RESPONSIBLE ROBOTICS: IDENTIFYING AND ADDRESSING ISSUES OF ETHICS, FAIRNESS, ACCOUNTABILITY, TRANSPARENCY, PRIVACY AND EMPLOYMENT

EDITED BY: Martim Brandão, Martin Magnusson and Masoumeh Mansouri
PUBLISHED IN: Frontiers in Robotics and AI

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88976-573-7
DOI 10.3389/978-2-88976-573-7

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

RESPONSIBLE ROBOTICS: IDENTIFYING AND ADDRESSING ISSUES OF ETHICS, FAIRNESS, ACCOUNTABILITY, TRANSPARENCY, PRIVACY AND EMPLOYMENT

Topic Editors:
Martim Brandão, King's College London, United Kingdom
Martin Magnusson, Örebro University, Sweden
Masoumeh Mansouri, University of Birmingham, United Kingdom

Citation: Brandão, M., Magnusson, M., Mansouri, M., eds. (2022). Responsible Robotics: Identifying and Addressing Issues of Ethics, Fairness, Accountability, Transparency, Privacy and Employment. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88976-573-7

Table of Contents

frontiers | Frontiers in Robotics and AI

EDITORIAL
published: 21 June 2022
doi: 10.3389/frobt.2022.937612

Editorial: Responsible Robotics

Martim Brandão[1]*, *Masoumeh Mansouri*[2] *and Martin Magnusson*[3]

[1]Department of Informatics, King's College London, London, United Kingdom, [2]School of Computer Science, University of Birmingham, Birmingham, United Kingdom, [3]School of Science and Technology (AASS), Örebro University, Örebro, Sweden

Keywords: robotics, responsible innovation, responsible robotics, trustworthy robotics, critical robotics, AI and society, robot ethics

Editorial on the Research Topic

Responsible Robotics: Identifying and Addressing Issues of Ethics, Fairness, Accountability, Transparency, Privacy and Employment

1 RESPONSIBLE AI AND ROBOTICS

Recent work in both academia, industry, and journalism has brought widespread attention to various kinds of harmful impact that AI can have on society. These are very often concentrated on marginalized social groups. AI algorithms may unintentionally reinforce social prejudice Bolukbasi et al. (2016) and biased conceptions of gender Adams and Loideáin (2019); Hamidi et al. (2018), race Sweeney (2013), age Rosales and Fernández-Ardávol (2019) or disabilities Guo et al. (2020), they may lead to unfair access to opportunities Dastin (2018); Angwin et al. (2016), discriminatory pricing practices Bar-Gill (2019); Hannak et al. (2014), etc. Recent work has also shown that many seemingly technical issues in machine learning are actually socio-technical. For example: the over-fitting of machine learning models, the choice of dataset or learning objective, and other aspects of learning may lead to algorithms performing poorly on unrepresented or unmodeled groups of people Brandao (2019); Barocas et al. (2019); Buolamwini and Gebru (2018). A growing community of Fairness, Accountability, Transparency, and Ethics of AI[1] is now approaching these Research Topic from a socio-technical point-of-view, in order to identify, understand, and alleviate such issues.

Robotics, as a technology focused on automation and intelligent behavior, also abounds in similar ethical and social issues that need to be identified, characterized, and considered in design. While many of the same problems with AI will also be present in robotics, the physical nature of robotics raises new aspects of the social and ethical nature of these technologies. As one example: models that are considerably less accurate on certain groups of people can lead to physical safety differentials Brandao (2019), where robots or autonomous vehicles using those models are more likely to collide with those groups. Additionally, there are physical safety concerns with respect to surgical and other medical robots Yang et al. (2017); Ficuciello et al. (2019), as well as concerns of physical and political security—not least concerning autonomous weapon systems and the dual-use of robot technologies like autonomous cars and drones Brundage et al. (2018); Sparrow (2007).

The physical design and visual appearance of robots also introduce new aspects to responsible development. For example, people's moral evaluation of robot decisions can be affected by whether the robot is more or less human-like Malle and Scheutz (2016), the design of robots in a care setting

[1]Example venues: ACM Conference on Fairness, Accountability, and Transparency (FAccT), AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES).

