Kim, Woo Jin; Ryoo, Yuhosua; Kim, Eunjin Anna; Stafford, Marla

Conference Paper

Hero or Villain: The Paradox of AI Algorithmic Disclosure in Utilitarian Versus Deontological Ethics

24th Biennial Conference of the International Telecommunications Society (ITS): "New bottles for new wine: digital transformation demands new policies and strategies", Seoul, Korea, 23-26 June, 2024

Provided in Cooperation with:
International Telecommunications Society (ITS)

Suggested Citation: Kim, Woo Jin; Ryoo, Yuhosua; Kim, Eunjin Anna; Stafford, Marla (2024) : Hero or Villain: The Paradox of AI Algorithmic Disclosure in Utilitarian Versus Deontological Ethics, 24th Biennial Conference of the International Telecommunications Society (ITS): "New bottles for new wine: digital transformation demands new policies and strategies", Seoul, Korea, 23-26 June, 2024, International Telecommunications Society (ITS), Calgary

This Version is available at:
https://hdl.handle.net/10419/302483

Hero or Villain:
The Paradox of AI Algorithmic Disclosure in Utilitarian vs. Deontological Ethics

WooJin Kim, Ph.D., University of Colorado Boulder

Yuhosua Ryoo, Ph.D., University of Minnesota Duluth

Eunjin (Anna) Kim, Ph.D., University of University of Southern California

Marla Royne Stafford, Ph.D., University of Nevada, Las Vegas

Keywords: AI algorithmic disclosure, AI message type, Self-efficacy, Empathy

Introduction

Artificial intelligence (AI) already permeates various aspects of our daily lives, with applications ranging from recommendation systems and autonomous vehicles to personal home assistants and educational support systems (Kaur et al., 2020). These AI systems meet many of our personal needs while also affecting different areas of our social interactions. Additionally, AI technologies are highly effective in several key areas, enabling them to promote prosocial behaviors and enhance social welfare (Efthymiou & Hildebrand 2023). First, AI can be programmed to make decisions free from the biases that typically affect human judgment, promoting fairer and more equitable outcomes (Lin et al., 2021). Thus, AI can help allocate resources efficiently, maximizing impact without the influence of personal biases that might sway human donors or organizations (Landers & Behrend, 2023). Moreover, the constant availability and scalability of AI make it ideal for addressing large-scale social challenges such as managing disaster responses or optimizing resource distribution during crises (Sun et al., 2020). By integrating these capabilities, AI not only supports individual well-being but also bolsters collective welfare through the promotion of ethical and efficient solutions to complex social issues.

However, despite these advantages, there are still significant uncertainties regarding the boundary conditions under which AI is or is not an effective tool for prosocial behavior, as well as the underlying mechanisms involved. To address these gaps, our research explores two main moral approaches that AI can use to encourage prosocial behaviors: deontological and utilitarian principles (Conway & Gawronski, 2013). In moral philosophy, prosocial behaviors—actions that aim to help or benefit others—are deeply rooted in moral reasoning, predominantly guided by deontological or utilitarian principles (Conway & Gawronski, 2013). Deontology evaluates the morality of actions based on a set of rules or duties, emphasizing the intrinsic nature of the actions themselves. In contrast, utilitarianism assesses actions based on their outcomes, advocating for actions that maximize overall well-being.

As AI systems increasingly undertake tasks traditionally performed by humans, the importance of incorporating moral principles in the context of

WooJin.Kim@colorado.edu
yryoo@d.umn.edu
eunjink@usc.edu
marla.stafford@unlv.edu

AI is crucial. AI agents can reach a broad audience consistently and at scale, making them an efficient tool for disseminating prosocial messages (Efthymiou & Hildebrand 2023). This growing role of AI raises significant questions regarding its capacity to replace human moral decision-making. While AI may not fully replace human moral judgment, it can certainly serve as a valuable tool to promote prosocial behaviors, leveraging its reach and consistency to foster positive societal impacts. However, to the best of our knowledge, there is virtually no research examining how these moral frameworks, when conveyed through AI, affect human prosocial responses. This gap in research becomes even more pressing when we consider the rapid integration of AI into various aspects of consumer life. AI's potential to act as a moral agent or advisor in prosocial decisions is immense but largely untapped. As AI begins to 'speak' on moral grounds, will consumers listen? And if so, how will the knowledge that an algorithm is behind the message shape their response?

