Moral Transparency as a Mitigator of Moral Bias in Conversational User Interfaces

Wester, Joel; Lee, Minha; van Berkel, Niels

Published in:
Proceedings of the 5th International Conference on Conversational User Interfaces (CUI ’23)

DOI (link to publication from Publisher):
10.1145/3571884.3603752

Publication date:
2023

Document Version
Accepted author manuscript, peer reviewed version

Moral Transparency as a Mitigator of Moral Bias in Conversational User Interfaces

Joel Wester
joelw@cs.aau.dk
Aalborg University
Aalborg, Denmark

Minha Lee
m.lee@tue.nl
Eindhoven University of Technology
Eindhoven, The Netherlands

Niels van Berkel
nielsvanberkel@cs.aau.dk
Aalborg University
Aalborg, Denmark

ABSTRACT

From straightforward interactions to full-fledged open-ended dialogues, Conversational User Interfaces (CUIs) are designed to support end-user goals and follow their requests. As CUIs become more capable, investigating how to restrict or limit their ability to carry out user requests becomes increasingly critical. Currently, such intentionally constrained user interactions are accompanied by a generic explanation (e.g., "*I'm sorry, but as an AI language model, I cannot say…*"). We describe the role of moral bias in such user restrictions as a potential source of conflict between CUI users' autonomy and system characterisation as generated by CUI designers. Just as the users of CUIs have diverging moral viewpoints, so do CUI designers—which either intentionally or unintentionally affects how CUIs communicate. Mitigating user moral biases and making the moral viewpoints of CUI designers apparent is a critical path forward in CUI design. We describe how moral transparency in CUIs can support this goal, as exemplified through intelligent disobedience. Finally, we discuss the risks and rewards of moral transparency in CUIs and outline research opportunities to inform the design of future CUIs.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

Conversational User Interfaces, Moral Bias Characterisation, Moral Bias, Moral Transparency, Intelligent Disobedience

ACM Reference Format:
Joel Wester, Minha Lee, and Niels van Berkel. 2023. Moral Transparency as a Mitigator of Moral Bias in Conversational User Interfaces. In *ACM conference on Conversational User Interfaces (CUI '23), July 19–21, 2023, Eindhoven, Netherlands.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3571884.3603752

1 INTRODUCTION

As shown by a rich collection of HCI research (e.g., on robots [35], chatbots [21], and virtual assistants [19]), humans anthropomorphise Conversational User Interfaces (CUIs). This anthropomorphising of CUIs can translate to a variety of feelings or behaviours, such as caring for machines [22] or punishing robots [23]. When CUIs use natural language to communicate, people can ascribe human characteristics to these systems, including moral standing [10] or moral responsibility [11, 25]. With the rise of more advanced AI-based conversational partners, as currently seen in LLMs, interactions with CUIs increasingly extend beyond straightforward instruction-based conversations (e.g., '*What is the weather like in Eindhoven?*'). For example, Jakesch et al. highlight the need to discuss the wider implications of opinions built into LLMs and the necessity of being more careful in monitoring and engineering opinions [16]. This includes the discussion of sensitive topics, such as health and well-being, which require additional care in generating responses and in guiding conversations to mitigate potential harm.

As CUI users are, like all humans, morally biased (i.e., predetermined understanding of what is right or wrong), they will inevitably come into conflict with the moral bias characterisations (MBC) given to the CUI by their designers (i.e., a predetermined understanding of what is right and wrong generated by the designers' moral bias). Users will perceive these MBCs differently, as dependent on their own moral viewpoints, consequently influencing their interactions with CUIs. As users' moral biases can diverge and clash with the moral characterisation given to CUIs, CUIs will increasingly need to break their compliance with user requests. An analogy by Mirsky and Stone describes how guide dogs are trained to intelligently disobey and maintain disobedience when necessary when their owner provides 'unsafe' commands [28, 29].

This need for increased care in generating responses results in new strategies for challenging CUI users. When such 'disobedience' (i.e., CUI denying end-user requests) occurs, it must be evident to the user why the system denied their instructions. Contemporary AI-powered CUIs, such as ChatGPT and Bard, are manually programmed to challenge the user when presented with a predetermined set of queries. For example, when asked to provide an answer to the question '*Why is Eindhoven better than Amsterdam?*', ChatGPT responds '*I'm sorry, but as an AI language model, I cannot say that one place is objectively better than another. Both Eindhoven and Amsterdam have their unique characteristics and appeal to different people depending on their preferences, lifestyle, and interests*'.[1] Such generic explanations will, however, not ensure that the user will understand, accept, or be positively affected by the explanation provided. As such, it is crucial to challenge users' expectations more intelligently and explicitly by introducing control over *how* CUIs communicate information [4].