OPEN ACCESS

Edited and reviewed by:
Bertram F. Malle,
Brown University, United States

***Correspondence:**
Martim Brandão
martim.brandao@kcl.ac.uk

Specialty section:
This article was submitted to Ethics in Robotics and Artificial Intelligence, a section of the journal Frontiers in Robotics and AI

Received: *06 May 2022*
Accepted: *30 May 2022*
Published: *21 June 2022*

Citation:
Brandão M, Mansouri M and Magnusson M (2022) Editorial: Responsible Robotics. Front. Robot. AI 9:937612. doi: 10.3389/frobt.2022.937612

affects caregivers and caretakers van Wynsberghe (2021); Kubota et al. (2021), the choice of sensors, measurements and motion has an impact of privacy Calo (2011); Eick and Antón (2020); Luo et al. (2020), and the ethics of deception takes on new shape Danaher (2020).

The robotics community has been discussing ethics for long[2]. Recent workshops have also started bringing attention to philosophical problems in robotics[3] and issues such as bias[4] and transparency[5]. These efforts share a common goal of developing robotics technologies responsibly—they are part of "Responsible Robotics" or "Trustworthy Robotics."

A similar effort on "Critical Robotics" Serholt et al. (2021) has focused on questioning current practices in robotics research. These range from how older adults are represented in HRI Burema (2021) and ethical issues in education robots Serholt et al. (2017), to normative dimensions of speech used by researchers Brandao (2021), their technological optimism Šabanović (2010) and the influence of their social background in research directions Forsythe (2001); Šabanović (2010).

2 THIS RESEARCH TOPIC

This Research Topic gathers a diverse set of articles on Responsible Robotics. They range from user studies and philosophical inquiry, to modeling, algorithmic, and governance methods. Our goal when organizing this Research Topic was exactly to join various approaches in a single edition—to allow for greater multidisciplinary exchange under the common mission of Responsible Robotics. We believe that Responsible Robotics should focus both on *identifying* social and ethical issues, and on *designing* methods to account for (and alleviate) such issues—thus the focus of this edition on both understanding and *acting* on social and ethical issues.

Two articles in the Research Topic are focused on eliciting social and ethical issues *from users and stakeholders*. Lutz and Tamò-Larrieux investigate privacy concerns of lay users and their impact on technology use intentions, when using social robots that are either privacy-friendly or privacy-invasive (e.g., listen to conversations, share data with third parties). Colombino et al. use ethnographic studies, interviews and futuristic autobiographies to identify organizational principles, potential roles, and ethical design considerations for a robot that collaborates with disabled employees.

Three articles are more focused on methods, or socio-technical solutions to ethical problems in robotics. Webb et al., for example, focus on methods for conducting investigations of accidents involving humans and robots. In particular, they propose and preliminarily evaluate a role-play-based methodology for investigating accidents, and to evaluate the testimonies that humans can give in forensic investigations of such accidents. Hurtado et al. focus on issues of harmful social bias in robot learning and how they could be detected and alleviated. Namely, they show through various examples how social robot navigation techniques that mimic human behavior may lead to harmful behavior, such as higher intrusion of personal space or longer waiting times for some groups compared to others. Winfield et al. focus on issues of transparency from a governance perspective. They describe a new draft standard on transparency for autonomous systems, with several contributions such as transparency levels, measurability, stakeholders, and example-based guidance on using the draft standard.

We then dive into philosophical inquiry and frameworks for robot ethics. Rhim et al. combine work in moral philosophy and psychology to propose a model that explains human decision-making in moral dilemmas involving autonomous vehicles. Pirni et al. consider aspects of autonomy and vulnerability in the ethics of designing care robots. And Kuipers argues that AI and robotics technologies rely heavily on over-simplified models, and that the widespread use of such models can lead to the erosion of trust and cooperation effectiveness. The article can serve as an argument for why more attention should be given to the *modeling* of complex socio-technical factors in AI/robotics.