This study is poised to fill this void by exploring how AI-generated messages framed within deontological and utilitarian contexts influence consumer behavior, specifically their willingness to help others. It examines the critical role of algorithmic disclosure—whether revealing or concealing AI's involvement moderates the message's impact. This research explores the psychological mechanisms at play, proposing that such disclosure may boost self-efficacy in the context of utilitarian messages, while potentially eroding empathy when associated with deontological messages.

Specifically, we posit that algorithmic disclosure differentially influences consumer perceptions of AI-generated messages rooted in utilitarian and deontological moral frameworks. Drawing on the work of Waytz et al. (2010), we suggest that revealing the algorithmic underpinnings of AI systems accentuates their machine-like nature, thereby invoking machine heuristics. Such heuristics frame AI agents as precise, objective, and aligned with outcome-focused reasoning, which may enhance the perceived appropriateness of AI delivering utilitarian messages, thus boosting perceived self-efficacy in these contexts. Conversely, Sundar (2020) indicates that algorithmic disclosure can also engender perceptions of AI as mechanistic and devoid of emotion, which might undermine the perceived suitability of AI in communicating deontological ethics, thereby diminishing empathy. From this perspective, we hypothesize that algorithmic disclosure will result in more favorable responses to utilitarian messages due to increased self-efficacy, whereas the disclosure in deontological contexts is likely to elicit less favorable responses because of reduced empathy.

We believe understanding these dynamics is not just academically intriguing; it has profound implications for how businesses and social organizations can harness AI to foster prosocial behavior effectively. By pinpointing why and how consumers react to AI's moral communications, we can better design AI systems that not only understand human values but also reinforce them, leading to greater societal benefits and paving the way for more ethically aware AI applications. The outcomes of this study could revolutionize the strategic use of AI in communications, policy-making, and beyond, making it a pivotal piece of research at the intersection of technology, psychology, and ethics.

Theoreticla Background

Utilitarian vs. deontological ethics

Moral judgements are traditionally assumed to be rooted in deliberate thought processes based on rational application of reasoned behaviors of basic abstract moral principles (Kohlberg, 1969). That is, the essence of moral judgment is grounded in rational acts attributed to the moral foundations of an individual and driven by acts guided by rational beliefs.

Scholars, however, have challenged these long-held assumptions with more recent beliefs moved toward moral approaches not necessarily based on a reasonable theory of action (Haidt,

2001). Such approaches are developed based on affective principles guided by more intuitive and emotional judgments (Greene & Haidt, 2002). Specifically, Gawronski and Beer (2017) acknowledge the integration of both reasoned and nonreasoned processes has become a prominent research paradigm. By integrating different perspectives into the moral judgement literature, two distinct paths related to helping behavior coalesce: utilitarian and deontological approaches. Merging these two approaches allows researchers to more holistically understand unique perspectives and what drives moral dilemmas. This argument is consistent with Gray and Schein (2012) who suggest that immorality encompasses two aspects of blame: acts and consequences. In fact, they specifically assert "Moral cognition simultaneously concerns acts and consequences."

The utilitarian approach is entrenched in outcomes. Its overarching philosophy is driven by the desire to have positive consequences and outcomes, specifically for overall well-being. As such, from a utilitarian perspective, a decision based on the positive consequences that would result is morally acceptable. The utilitarian approach is also based on cognitive factors aligned with objective evaluations that rationally drive outcomes that result in positive well-being. Utilitarian message appeals are generally considered rational and objective and seek the greatest amount of good (Playford et al., 2015) and for the greatest number of people (Anderson & Anderson, 2011). That is, if the decision results in overall well-being for a large number of people, it is considered moral (Hennig & Hutter, 2020).

In contrast, deontological moral approaches are guided by the particular situation faced, with priority given to consistency with moral norms. That is, the decision is driven by adherence to the rules and norms used to actually make the specific decision, as opposed to the outcomes or consequences that arise from the decision. Hence, deontological approaches are associated with the act itself and whether or not it is moral. It generally disregards whether the consequences are negative or positive. Deontology stresses treating people right because it's the right thing to do, as opposed to treating people right to achieve a target positive outcomes (Playford et al., 2015).