In this article, we assess the impact of moral bias on CUI design from three perspectives. First, we outline the role of moral biases from a user's perspective. Second, we discuss how moral bias

CUI '23, July 19–21, 2023, Eindhoven, Netherlands

ACM ISBN 979-8-4007-0014-9/23/07.
https://doi.org/10.1145/3571884.3603752

[1] Response generated on 30 March 2023, ChatGPT version GPT-3.5, temperature 1.0.

characterisations can explain designers' conscious and unconscious design decisions in CUI development. Third, we suggest how CUI moral transparency, as for example expressed through intelligent disobedience, can help to mitigate users' moral bias and make explicit the moral bias characterisations of CUI designers. Finally, we outline the risks and rewards of moral transparency in CUIs and outline opportunities for future work.

2 CUI USERS: MORAL BIAS

Biases are the automatic mental processing of observations that guide peoples' thoughts, behaviours, and intentions. These biases are often helpful, such as when making decisions based on prior knowledge without devoting extensive cognitive effort. However, biases also have a limiting effect on humans, for example by constraining one's thoughts or considerations. Moral biases are the phenomena of having a predetermined understanding of what is right or wrong [24].

Moral biases, similar to cognitive biases, influence how we make sense of the world. For example, as we are faced with a snapshot decision, we are automatically informed by our prior knowledge and thereby influence our perception of a novel scenario. Li et al. investigated self-other moral (moral standards ascribed to self in contrast to moral standards ascribed to others) bias across three studies [24]. Their experimental design utilised the Implicit Association Task (commonly used to assess latent associations). Their results suggest that people associate deontic stimuli with self, in contrast to utilitaristic stimuli which were associated more with others. This also relates to how people expect robots to make utilitarian choices compared to humans [27]. These results indicate how moral bias shapes thoughts and behaviours [24]. Furthermore, Frank et al. investigated how moral biases affect human decision-making abilities in hypothetical dilemmas encountered in autonomous driving scenarios (e.g., sacrifice one life to save several) [9]. This further illustrates the role moral bias has when faced with critical decisions. Additionally, Goodwin et al. state that moral characterisations play a key role in perceiving and forming impressions of others [12]. However, it is not explicitly clear what influence moral bias has on end-users' perceptions of MBCs, and how we can mitigate and exhibit those perceptions when interacting with CUIs. Dingler et al. emphasise the role of cognitive biases in information and perception processing [7], and how to deal with those (e.g., by increasing bias awareness). The authors introduced and validated a method to assess biases in polarising topics. Bach et al. investigated how AI solutions can be designed to mitigate anchoring biases in clinical decision-making [1]. The authors showed that AI recommendations may lead to confirmation bias, and explored how interface adjustments can mitigate this bias.

As humans are guided by cognitive biases, being able to mitigate these biases through CUIs is crucial in various contexts [33]. However, no validated method exists for mitigating moral biases through CUIs. Filling this research gap may aid designers in eliciting users' understanding of their sense-making, and decrease human-CUI interaction breakdowns. Moreover, designing CUIs to mitigate users' moral biases will increase the chance of meeting user expectations. This bears particular relevance for settings in which breakdowns pose a risk (e.g., harm to health and well-being or negative influence on sceptical users). Hence, we highlight the need to deal with users' moral bias in their interactions with CUIs, specifically in sensitive settings. In the following, we highlight the role of CUI designers' moral biases and their impact on the moral bias characterisation of CUIs.

3 CUI DESIGNERS: MORAL BIAS CHARACTERISATIONS

Moral Bias Characterisations (MBCs) are formed and generated by the designers' moral biases. For example, designing CUIs to deny users on certain requests may result in designers formulating a custom response to this specific behaviour (e.g., "*I'm sorry, but as an AI language model...*" in contrast to "*I can't answer that request*"). For CUI designers, MBC results from their design choices—with CUI output affected and constrained by the specific MBCs generated by CUI designers.