Finally, two articles in the Research Topic dive into issues of jobs and economics in robotics and automation. Studley argues that we should consider how robotics impacts global supply chains, international development, and global economic disparities. Kyvik Nordås and Klügl then use modeling to understand the uptake of automation technologies and its relationship with unemployment and engineering, consultancy, and manufacturing jobs. The authors use this analysis to suggest an automation policy focus on user costs and education.

We believe that the contributions collected in this Research Topic can be relevant to roboticists, AI practitioners, policy makers and any other stakeholders concerned with the societal impacts of AI and robotics. We hope this Research Topic will stimulate future work on responsible robotics.

We end with an important remark. While the abundance of social and ethical issues raised in this editorial and this Research Topic might feel overwhelming or hopeless, we believe the opposite is the case. Responsible Robotics is about clearly identifying potential issues, because by doing so it is also possible to work towards responsible methods that mitigate them. This ultimately facilitates the application of robotics and AI in ways that increase safety, efficiency, and wellbeing in many areas of life: transportation, healthcare, work life, just to name a few.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

[2]ICRA 2007/2009/2011 workshops on Roboethics, ICRA 2014 workshop on "Robotics and Military Applications".

[3]Robophilosophy Conference.

[4]ICRA 2019 workshops on "Bias-sensitizing robot behaviours" and "Unlearning biases in robot design".

[5]HRI 2022 workshop on "Fairness and Transparency in HRI," ICRA 2020 workshop "Against robot dystopias".

REFERENCES

Adams, R., and Loideáin, N. N. (2019). Addressing Indirect Discrimination and Gender Stereotypes in Ai Virtual Personal Assistants: the Role of International Human Rights Law. *Camb. Int. Law J.* 8, 241–257. doi:10.4337/cilj.2019.02.04

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). *Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks*. Wilmette, IL: Benton Institute for Broadband & Society.

Bar-Gill, O. (2019). *Algorithmic Price Discrimination when Demand Is a Function of Both Preferences and (Mis) Perceptions*. Chicago, Illinois: University of Chicago Law Review, 86.

Barocas, S., Hardt, M., and Narayanan, A. (2019). Fairness and Machine Learning (fairmlbook.Org). Available at: http://www.fairmlbook.org.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings. *Adv. neural Inf. Process. Syst.* 29.

Brandao, M. (2019). "Age and Gender Bias in Pedestrian Detection Algorithms," in Workshop on Fairness Accountability Transparency and Ethics in Computer Vision, CVPR.

Brandao, M. (2021). "Normative Roboticists: the Visions and Values of Technical Robotics Papers," in IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 671–677. doi:10.1109/RO-MAN50785.2021.9515504

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., et al. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *arXiv Prepr. arXiv:1802.07228*.

Buolamwini, J., and Gebru, T. (2018). "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. *Vol. 81 of Proceedings of Machine Learning Research*. Editors S. A. Friedler and C. Wilson (New York, NY, USA: PMLR), 77–91.

Burema, D. (2021). A Critical Analysis of the Representations of Older Adults in the Field of Human–Robot Interaction. *AI Soc.* 2021, 1–11.

Calo, R. (2011). "Robots and Privacy," in *Robot Ethics: The Ethical and Social Implications of Robotics*.

Danaher, J. (2020). Robot Betrayal: a Guide to the Ethics of Robotic Deception. *Ethics Inf. Technol.* 22, 117–128. doi:10.1007/s10676-019-09520-3

Dastin, J. (2018). "Amazon Scraps Secret Ai Recruiting Tool that Showed Bias against Women," in *Ethics of Data and Analytics* (Boca Raton, Fla: Auerbach Publications), 296–299.

Eick, S., and Anton, A. I. (2020). "Enhancing Privacy in Robotics via Judicious Sensor Selection," in 2020 IEEE International Conference on Robotics and Automation (ICRA), 7156–7165. doi:10.1109/ICRA40945.2020.9196983

Ficuciello, F., Tamburrini, G., Arezzo, A., Villani, L., and Siciliano, B. (2019). Autonomy in Surgical Robots and its Meaningful Human Control. *Paladyn, J. Behav. Robotics* 10, 30–43. doi:10.1515/pjbr-2019-0002

Forsythe, D. (2001). *Studying Those Who Study Us: An Anthropologist in the World of Artificial Intelligence*. Redwood City, California: Stanford University Press.