The integration of these two moral philosophies is the underlying principle of the well-known Dual-Process Theory of Moral Judgement studied by Greene and several of his colleagues (e.g., Cushman, Young, & Greene, 2010; Greene, et al., 2008; Greene, et al., 2004; Greene, et al., 2001; Paxton, Ungar, & Greene, 2012).) Greene and others note their work has been supported by the well-known Trolley Problem (e.g. Fischer & Ravizza, 1992; Thomson, 1985) well documented in the moral philosophy research. Also referred to as the trolley dilemma, the situation questions whether it is morally acceptable to "divert a runaway trolley that threatens five lives onto a side track, where it will run over and kill only one person instead" (e.g., Greene, 2009; Greene et al., 2001; Mikhail, 2000). Generally, people believe it is morally acceptable to divert a runaway trolley that threatens five lives onto a side track, where it will run over and kill only one person instead (Greene et al., 2001; Mikhail, 2000). The trolley problem also reinforces the belief that automatic emotional responses are indicative of the deontological approach (e.g. disapproving of killing one person to save several others) while cognitive processes drive utilitarian judgments (e.g. approving of killing one to save several others).

Although Greene (2009) acknowledges the dissenting work by McGuire et al (2009), he reaffirms his original findings, by clearly detailing the flaws in the others" work. Greene (2009 p. 583) concludes "… McGuire and colleagues conflate the dual-process theory of moral judgment with the personal/ impersonal distinction, too hastily dismiss more recent convergent evidence for the dual-process theory, and completely ignore the evidence that bears most directly on the issues they raise." Greene further provides additional support for the dual process theory.

Interestingly, Gawronsky and Beer (2017) argue that moral research has only minimally

manipulated and, thus, investigated outcomes in experimental work, suggesting the difficulty in accurately interpreting research investigating the approaches to moral dilemmas and that such ambiguity exists in research on both utilitarian and deontological judgments. These assertions signal the need for more experimental work in this area. Gawronsky and Beer (2017) further note the need for experiments manipulating moral norms to better understand and resolve such interpretational ambiguities. Such research has the power to understand patterns of utilitarian and deontological responses such as moral norm consistency and actions related to outcomes of well-being. Moreover, understanding differences in these two perspectives can offer an explanation as to why individuals make different choices based on their own moral perspectives.

Algorithmic disclosure

AI systems are powered by sophisticated algorithms (Khaleel et al., 2023). Prosocial chatbots, in particular, rely on these algorithms to encourage prosocial behaviors by analyzing user interactions and employing persuasive communication techniques that are tailored to individual preferences and contexts (Namkoong et al., 2023). Designed to promote actions such as community involvement, environmental conservation, and charitable giving, these chatbots leverage data-driven insights to identify the most effective strategies for influencing users towards making decisions that benefit society (Park et al., 2023). Despite the notable benefits of using algorithms for prosocial encouragement, their integration raises some concerns. Algorithms operate behind the interface, making the decision-making process opaque (Dwivedi et al., 2021).

This lack of transparency means that consumers are often unaware of the specific moral reasoning AI chatbots use to encourage prosocial behaviors. This issue highlights the necessity for algorithmic disclosure that ensures users are informed about how underlying processes are used to recommend specific prosocial behaviors (Eslami et al., 2015). Thus, we suggest that algorithmic disclosure should be an integral part of algorithm operations in prosocial chatbots.

Algorithmic disclosure refers to the practice of making the decision-making processes, criteria, and underlying data used by algorithms transparent to users (Di Porto, 2023). This transparency may include explanations of how algorithms process user data, how decisions are made, and what factors influence these decisions (Bell et al., 2023). By facilitating this disclosure, consumers can better understand and recognize the mechanisms, functions, and impacts of algorithms within technological systems (Wang, 2023). Specifically, algorithmic disclosure helps consumers make educated judgments about their interactions with algorithmic platforms, leading to more informed decisions (Zarouali et al., 2021).

In our research, we posit that algorithmic disclosure differentially influences consumer perceptions of AI-generated prosocial messages based on utilitarian and deontological moral frameworks. We suggest that revealing the algorithmic foundations of AI systems emphasizes their machine-like nature, which invokes machine heuristics (Sundar, 2008). These heuristics portray AI agents as precise, objective, and aligned with outcome-focused reasoning, potentially enhancing the perceived appropriateness of AI for delivering utilitarian messages by boosting perceived self-efficacy in these contexts. Conversely, we predict that algorithmic disclosure can also lead to perceptions of AI as cold, mechanical, and devoid of emotional depth, which might undermine the perceived suitability of AI for communicating deontological ethics by diminishing empathy. From this perspective, we generated the following hypotheses: algorithmic disclosure will result in more favorable responses to utilitarian messages due to increased self-efficacy, while disclosure in deontological contexts is likely to elicit less favorable responses due to reduced empathy.