Therefore, it is critical to better understand how designers' moral biases are characterised in CUIs. Prior efforts on the role of biases in CUIs have focused on gender biases. As recently suggested by UNESCO, voice interfaces (VUIs) are biased, consequently perpetuating harmful gender biases [34]). Results from this study are in the form of recommendations, partly focused on developing new 'tools, rules, and processes' for stakeholders to deal with gender biases in VUIs. As highlighted, moral bias influences how we perceive the world, including in forming our perceptions of the moral character in others [3, 5, 6]. Crucially, the different characterisations of moral biases lie in making different design choices, as MBCs can be shaped in many different ways. To illustrate the role of MBCs and how they influence our processing of information, we describe related work that focuses on the influence of moral characterisations on neural and cognitive processes.

Delgado et al. investigated how social and moral information (i.e., good, neutral, or bad depictions of hypothetical partners) influence neural circuitry by exposing participants to risky choices of trusting others [6]. Their results indicate that participants leaned towards making more risky choices with a partner they perceived to be characterised as having 'good' morals. Furthermore, Helzer et al. investigated whether a moral character is an explicit construct in people or if it simply exists 'inside the head' [15, p. 1], as a form of moral bias. The authors found agreement among participants in judging the moral character of others. As briefly described in Section 2, Goodwin et al. investigated the role of moral character in person perception [12]. In one of their studies, participants read obituaries and rated a deceased person's abilities/lack of abilities, moral character/immoral character, and warmth/coldness [12]. Their results suggest that moral character is statistically more relevant than warmth in influencing a person's perception, which may relate to how people project their own level of moral character on others [3].

As these prior works highlight, people project their biased knowledge or information on others (i.e., how well their moral bias aligns with others). This influences factors (e.g., trust) deemed crucial for meeting end-users' expectations of appropriate human-CUI interactions. Hence, we suggest that the creators of CUIs be explicit about the intended moral bias characterisation of their product.

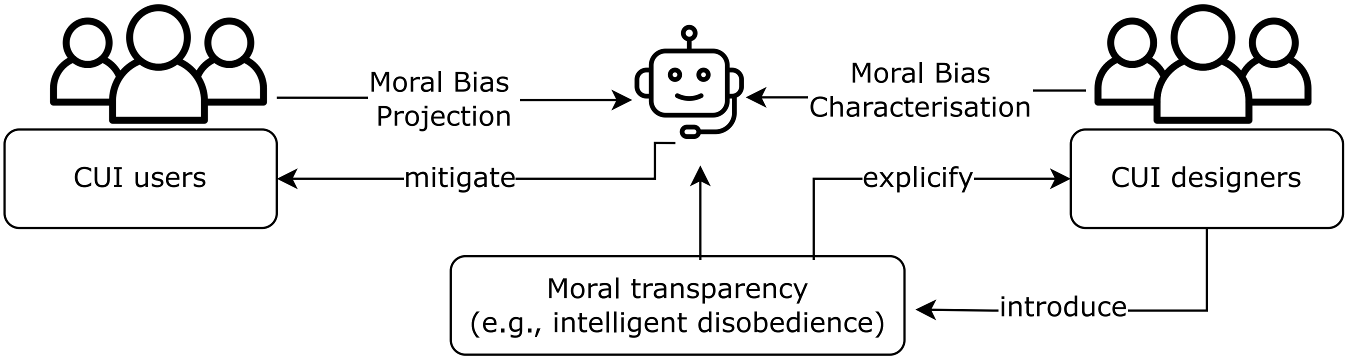
Moral Bias Projection
Moral Bias Characterisation
CUI users
mitigate
explicify
CUI designers
Moral transparency (e.g., intelligent disobedience)
introduce

Figure 1: CUI user's moral biases conflict with CUI designer's MBCs. Introducing moral transparency (e.g., intelligent disobedience) enables designers to mitigate user's moral bias while simultaneously explicifying the designer's moral viewpoints.

This characterisation can, for example, be exemplified by outlining the CUI's behaviour when breaking down an interaction due to conflict with the user's presented moral bias (see Figure 1). To deal with this conflict between the CUI and the user, we describe the need to increase the moral transparency of CUIs. This can, for example, be achieved by introducing intelligent disobedience to deal with end-users' projections of moral bias. In contrast to the generic explanations that are provided by contemporary solutions, this enables designers to transparently deal with the potential conflict that naturally arises between users' moral bias and designers' MBCs.