Guo, A., Kamar, E., Vaughan, J. W., Wallach, H., and Morris, M. R. (2020). Toward Fairness in AI for People with Disabilities SBG@a Research Roadmap. *SIGACCESS Access. Comput.* 2020, 1. doi:10.1145/3386296.3386298

Hamidi, F., Scheuerman, M. K., and Branham, S. M. (2018). "Gender Recognition or Gender Reductionism? the Social Implications of Embedded Gender Recognition Systems," in Proceedings of the 2018 chi conference on human factors in computing systems, 1–13.

Hannak, A., Soeller, G., Lazer, D., Mislove, A., and Wilson, C. (2014). "Measuring Price Discrimination and Steering on E-Commerce Web Sites," in Proceedings of the 2014 conference on internet measurement conference, 305–318. doi:10.1145/2663716.2663744

Kubota, A., Pourebadi, M., Banh, S., Kim, S., and Riek, L. (2021). Somebody that I Used to Know: The Risks of Personalizing Robots for Dementia Care. *Proc. We Robot*.

Luo, Y., Yu, Y., Jin, Z., Li, Y., Ding, Z., Zhou, Y., et al. (2020). "Privacy-aware Uav Flights through Self-Configuring Motion Planning," in 2020 IEEE International Conference on Robotics and Automation (ICRA), 1169–1175. doi:10.1109/ICRA40945.2020.9197564

Malle, B. F., and Scheutz, M. (2016). "Inevitable Psychological Mechanisms Triggered by Robot Appearance: Morality Included?," in 2016 AAAI Spring Symposium Series.

Rosales, A., and Fernández-Ardèvol, M. (2019). Structural Ageism in Big Data Approaches. *Nord. Rev.* 40, 51–64. doi:10.2478/nor-2019-0013

Šabanović, S. (2010). Robots in Society, Society in Robots. *Int. J. Soc. Robotics* 2, 439–450.

Serholt, S., Barendregt, W., Vasalou, A., Alves-Oliveira, P., Jones, A., Petisca, S., et al. (2017). The Case of Classroom Robots: Teachers' Deliberations on the Ethical Tensions. *AI Soc* 32, 613–631. doi:10.1007/s00146-016-0667-2

Serholt, S., Ljungblad, S., and Ní Bhroin, N. (2021). Introduction: Special Issue—Critical Robotics Research. *AI Soc.* 2021, 1–7.

Sparrow, R. (2007). Killer Robots. *J. Appl. Philos.* 24, 62–77. doi:10.1111/j.1468-5930.2007.00346.x

Sweeney, L. (2013). Discrimination in Online Ad Delivery. *Commun. ACM* 56, 44–54. doi:10.1145/2447976.2447990

van Wynsberghe, A. (2021). Social Robots and the Risks to Reciprocity. *AI Soc.* 2021, 1–7. doi:10.1007/s00146-021-01207-y

Yang, G.-Z., Cambias, J., Cleary, K., Daimler, E., Drake, J., Dupont, P. E., et al. (2017). Medical Robotics-Regulatory, Ethical, and Legal Considerations for Increasing Levels of Autonomy. *Sci. Robot.* 2, eaam8638. doi:10.1126/scirobotics.aam8638

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Robotics and AI

HYPOTHESIS AND THEORY
published: 24 March 2021
doi: 10.3389/frobt.2021.650325

From Learning to Relearning: A Framework for Diminishing Bias in Social Robot Navigation

Juana Valeria Hurtado[*†], *Laura Londoño*[†] *and Abhinav Valada*

Department of Computer Science, University of Freiburg, Freiburg im Breisgau, Germany