H1: Algorithmic disclosure will result in more (less) favorable responses to utilitarian

(deontological) messages in prosocial campaigns.

H2: Enhanced self-effifacy and reduced emphasy will mediate such effects.

<Insert Figure 1 about here >

Methods

Our objective was to find preliminary evidence that an interaction occurs between algorithmic disclosure and AI message type (H1), mediated by self-efficacy and empathy (H2). For this research, we developed four different chatbots—crossing algorithmic disclosure with non-disclosure, and utilitarian messages with deontological messages—using web-based algorithms, and integrated them into Facebook Messenger. Participants engaged with the chatbots by logging into their Facebook accounts to enhance realism.

Sample and research design

We used a between-subjects design with a 2 (algorithmic disclosure: yes vs. no) × 2 (AI message type: utilitarian vs. deontological). We recruited 486 participants (M = 33.84 years, 370 women) from Amazon's Mechanical Turk (MTurk), a platform known for offering diverse and nationally representative samples (Chandler et al. 2019). Our sample demonstrated a robust statistical power of .9, surpassing the commonly accepted threshold of .8 (Faul et al., 2009).

Procedures and measures

After obtaining participant consent for the online experiment, we clarified that the aim was to gather feedback on the conversational style of a chatbot. We informed participants that our chatbots had been integrated into Facebook Messenger and required them to verify their Facebook login status and log into their personal Facebook accounts. Participants who did not have a Facebook account or experienced issues with the platform were redirected to a final survey page. There, we expressed our gratitude for their willingness to participate but explained that they did not meet the eligibility criteria. As a result, their data were excluded from the dataset used for analysis.

Initially, participants were directed to a Qualtrics page where, upon agreeing to the consent terms, they received an external link to Facebook Messenger to engage with our chatbots. After the interaction, at the conclusion of the conversation, the chatbots offered an additional external link that redirected participants back to Qualtrics. There, participants were asked to complete questionnaires to finalize the experiment.

Algorithmic disclosure: Prior research has utilized algorithmic disclosure manipulation by explaining to participants the mechanistic processes through which a set of algorithms operates a system (Di Porto, 2023). Similarly, Diakopoulos and Koliska (2017) advanced algorithmic disclosure by revealing details on how algorithms function and perform. In line with previous research, our study manipulated levels of algorithmic disclosure by providing descriptions of the underlying reasoning mechanisms that the prosocial behavior chatbot relies on:

The chatbot is powered by algorithms specifically tailored to encourage prosocial behavior, analyzing extensive data sets and employing machine learning techniques to enhance its ability to promote positive social actions effectively.

In contrast, participants were not provided with any information about the underlying mechanisms of how the AI chatbot encourages prosocial behaviors.

AI message type: Prior research indicates that utilitarian messages prioritize outcomes or consequences and aim to maximize overall well-being or utility. They might emphasize benefits to society or the greater good. For example, a utilitarian message might say, "Helping others improves community harmony and happiness, leading to a better society for everyone" (Conway & Gawronski, 2013). In contrast, deontological messages focus on moral rules, duties, or

obligations. They might emphasize principles like fairness, justice, or respecting individual rights. For example, a deontological message might say, "It is important to always treat others with kindness and respect, regardless of personal gain" (Conway & Gawronski, 2013). Building on these arguments, we designed two different chatbot messages focusing on the utilitarian or deontological framework. For the utilitarian framework, the chatbot highlights campaigns designed to achieve the greatest good for the greatest number of people. In contrast, for the deontological framework, the chatbot promotes campaigns that align with values and principles.

Results

Manipulation checks: Participants in the algorithmic disclosure condition (M = 5.43, SD = .94) perceived that that the chatbot disclosed more information about the mechanical processes and algorithms it uses, compared to those in the non-disclosure condition (M = 5.09, SD = 1.22; $t(488) = 3.35$, $p < .001$). Participants also evaluated the utilitarian message as being significantly more focused on maximizing outcomes rather than adhering to moral principles ($M_{utilitarian}$ = 5.52 vs. $M_{deontological}$ = 5.17, $t(270) = -3.58$, $p < .001$), while they judged the deontological message as being more centered on moral rules rather than outcomes ($M_{utilitarian}$ = 5.02 vs. $M_{deontological}$ = 5.47, $t(218)$ = 5.06, $p < .001$). Thus, all manipulations were successful.