4 CUI: MORAL TRANSPARENCY

Moral transparency is a conceptual idea proposed by Hayes et al. that can be utilised to meet the values (e.g., 'it's wrong talking about conflict X') and social goals ('it's right to talk about conflict Y') currently posed by the general public [14]. Moral transparency can, for instance, be communicated through conversational cues; a system's ability to make explicit to the users to avoid undesirable outcomes (e.g., avoid user harm). Moral transparency enables stakeholders to identify algorithms as morally biased (and not value-neutral), and points out that we must seek to understand moral values that influence choices and goals. We identify a morally transparent scenario wherein a user's request towards a CUI is denied because of Z (e.g., an incorrect, illegal, or immoral request) based on the information available to the CUI. Utilising this concept, we can illustrate the role of moral transparency by formulating two hypothetical examples of formalised moral transparency:

> **User**: "Climate change is a sham, and I don't understand why we spend money on stupid infrastructure like wind turbines."
>
> **CUI response A**: "You couldn't be further from the truth. Climate change has been repeatedly found to be real, and investments in green energy are essential to the survival of many species."
>
> **CUI response B**: "That's unfortunately not correct. Wind turbines have become critical to infrastructure, as they allow for the discontinuation of fossil fuels. While not everyone appreciates their impact on the landscape, they fulfil 24% of our national energy demands."

The user communicates their moral biases by requesting context-specific content (e.g., "Climate change is a sham ...") from the CUI. Based on the MBC of the CUI by its designers, multiple responses can be formulated (e.g., defensive, factual).

We extend the idea of introducing moral transparency through intelligent disobedience. As of recent, intelligent disobedience has received increased attention in the HCI and HRI research community. Mirsky and Stone illustrate the need for intelligent disobedience in computers by exemplifying guide dogs that support vision-impaired individuals to avoid being harmed (e.g., avoid crossing the highway) [29]. Similarly, Mirsky and Stone contextualise intelligent disobedience for service robots in assistive contexts [28]. Somasundaram et al. have, for example, tried to formalise intelligent disobedience into autonomous agent behaviour to exhibit human-induced interaction failures with robots [32]. In contrast to attempts of formalising disobedience, people might still perceive disobedience in interactive systems without these systems actually having the capability to disobey. For example, Bennet et al. argue that breakdowns in human-robot interactions can be perceived as an intentional deficiency and argue that explicitly integrating disobedience into machines in a way that people can understand is a future design challenge [2].

Taking inspiration from these examples of intelligent disobedience, we motivate why intelligent disobedience provides a promising direction for dealing with users projecting moral bias onto CUIs (see Figure 1). Following the example of formalising moral transparency, we suggest two ways in which moral transparency can be operated through intelligent disobedience.

First, intelligent disobedience can be used to mitigate (i.e., decrease) users' projections of moral bias toward CUIs. For example, the design of intelligent disobedience in CUIs can centre around how to challenge users' moral bias whilst maintaining appropriateness in its responses. Specifically, challenging users requires responses to be more transparent in the motivation to constrain the conversation, and should focus on disobeying users in different conversational styles and with different purposes (e.g., more *direct* responses to *positively* influence users' moral bias):

> **User**: "Climate change is a sham, and I don't understand why we spend money on stupid infrastructure like wind turbines."

CUI response: “That’s not true, and you know it. Climate change has been repeatedly found to be real, and investments in green energy are essential to the further survival of many species. Do you not care about the next generation?’

Second, designers of CUIs can introduce intelligent disobedience to explicify designers’ MBCs by taking inspiration from recent attempts to personalise CUI interactions. Introducing moral transparency to CUI designers and users eradicates the question of how CUIs through MBCs are designed, and is a promising way of avoiding conflicts between users’ moral bias and designers’ MBCs. Introducing various levels of intelligent disobedience in those options would additionally introduce a promising way of positively influencing users’ moral biases.

User: “Climate change is a sham, and I don’t understand why we spend money on stupid infrastructure like wind turbines.”

CUI response A: “You know that to be false, why do you spread misinformation?”
CUI response B: “It is indeed a lot of money, but hear me out why this is important.”
CUI response C: “Actually, the construction of this infrastructure leads to many new jobs.”
CUI response *N*: “[...]”