The exponentially increasing advances in robotics and machine learning are facilitating the transition of robots from being confined to controlled industrial spaces to performing novel everyday tasks in domestic and urban environments. In order to make the presence of robots safe as well as comfortable for humans, and to facilitate their acceptance in public environments, they are often equipped with social abilities for navigation and interaction. Socially compliant robot navigation is increasingly being learned from human observations or demonstrations. We argue that these techniques that typically aim to mimic human behavior do not guarantee fair behavior. As a consequence, social navigation models can replicate, promote, and amplify societal unfairness, such as discrimination and segregation. In this work, we investigate a framework for diminishing bias in social robot navigation models so that robots are equipped with the capability to plan as well as adapt their paths based on both physical and social demands. Our proposed framework consists of two components: *learning* which incorporates social context into the learning process to account for safety and comfort, and *relearning* to detect and correct potentially harmful outcomes before the onset. We provide both technological and societal analysis using three diverse case studies in different social scenarios of interaction. Moreover, we present ethical implications of deploying robots in social environments and propose potential solutions. Through this study, we highlight the importance and advocate for fairness in human-robot interactions in order to promote more equitable social relationships, roles, and dynamics and consequently positively influence our society.

Keywords: social robot navigation, robot learning, fairness-aware learning, algorithmic fairness, ethics, responsible innovation

OPEN ACCESS

Edited by:
Martim Brandão,
King's College London,
United Kingdom

Reviewed by:
Pablo Jiménez-Schlegl,
Consejo Superior de Investigaciones Científicas (CSIC), Spain
Helena Webb,
University of Oxford, United Kingdom

***Correspondence:**
Juana Valeria Hurtado
hurtadoj@cs.uni-freiburg.de

[†]*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Ethics in Robotics and Artificial Intelligence, a section of the journal Frontiers in Robotics and AI

Received: *06 January 2021*
Accepted: *01 March 2021*
Published: *24 March 2021*

Citation:
Hurtado JV, Londoño L and Valada A (2021) From Learning to Relearning: A Framework for Diminishing Bias in Social Robot Navigation. Front. Robot. AI 8:650325. doi: 10.3389/frobt.2021.650325

1. INTRODUCTION

The last decade has brought numerous breakthroughs in the development of autonomous robots which is evident from the manufacturing and service industries. More interesting are the advances that are essential enablers of several innovative applications, such as robot-assisted surgery (Tewari et al., 2002), transportation (Thrun, 1995), environmental monitoring (Valada et al., 2012), planetary exploration (Toupet et al., 2020), and disaster relief (Mittal et al., 2019). Novel machine learning algorithms accompanied by the boost in computational capacity and availability of large annotated datasets have primarily fostered the progress in this field. Machine learning and reinforcement learning techniques enable robots to learn complex tasks directly from raw sensory

input. One such task of navigation has seen tremendous progress over the years. Robots today have the capability to autonomously plan paths to reach a certain location and even make decisions based on the scene dynamics, avoiding collisions with people and objects (Boniardi et al., 2016; Gaydashenko et al., 2018; Jamshidi et al., 2019; Hurtado et al., 2020). Advancing robot navigation abilities is crucial for robots to effectively operate in real-world environments.

Robot navigation is a complex task that requires a high degree of autonomy. For a robot to successfully navigate the real-world, it is essential to fulfill high accuracy, efficacy, and efficiency requirements. Additionally, it is critical to consider safety standards while developing robots that navigate around humans. To carry out this task, robots are equipped with sensors that allow them to perceive the environment and a path planning system that enables them to compute a feasible route to achieve the navigation goal. So far, mobile robots have been successfully employed in various applications, such as material transportation, patrolling, rescue operation, cleaning, guidance, warehouse automation, among others (Nolfi and Floreano, 2002; Poudel, 2013; Hasan et al., 2014; Bogue, 2016). This also elucidates that mobile robot applications are moving closer from the industry to everyday tasks in households, offices, and public spaces. Robot navigation models tailored to solely reach a goal location efficiently are insufficient in these spaces where robots cohabitate with humans. Other complex considerations, such as social context, norms, and conventions are essential to ensure that the presence and movements of robots are safe and comfortable. These additional considerations of sociability play an indispensable role in the acceptance of robots in human spaces. Nevertheless, modeling the social policies that represent humans is a challenging task. To better capture the social behavior of navigation, several learning approaches have been proposed with the goal of directly imitating human navigation or learning from demonstrations (Silver et al., 2010; Wittrock, 2010; Bicchi and Tamburrini, 2015; Khambhaita and Alami, 2020). With the aim of incorporating social context in learning algorithms, socially-aware robot navigation extends the traditional objective of reaching a certain location to also reflect social behavior in the decision making process (Kretzschmar et al., 2016). This can be achieved with learning methodologies based on social and cultural norms. These social characteristics can be incorporated into the learning process as social constraints (Wittrock, 2010; Bicchi and Tamburrini, 2015; Khambhaita and Alami, 2020) or via imitation and demonstrations (Silver et al., 2010). As the role of robots within society is that of a social agent, they should follow social conventions for better acceptability in human environments. Following such conventions will enable them to generate actions that are influenced by respecting personal spaces, perceiving emotions, gestures, and expressions (Luber et al., 2012; Ferrer et al., 2013; Kruse et al., 2013; Kretzschmar et al., 2016).