Hypothesis testing: We ran a 2 (algorithmic disclosure: yes vs. no) × 2 (AI message type: utilitarian vs. deontological) ANOVA with WTS as a dependent variable. The results revealed a significant interaction was found between disclosure and AI message type on willingness to support a prosocial campaign (WTS; $F(1, 486)$ = 20.85, p < .001). Follow-up analyses revealed that algorithmic disclosure increased WTS when AI agents employed a utilitarian message (Myes = 5.89 vs. Mno = 5.55, $t(269) = 3.79$, $p < .001$), but decreased WTS when AI agents adopted a deontological message (Myes = 5.06 vs. Mno = 5.45, $t(217) = -2.79$, $p < .01$).

A moderated mediation model (Hayes 2017, Model 8) was estimated to test for the interaction effect of algorithmic disclosure and AI message type on WTS via self-efficacy and empathy using the bootstrapping procedure (10,000 samples). The bootstrap results confirmed a significant moderated mediation for self-efficacy (b = .08, SE = .03, CI = .02 to .15). Specifically, the results revealed that the mediating role of self-efficacy was only significant when AI agents delivered utilitarian messages with algorithmic disclosure (b = .05, SE = .02, CI = .02 to .09) rather than deontological message (b = -.02, SE = .02, CI = -.08 to .02). Regarding empathy, the bootstrap results also confirmed a significant moderated mediation (b = .08, SE = .03, CI = .02 to .15). Specifically, empathy played a significant mediating role only for when AI agents delivered deontological messages with algorithmic disclosure (b = -.09, SE = .03, CI = -.15 to -.04) rather than utilitarian (b = .02, SE = .02, CI -.02 to .07).

Discussion

This study contributes to the existing literature by exploring how the ethical principles guiding prosocial behavior—utilitarian versus deontological—interact with algorithmic disclosure to influence consumer perceptions and acceptance of chatbot prosocial recommendations. The findings suggest that the effect of the ethical framework adopted by a chatbot varies depending on whether the underlying mechanism of how the prosocial chatbot operates is revealed. Specifically, our research findings indicate that when chatbots employing the utilitarian principle disclose their underlying algorithms to promote prosocial behaviors, consumers experience enhanced self-efficacy. This increased perception of self-efficacy subsequently leads to greater acceptance of the chatbots' suggestions. Conversely, when chatbots using the deontological principle reveal their algorithms for prosocial purposes, consumers tend

to perceive lower levels of empathy, resulting in the rejection of the chatbots' recommendations.

These insights provide a nuanced understanding of the intersection between ethical principles and algorithmic disclosure in AI-mediated communication, highlighting the importance of aligning ethical frameworks with disclosure strategies to optimize consumer engagement and prosocial outcomes.

References

Anderson, S. L., & Anderson, M. (2011, August). A prima facie duty approach to machine ethics and its application to elder care. In Workshops at the twenty-fifth AAAI conference on artificial intelligence.

Bell, A., Nov, O., & Stoyanovich, J. (2023). Think about the stakeholders first! Toward an algorithmic transparency playbook for regulatory compliance. Data & Policy, 5, e12.

Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., & Litman, L. (2019). Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. Behavior Research Methods, 51, 2022-2038.

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: a process dissociation approach. Journal of Personality and Social Psychology, 104(2), 216.

Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. Cognition, 111(3), 364-371.

Cushman, F., Young, L., & Greene, J. D. (2010). Our multi-system moral psychology: Towards a consensus view. In J. M. Doris (Ed.), The Oxford handbook of moral psychology (pp. 47–71). Oxford University Press.

Di Porto, F. (2023). Algorithmic disclosure rules. Artificial Intelligence and Law, 31(1), 13-51.Diakopoulos, N., & Koliska, M. (2017). Algorithmic transparency in the news media. Digital Journalism, 5(7), 809-828.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., ... & Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. International Journal of Information Management, 57, 101994.

Efthymiou, F., & Hildebrand, C. (2023). Designing Vulnerable Conversational AI: The Impactof Trembling Voice on Empathic Concern and Prosocial Behavior. IEEE Transactions on Affective Computing, (01), 1-12.