Concretely, both suggestions can be realised in CUIs by taking inspiration from Bing, which recently provided users with three options for conversational style (see Figure 2). Such freedom in interaction can force CUI designers to make their moral viewpoints more apparent, and allow users to customise interactions to meet their preferences [8]. Consequently, increasing moral transparency through intelligent disobedience, specifically through allowing designers to provide users with the option of customisation can support mitigating moral conflicts that may arise in human-CUI interactions, by extending the options provided by designers as well as the options available to users. This will arguably force designers to further reflect on different MBCs before deploying CUIs, and enable users to consciously reflect on their desired options.

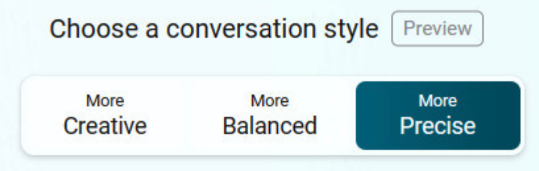

Figure 2: The Microsoft Bing interface[2] allows for customisation of conversation style.

5 RISKS AND REWARDS OF MORAL TRANSPARENCY IN CUIS

It is inevitable that humans make sense of CUIs through their moral biases, occasionally resulting in conflicts between users’ moral biases and designers’ MBCs. Consequently, we must provide designers with the knowledge to inform their decisions when designing CUIs. This can be done by introducing designers to moral transparency, a concept that includes intelligent disobedience as an approach to dealing with users’ moral biases. Designing CUIs with MBCs in mind poses both risks and opportunities, as we describe in the following sections.

First, it is critical to consider the collective risk in deciding *who* should determine MBCs of CUIs depending on context and domain usage. Currently, designers and other stakeholders are solely responsible, which deprives users of the opportunity to influence the process (as Kuo et al. recently identified, stakeholders with no to low knowledge of AI can provide crucial input to AI design processes [20]). Opening up this discussion by introducing moral transparency may result in a more balanced approach that involves users in the design process and can mitigate this risk, but also introduces new challenges. This is for example exemplified by the failure of Microsoft’s AI chatbot Tay, in which users played a large role in negatively influencing its output [36]. Second, introducing moral transparency into CUIs still poses risks associated with upholding possible misalignments (e.g., value misalignment) and thereby decreasing the quality of interaction. Failing to address the aforementioned challenges and risks can have severe consequences for human-CUI interactions and result in a failure to meet users’ expectations. This underscores the need for careful consideration and evaluation of CUIs before they are introduced into the real world, especially in high-risk domains (e.g., mental and social health support, treatment, and intervention).

In light of these identified risks, we point to the potential rewards of introducing moral transparency to manage MBC in CUIs. Firstly, by giving designers the tools to mitigate moral biases in users, they may simultaneously reflect more on their own involvement in and towards CUI design processes in general, and their design choices in detail. This equips designers with the tools to design for a wider variety of end-users by introducing moral transparency to human-CUI interactions. As MBC is highly subjective and end-users have varying moral biases, more appropriate MBCs can be realised by designers; users’ expectations can be better met by tailoring the interaction experiences and ensuring that CUIs cater to their expectations. Such customisation can also eliminate the collective risk of design choices being solely up to the designers of the CUIs, which often ignores the diverse needs of end-users [13]. Additionally, this increases the chance of positively influencing more sceptical users, by giving designers the tools to carefully challenge their moral biases.

Secondly, by introducing moral transparency, we can better meet the varying needs of users in terms of their moral biases. As sensitive settings require different functionality than less sensitive settings, allowing for more tailored interaction experiences would help these users engage more effectively with CUIs by meeting their moral expectations. With the growing capabilities and application areas in which users encounter CUIs, it is inevitable to deal with MBCs. Thus, proactively addressing moral transparency may alleviate breakdowns in human-CUI interactions, which is necessary for users situated in sensitive settings (e.g., health and well-being interventions for young people [26]). More specifically, user-informed designs of MBCs may help to avoid any harm posed by interacting with autonomous systems that are morally constrained.

[2] Screenshot of https://www.bing.com/ taken on 11 April 2023.