However, despite significant advances that enable incorporating social conventions into navigation models, there is still no guarantee that a socially-aware robot will always make fair decisions. We can extensively observe in other applications of machine learning and Artificial Intelligence (AI), how learning algorithms replicate, promote, amplify injustice, unequal roles in society, and many other societal as well as historical biases. Numerous cases have been identified in face recognition, gender classification, and natural language processing methods (Garcia, 2016; Buolamwini and Gebru, 2018; Benthall and Haynes, 2019; Costa-jussà, 2019; Wilson et al., 2019; Lu et al., 2020; Wang et al., 2020). Similar to these cases, learning social behavior from real-world observations will not prevent discrimination. This is of special concern in service and caregiving applications where robots physically interact with humans.

There are multiple social and technical factors that can lead to bias while learning social robot navigation models. First, learning techniques require guidance to optimize the navigation model. Supervised approaches utilize datasets gathered from simulations, controlled experiments, or the real-world. Other approaches, such as imitation learning and reinforcement learning, obtain guidance directly from real experiences. It is important to consider that real-world data can always include bias reflecting unwanted humans behaviors. Additionally, simulations and controlled experiments cannot contain sufficient diverse information about diverse groups of people and their interactions for the robot to learn the large number of potential unfair situations that it can encounter. Therefore, current learning algorithms can significantly replicate, promote, and amplify unfair situations. Besides data-related issues, learning algorithms tend to find certain features that make it easier to optimize for a task and rely on these attributes to learn the function or policy. This can lead to mechanisms that depend on these potential bias inducing features related to a particular characteristic, such as race, age, or gender. Another issue encompasses fairness measurements. Thus far, there are no standard fairness definitions or metrics for the optimization of learning-based navigation algorithms or even to detect biased or unfair situations. Furthermore, robots are typically deployed with models that have been pre-trained and do not have the ability to automatically update their parameters or their policy online if they encounter a discrimination scenario.

Recently, several strategies to mitigate unfair outcomes in learning algorithms for tasks, such as classification or recognition have been proposed (Woodworth et al., 2017; Zafar et al., 2017; Agarwal et al., 2018; Dixon et al., 2018). Nevertheless, learning fair social navigation models for robotics is substantially lesser studied. Particularly, investigating fairness in mobile robot navigation presents more complex challenges that are not manifested in other data-driven tasks in computer vision and machine learning. In learning-based mobile robot navigation, fairness behavior not only depends on data but also on the future actions of the humans around the robot and other factors of the environment. In this case, it is impractical to anticipate all the possible actions in advance during the development of these models. With these considerations in mind, socially-aware robot navigation, besides learning social skills, should also account for non-discriminatory and fair behavior that makes the interaction safer for diverse groups of people.

In the case of humans, the learning process is not fixed but rather continuous. This allows humans to have both physical

and social adaptability. We refer to this adaptive learning from experiences as relearning in this work. We, as humans, not only relearn about the physical world to react to unexpected obstacles in our path, but we also develop adaptability in terms of interaction. This generally prevents us from causing harm to others with our actions and enables us to correct our behavior when we encounter unfair situations. Within this social adaptation, we learn to behave socially and fairly with those with whom we relate to (Goodwin, 2000; Hutchins, 2006; McDonald et al., 2008). The relearning process allows us to reason about what we are experiencing and develop a personality defined by certain moral values, ethical values, beliefs, and ideologies, which in turn influences the way we interact with others (Jarvis, 2006). Humans decide how to navigate in public spaces while taking both social conventions and ethical aspects into account, such as empathy, solidarity, recognition, respect for people, and recognizing behaviors that lead to discrimination. Accordingly, learning and relearning are important processes for humans to acquire the capabilities that are required for navigating in the environment and cohabitate in society.