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., … & Sandvig, C. (2015, April). "I always assumed that I wasn't really that close to [her]" Reasoning about Invisible Algorithms in News Feeds. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (pp. 153-162).

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41(4), 1149-1160.

Fischer, J. M., & Ravizza, M. (Eds.). (1992). Ethics: Problems and principles. Harcourt Brace Jovanovich College Publishers.Gawronski, B., & Beer, J. S. (2017). What makes moral dilemma judgments "utilitarian" or "deontological"? Social Neuroscience, 12(6), 626-632.

Gray, K., & Schein, C. (2012). Two minds vs. two philosophies: Mind perception defines morality and dissolves the debate between deontology and utilitarianism. Review of Philosophy and Psychology, 3, 405–423.

Greene, J., Morelli, S., Lowenberg, K., Nystrom, L., & Cohen, J. (2008). Cognitive load selectively interferes with utilitarian moral judgment. Cognition, 107(3), 1144–1154.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral

judgment. Neuron, 44(2), 389–400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. Science, 293(5537), 2105–2108.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? Trends in Cognitive Sciences, 6(12), 517–523.

Hayes, A. F. (2017). Introduction to mediation, moderation, and conditional process analysis: A regression-based approach. Guilford publications.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. Psychological Review, 108, 814–834.

Hennig, M., & Hutter, M. (2020). Revisiting the divide between deontology and utilitarianism in moral dilemma judgment: A multinomial modeling approach. Journal of Personality and Social Psychology: Attitudes and Social Cognition, 118(1), 22–56.

Kaur, G., Tomar, P., & Tanque, M. (Eds.). (2020). Artificial intelligence to solve pervasive internet of things issues. Academic Press. Khaleel, M., Ahmed, A. A., & Alsharif, A. (2023). Artificial Intelligence in Engineering. Brilliance: Research of Artificial Intelligence, 3(1), 32-42.

Kohlberg, L. (1969). Stage and sequence: The cognitive-developmental approach to socialization. In D. A. Goslin (Ed.), Handbook of socialization theory and research (pp. 347–480). Rand McNally.

Landers, R. N., & Behrend, T. S. (2023). Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. American Psychologist, 78(1), 36.

Lin, Y. T., Hung, T. W., & Huang, L. T. L. (2021). Engineering equity: How AI can help reduce the harm of implicit bias. Philosophy & Technology, 34, 65-90.

McGuire, J., Langdon, R., Coltheart, M., & Mackenzie, C. (2009). A reanalysis of the personal/impersonal distinction in moral psychology research. Journal of Experimental Social Psychology, 45(3), 581–584.

Mikhail, J. (2000). Rawls' linguistic analogy: A study of the generative grammar model of moral theory described by John Rawls in A Theory of Justice. (Unpublished doctoral dissertation). Cornell University.

Namkoong, M., Park, G., Park, Y., & Lee, S. (2023). Effect of Gratitude expression of AI Chatbot on willingness to Donate. International Journal of Human–Computer Interaction, 1-12.

Park, G., Yim, M. C., Chung, J., & Lee, S. (2023). Effect of AI chatbot empathy and identity disclosure on willingness to donate: the mediation of humanness and social presence. Behaviour & Information Technology, 42(12), 1998-2010.

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. Cognitive Science, 36, 163–177. https://doi.org/10.1111/j.1551-6709.2011.01210.x Playford, R. C., Roberts, T., & Playford, D. (2015). Deontological and utilitarian ethics: A brief introduction in the context of disorders and consciousness. Disability and Rehabilitation, 37(21), 2006-2011.

Sun, W., Bocchini, P., & Davison, B. D. (2020). Applications of artificial intelligence for disaster management. Natural Hazards, 103(3), 2631-2689.

Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility (pp. 73-100). Cambridge, MA: MacArthur Foundation Digital Media and Learning Initiative.

Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAII). Journal of Computer-Mediated Communication, 25(1), 74-88.

Thomson, J. (1985). The trolley problem. Yale Law

Journal, 94(6), 1395–1415.Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., ... & Zitnik, M. (2023). Scientific discovery in the age of artificial intelligence. Nature, 620(7972), 47-60.

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. Perspectives on Psychological Science, 5(3), 219-232.

Zarouali, B., Boerman, S. C., & de Vreese, C. H. (2021). Is this recommended by an algorithm? The development and validation of the algorithmic media content awareness scale (AMCA-scale). Telematics and Informatics, 62, 101607.

Figure 1. Conceptual design