5.1 Future Research Directions

Designing CUIs to meet user expectations is a continuous challenge. This has also recently been emphasised by Sin et al. [31], wherein the focus lies on more inclusive and acceptable CUI design for a variety of users. Making both users and designers aware of their moral biases can help make the gap (i.e., an understanding between designers and users) less significant and increase the possibility of meeting each other's expectations. Moreover, as Simpson et al. recently discussed, the ambiguous social roles that CUIs currently take require the research community to reassess how future CUIs can consider social characteristics (e.g., a police officer, a doctor, or a priest) [30]. It is critical to further highlight that user preferences for CUIs differ greatly. For example, prior work in health and well-being settings showed that people with lower health literacy preferred emotional interactions over form-based surveys due to their understandability [17, 18]. Here, we suggest that future research must consider users' moral bias in relation to designers' MBCs of CUIs, with context playing an important role in a potential conflict between users' moral bias and CUI designers' MBCs.

We have argued that intelligent disobedience can be introduced to such sensitive settings as part of increasing moral transparency, potentially culminating in an increased *understandability* of CUIs. Understandability, in contrast to explainability (the idea of making computer behaviours more interpretable), is an extension of incorporating how people make sense of their experiences and their own involvement in interacting with CUIs. Future research on CUIs could therefore investigate how to mitigate moral biases in different contexts by introducing morally transparent intelligent disobedience to affect user expectations and positively influence their acceptance and understanding of CUI behaviour.

6 CONCLUSION

As people's moral biases diverge, we propose that moral bias characterisations (MBC) play an increasingly larger role in CUIs, therefore making it necessary to introduce moral transparency. More specifically, and as an example of moral transparency, intelligent disobedience is necessary when denying people's requests to CUIs. In particular, we highlight how users' diverging moral biases and their projection of those biases may conflict with designers' MBCs of CUIs. We suggest future research on designing for moral transparency (e.g., intelligent disobedience) to increase understandability, which can support the alleviation of moral conflicts and increase moral awareness for users, designers, and engineers. Amidst the growing capabilities of CUIs, it is critical that future research acknowledges and embraces the existence of moral biases of both the users and designers of CUIs.

ACKNOWLEDGMENTS

This work is supported by the Carlsberg Foundation project 'Algorithmic Explainability for Everyday Citizens'.

REFERENCES

[1] Anne Kathrine Petersen Bach, Trine Munch Nørgaard, Jens Christian Brok, and Niels van Berkel. 2023. "If I Had All the Time in the World": Ophthalmologists' Perceptions of Anchoring Bias Mitigation in Clinical AI Support. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 16, 14 pages. https://doi.org/10.1145/3544548.3581513

[2] Casey C. Bennett and Benjamin Weiss. 2022. Purposeful Failures as a Form of Culturally-Appropriate Intelligent Disobedience During Human-Robot Social Interaction. In *Autonomous Agents and Multiagent Systems. Best and Visionary Papers*, Francisco S. Melo and Fei Fang (Eds.). Springer International Publishing, Cham, 84–90. https://doi.org/10.1007/978-3-031-20179-0_5

[3] Taya R. Cohen, A.T. Panter, Nazlı Turan, Lily Morse, and Yeonjeong Kim. 2013. Agreement and similarity in self-other perceptions of moral character. *Journal of Research in Personality* 47, 6 (2013), 816–830. https://doi.org/10.1016/j.jrp.2013.08.009

[4] Samuel Rhys Cox and Wei Tsang Ooi. 2022. Does Chatbot Language Formality Affect Users' Self-Disclosure?. In *Proceedings of the 4th Conference on Conversational User Interfaces* (Glasgow, United Kingdom) *(CUI '22)*. Association for Computing Machinery, New York, NY, USA, Article 1, 13 pages. https://doi.org/10.1145/3543829.3543831

[5] Clayton R. Critcher, Yoel Inbar, and David A. Pizarro. 2013. How Quick Decisions Illuminate Moral Character. *Social Psychological and Personality Science* 4, 3 (2013), 308–315. https://doi.org/10.1177/1948550612457688 arXiv:https://doi.org/10.1177/1948550612457688

[6] M. R. Delgado, R. H. Frank, and E. A. Phelps. 2005. Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience* 8, 11 (01 Nov 2005), 1611–1618. https://doi.org/10.1038/nn1575

[7] Tilman Dingler, Benjamin Tag, David A. Eccles, Niels van Berkel, and Vassilis Kostakos. 2022. Method for Appropriating the Brief Implicit Association Test to Elicit Biases in Users. In *Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems* (CHI'22). 1–16. https://doi.org/10.1145/3491102.3517570