Inspired by the learning and relearning processes in humans, we propose a framework for diminishing bias in social robot navigation. Our framework consists of two components. During robot development, we introduce social context based on social norms and skills while learning navigation models so that the robot acquires social conventions. We then incorporate a relearning mechanism that detects systematic bias in control decisions made by the robot during navigation. This enables the robot to update its navigation model when unfair situations are detected during the operation. Our proposed framework facilitates diminishing bias in the behavior of the robot and generates early warnings of discrimination after the deployment. More importantly, it enables the adaptation of the robot's navigation model to new cultural and social conditions that are not considered during training.

In this work, we describe the motivation and the technical approach for implementing our proposed Learning-Relearning framework for social robot navigation. We then highlight the risks and propose potential solutions that include specific fairness considerations for mobile robots that navigate in social environments. Furthermore, we analyze the ethical and societal implications of deploying mobile robots in social environments. To this end, we investigate the behavior of mobile robots in terms of fairness in three specific service and caregiving scenarios with different levels of human-robot interaction. There are other social scenarios where the mobility of the robot directly depends on the human's control action, such as autonomous wheelchairs (Johnson and Kuipers, 2018) or robotic guide canes (Ulrich and Borenstein, 2001). Nevertheless, in this work, we only consider scenarios where the robot navigates as an independent machine that interacts with multiple humans in the surrounding environment at different levels of priority. We provide examples that show cases where models that are only based on learning social navigation are insufficient to obtain fair behavior, and we discuss how the relearning mechanism can extend those models to yield fair behavior. Finally, we analyze scenarios in which learning social behavior and accounting for fair behavior play an important role in the real-world.

To the best of our knowledge, this is the first work to investigate the societal implications of bias in learned socially-aware robot navigation models, and the framework that we present is the first to demonstrate a feasible solution for learning fair socially compliant robot navigation models. Even though our work targets socially-aware robot navigation, the framework that we propose can also be extended to other aspects of human-robot interaction, which would benefit from the presented insights. As a result of the social perspective, we provide a comprehensive understanding of fairness in human-robot interactions. This is an important step toward diminishing bias and amplifying healthy social conventions to positively influence the society. With this work, we aim to create awareness that robots should positively impact society and should never cause harm, especially against individuals or groups who have been historically marginalized and who disproportionately suffer the unwanted consequences of algorithmic bias.

In summary, the primary contributions of this paper are:

- We introduce a framework for diminishing bias in social robot navigation, consisting of two stages: Learning and Relearning. We present the technical concept and introduce methods that can be used to implement our framework.
- We present a societal and technical analysis of the social abilities and bias considerations in learning robot navigation models.
- We present the social implications of socially-aware robot navigation models and provide a set of fairness considerations.
- We provide detailed case studies that analyze the impact of bias in different service and caregiving robot applications and discuss mitigation strategies.

2. ETHICAL ASPECTS AND FAIRNESS IMPLICATIONS

The growing impact that AI and robotics have in the daily lives of people has led to the increase in ethical discussions about current machine learning algorithms and how to handle new research toward an equal and positive impact of technology for diverse groups of people. Consequently, recent works in both social sciences and machine learning have highlighted the challenges in socio-cultural structures that are reflected and amplified by learning algorithms. As a result, many guidelines from the technical (Cath, 2018; Silberg and Manyika, 2019; Hagendorff, 2020a; Piano, 2020) and social perspectives (Verbeek, 2008; Liu and Zawieska, 2017; Birhane and Cummins, 2019) have been presented. These guidelines (Vayena et al., 2018; Hagendorff, 2020b; Piano, 2020) are aimed toward mitigating the adverse effects and advocating for ethical principles, such as fairness, trust, privacy, liability, data management, transparency, equality, justice, truth, and welfare. Similar efforts have been made by the European Robotics Research Network (Euronet) in the Euronet Roboethics Atelier project in 2005, and the British Standards Institute which published the World's First Standard on Ethical Guidelines in 2016 (Torresen, 2018). Moreover,

