of

internal good, practical wisdom, and voluntary action (Gal et al. 2020). Building on the theoretical foundations described above, Figure 1 depicts our research model, which is drawn from virtue ethics. Ambiguity cues refer to the information regarding the confidence level of the AI, and bias cues refer to the information regarding the set-up of the underlying data.

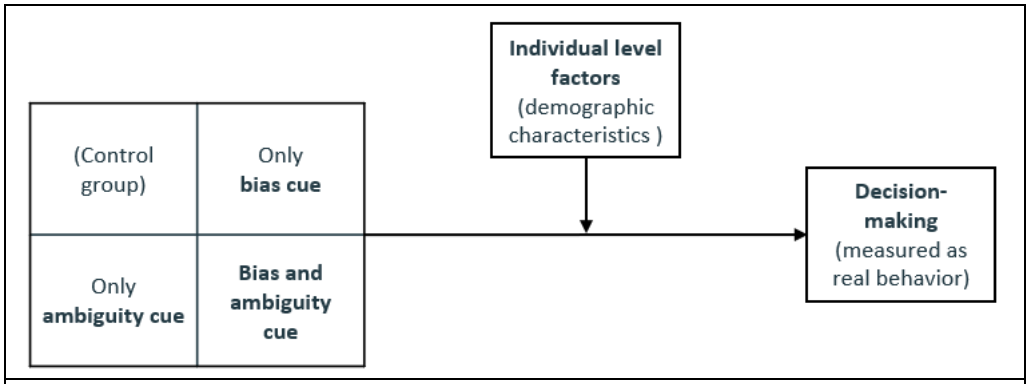

Figure 1. Proposed Research Model

Methodology

We will apply a 2 (bias cues: Yes vs. No) × 2 (ambiguity cues: Yes vs. No) between-subjects, full factorial multi-study design including an online experiment (study 1) and a field experiment (study 2). The online experiment will be based on vignettes.

Procedure: In group 1, the control condition (absence of bias and ambiguity cues), a general task description will be shown. In group 2 (provision of bias cues but lack of ambiguity cues), participants will be confronted with information regarding the reference data of the AI. For example, the AI could transparently state that reference data were 90% based on Asian and Caucasian men. In group 3 (provision of ambiguity cues but lack of bias cues), subjects will receive information on the confidence level of the AI recommendation, e.g., the recommendation is based on a confidence level of 73%. Finally, in group 4 (provision of bias and ambiguity cues), both cues are combined and will be presented.

Data Collection: For the pretests, we intend to invite students from different universities who are enrolled in management courses and have a prior understanding of how organizations are structured and which strategic role artificial intelligence technologies have in decision making. For the main data collection, we are planning to run a similar experiment with workers from the crowdsourcing platform Amazon Mechanical Turk (AMT) who are active for at least one year. The field experiment will be conducted in an organizational setting to collect decision-making as real behavior, in the context of AI and virtue ethic cues. Two possible contexts in consideration are hiring decisions and human resource management and diagnosis decisions in radiology.

Expected contribution

Theoretical implication

Drawing on virtue ethics, this study offers a theoretical framework to investigate how the two types of virtue ethics cues influence individuals' decision makings and provide a novel way to scrutinize AI biases and discriminations issue. Specifically, this study contributes to (1) extend virtue ethics in IS research, especially to understand how virtue ethics design principles affect individuals decision making (2) address lack of theoretically grounded empirical social-technological studies to investigate the topic of AI biases, (3) provide how virtue ethics design principles can mitigate different sources of algorithmic biases, and (4) explore whether the decision making varies among users from different experiences, ethnicity, gender, age, etc. This theory-driven model to individuals' decision-making behaviors will complement existing IS literature.

Practical Implications

From a practical perspective, we hope to achieve a better understanding of different individuals' behaviors with the guidance of AI. This study elaborates on informed and objective decision making and provides insights on how to design AI-based recommendations ethically. We need to highlight the role of information cues in our software and algorithm designs. In addition, this study improves confidence in the application of AI and stresses the need to mitigate adverse employee reactions. It is also essential for governments to make AI technology regulations more proximal to users.

Acknowledgments

We appreciate the insightful comments from session presentation and reviewers, as well as the kind support and guidance of the track chairs James Gaskin, Panayiotis Constantinides, and Anjana Susarla.

References

Ajunwa, I. 2019. "The Paradox of Automation as Anti-Bias Intervention," *Cardozo L. Rev*. (41), p. 1671.

Gal, U., Jensen, T. B., and Stein, M.-K. 2020. "Breaking the Vicious Cycle of Algorithmic Management: A Virtue Ethics Approach to People Analytics," *Information and Organization* (30:2), p. 100301.

Goldsmith, J., and Burton, E. 2017. "Why Teaching Ethics to Ai Practitioners Is Important," *Proceedings of the AAAI Conference on Artificial Intelligence*.

Gonzalez, G. 2018. "How Amazon Accidentally Invented a Sexist Hiring Algorithm." Retrieved 12/11/2020, 2020, from https://www.inc.com/guadalupe-gonzalez/amazon-artificial-intelligence-ai-hiring-tool-hr.html

Hmoud, B., & Laszlo, V. (2019). Will Artificial intelligence Take Over HumanResources Recruitment and Selection. *Network Intelligence Studies, 7*(13), 21-30.

Jussupow, E., Spohrer, K., Heinzl, A., and Gawlitza, J. 2020. "Augmenting Medical Diagnosis Decisions? An Investigation into Physicians' Decision Making Process with Artificial Intelligence," *Information Systems Research* (1:1), p. tba.

LaBrie, R. C., and Steinke, G. 2019. "Towards a Framework for Ethical Audits of Ai Algorithms," *Twenty-fifth Americas Conference on Information Systems*, Cancun: AIS.

Manyika, J., Silberg, J., and Presten, B. 2019. "What Do We Do About the Biases in Ai?" Retrieved 12/11/2020, 2020, from https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai

Pietri, E. S., Johnson, I. R., Ozgumus, E., & Young, A. I. (2018). Maybe she is relatable: Increasing women's awareness of gender bias encourages their identification with women scientists. *Psychology of Women Quarterly, 42*(2), 192-219.

Reis, L., Maier, C., Mattke, J., Creutzenberg, M., and Weitzel, T. 2020. "Addressing User Resistance Would Have Prevented a Healthcare Ai Project Failure," *MIS Quarterly Executive (19*:4), pp. 279–296.

Scheiber, N. 2018. "Facebook Accused of Allowing Bias against Women in Job Ads." Economy Retrieved 12/10/2020, 2020, from https://www.nytimes.com/2018/09/18/business/economy/facebook-job-ads.html

Tambe, P., Cappelli, P., and Yakubovich, V. 2019. "Artificial Intelligence in Human Resources Management: Challenges and a Path Forward," *California Management Review (61*:4), pp. 15-42.

Tarafdar, M., Teodorescu, M., Tanriverdi, H. s., Robert, L., and Morse, L. 2020. "Seeking Ethical Use of Ai Algorithms: Challenges and Mitigations," *ICIS 2020*, India: AIS.