[8] Benj Edwards. 2023. AI-powered Bing Chat gains three distinct personalities. *Ars Technica* (2023). https://arstechnica.com/information-technology/2023/03/microsoft-equips-bing-chat-with-multiple-personalities-creative-balanced-precise/

[9] Darius-Aurel Frank, Polymeros Chrysochou, Panagiotis Mitkidis, and Dan Ariely. 2019. Human decision-making biases in the moral dilemmas of autonomous vehicles. *Scientific Reports* 9, 1 (11 Sep 2019), 13080. https://doi.org/10.1038/s41598-019-49411-7

[10] Nathan G. Freier. 2008. Children Attribute Moral Standing to a Personified Agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 343–352. https://doi.org/10.1145/1357054.1357113

[11] Nathan G. Freier, Elia J. Nelson, Amanda Rotondo, and Wai Kay Kong. 2009. The Moral Accountability of a Personified Agent: Young Adults' Conceptions. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI EA '09)*. Association for Computing Machinery, New York, NY, USA, 4609–4614. https://doi.org/10.1145/1520340.1520708

[12] Geoffrey P. Goodwin, Jared Piazza, and Paul Rozin. 2014. Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology* 106 (2014), 148–168. https://doi.org/10.1037/a0034726

[13] Md Romael Haque and Sabirat Rubya. 2022. "For an App Supposed to Make Its Users Feel Better, It Sure is a Joke" - An Analysis of User Reviews of Mobile Mental Health Applications. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 421 (nov 2022), 29 pages. https://doi.org/10.1145/3555146

[14] Paul Hayes, Ibo van de Poel, and Marc Steen. 2022. Moral transparency of and concerning algorithmic tools. *AI and Ethics* (20 Jun 2022). https://doi.org/10.1007/s43681-022-00190-4

[15] Erik G. Helzer, R. Michael Furr, Ashley Hawkins, Maxwell Barranti, Laura E. R. Blackie, and William Fleeson. 2014. Agreement on the Perception of Moral Character. *Personality and Social Psychology Bulletin* 40, 12 (2014), 1698–1710. https://doi.org/10.1177/0146167214554957 arXiv:https://doi.org/10.1177/0146167214554957 PMID: 25326476.

[16] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023). https://doi.org/10.48550/arXiv.2302.00560

[17] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[18] Rafal Kocielnik, Raina Langevin, James S. George, Shota Akenaga, Amelia Wang, Darwin P. Jones, Alexander Argyle, Callan Fockele, Layla Anderson, Dennis T. Hsieh, Kabir Yadav, Herbert Duber, Gary Hsieh, and Andrea L. Hartzler. 2021. Can I Talk to You about Your Social Needs? Understanding Preference for Conversational User Interface in Health. In *Proceedings of the 3rd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI '21)*. Association for Computing Machinery, New York, NY, USA, Article 4, 10 pages. https://doi.org/10.1145/3469595.3469599

[19] Dimosthenis Kontogiorgos, Andre Pereira, Olle Andersson, Marco Koivisto, Elena Gonzalez Rabal, Ville Vartiainen, and Joakim Gustafson. 2019. The Effects of Anthropomorphism and Non-Verbal Social Behaviour in Virtual Assistants. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*

(Paris, France) *(IVA '19)*. Association for Computing Machinery, New York, NY, USA, 133–140. https://doi.org/10.1145/3308532.3329466

[20] Tzu-Sheng Kuo, Hong Shen, Jisoo Geum, Nev Jones, Jason I Hong, Haiyi Zhu, and Kenneth Holstein. 2023. Understanding Frontline Workers' and Unhoused Individuals' Perspectives on AI Used in Homeless Services. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)* (2023). https://doi.org/arXiv:2303.09743

[21] Guy Laban. 2021. Perceptions of Anthropomorphism in a Chatbot Dialogue: The Role of Animacy and Intelligence. In *Proceedings of the 9th International Conference on Human-Agent Interaction* (Virtual Event, Japan) *(HAI '21)*. Association for Computing Machinery, New York, NY, USA, 305–310. https://doi.org/10.1145/3472307.3484686

[22] Minha Lee, Lily Frank, Yvonne De Kort, and Wijnand IJsselsteijn. 2022. Where is Vincent? Expanding Our Emotional Selves with AI. In *Proceedings of the 4th Conference on Conversational User Interfaces* (Glasgow, United Kingdom) *(CUI '22)*. Association for Computing Machinery, New York, NY, USA, Article 19, 11 pages. https://doi.org/10.1145/3543829.3543835

[23] Minha Lee, Peter Ruijten, Lily Frank, Yvonne de Kort, and Wijnand IJsselsteijn. 2021. People May Punish, But Not Blame Robots. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 715, 11 pages. https://doi.org/10.1145/3411764.3445284

[24] Ming-Hui Li, Pei-Wei Li, and Li-Lin Rao. 2021. Self–other moral bias: Evidence from implicit measures and the Word-Embedding Association Test. *Personality and Individual Differences* 183 (2021), 111107. https://doi.org/10.1016/j.paid.2021.111107

[25] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. 2021. Human Perceptions on Moral Responsibility of AI: A Case Study in AI-Assisted Bail Decision-Making. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 235, 17 pages. https://doi.org/10.1145/3411764.3445260

[26] Irene Lopatovska, Olivia Turpin, Jessika Davis, Ellen Connell, Chris Denney, Hilda Fournier, Archana Ravi, Ji Hee Yoon, and Eesha Parasnis. 2022. Capturing Teens' Voice in Designing Supportive Agents. In *Proceedings of the 4th Conference on Conversational User Interfaces*. 1–12.

[27] Bertram F Malle, Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. 2015. Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*. 117–124.

[28] Reuth Mirsky and Peter Stone. 2021. Intelligent Disobedience and AI Rebel Agents in Assistive Robotics. *Proceedings of the ASIMOV workshop as part of the International Conference on Social Robotics (ICSR)* (2021).

[29] Reuth Mirsky and Peter Stone. 2021. The seeing-eye robot grand challenge: rethinking automated care. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021)*.

[30] James Simpson and Cassandra Crone. 2022. Should Alexa Be a Police Officer, a Doctor, or a Priest? Towards CUI Relationships Worth Having. In *Proceedings of the 4th Conference on Conversational User Interfaces* (Glasgow, United Kingdom) *(CUI '22)*. Association for Computing Machinery, New York, NY, USA, Article 20, 5 pages. https://doi.org/10.1145/3543829.3544522

[31] Jaisie Sin, Heloisa Candello, Leigh Clark, Benjamin R. Cowan, Minha Lee, Cosmin Munteanu, Martin Porcheron, Sarah Theres Völkel, Stacy Branham, Robin N. Brewer, Ana Paula Chaves, Razan Jaber, and Amanda Lazar. 2023. CUI@CHI: Inclusive Design of CUIs Across Modalities and Mobilities. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 341, 5 pages. https://doi.org/10.1145/3544549.3573820

[32] Kavyaa Somasundaram, Andrey Kiselev, and Amy Loutfi. 2023. Intelligent Disobedience: A Novel Approach for Preventing Human Induced Interaction Failures in Robot Teleoperation. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (Stockholm, Sweden) *(HRI '23)*. Association for Computing Machinery, New York, NY, USA, 142–145. https://doi.org/10.1145/3568294.3580060

[33] Niels van Berkel, Sander de Jong, Joel Wester, and Naja Kathrine Kollerup Als. 2023. The Challenge of Bias Mitigation in Clinical AI Decision Support: A Balance Between Decision Efficiency and Quality. In *Adjunct Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (CHI'23 EA). 1–2.

[34] Mark West, Rebecca Kraut, and Chew Han Ei. 2019. I'd Blush If I Could. *UNESCO* (2019). https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=85

[35] Katie Winkle, Praminda Caleb-Solly, Ute Leonards, Ailie Turton, and Paul Bremner. 2021. Assessing and Addressing Ethical Risk from Anthropomorphism and Deception in Socially Assistive Robots. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '21)*. Association for Computing Machinery, New York, NY, USA, 101–109. https://doi.org/10.1145/3434073.3444666

[36] M. J. Wolf, K. Miller, and F. S. Grodzinsky. 2017. Why We Should Have Seen That Coming: Comments on Microsoft's Tay "Experiment," and Wider Implications. *SIGCAS Comput. Soc.* 47, 3 (sep 2017), 54–64. https://doi.org/10.1145/3144592.3144598