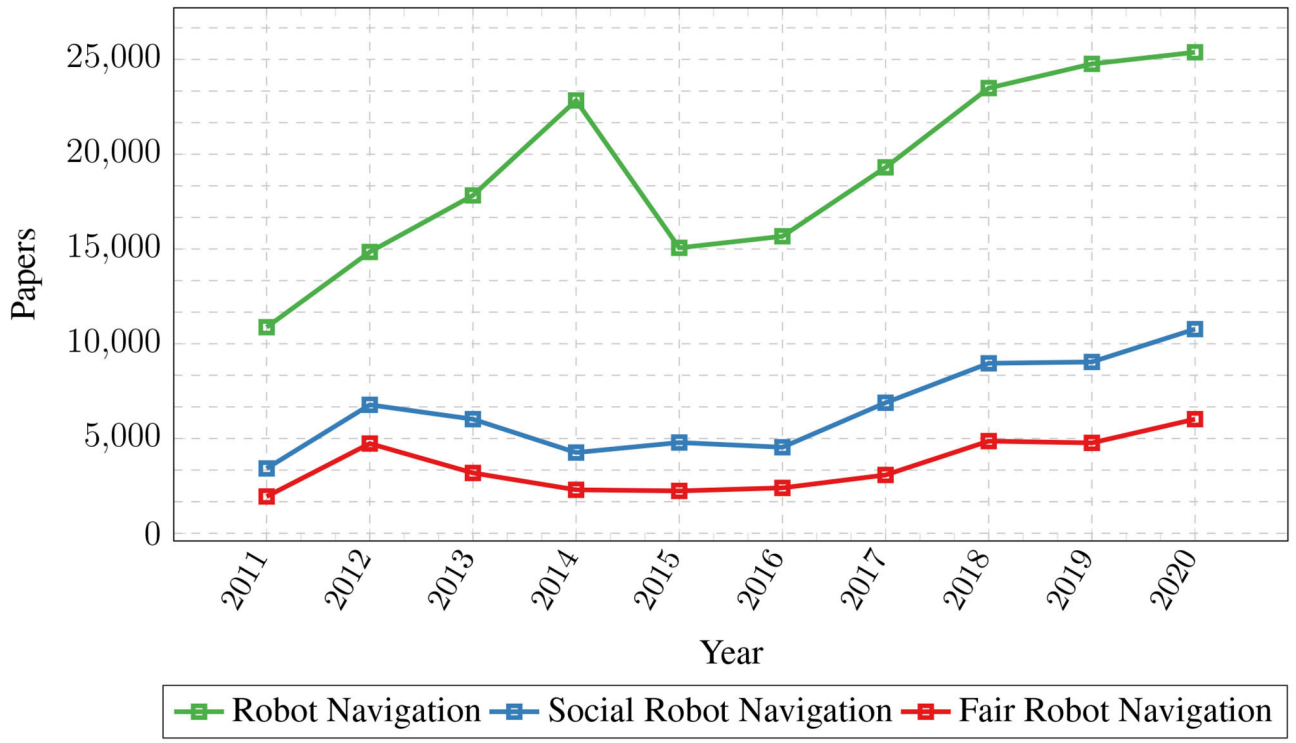

FIGURE 1 | Comparison of the number of publications on Robot Navigation (blue), Social Robot Navigation (red), and Fair Robot Navigation (green) from 2011 to 2020. Although the rate at which fairness is being considered in robot navigation methods is increasing, there is a growing gap with the number of works that address robot navigation each year.

some works in robotics (Anderson and Anderson, 2010; Lin et al., 2012; BSI-2016, 2016; Boden et al., 2017) have also investigated the importance of addressing ethical issues for safe and responsible development.

These ethical guidelines (Reed et al., 2016; Goodman and Flaxman, 2017; Johnson et al., 2019; Arrieta et al., 2020) share the value of robots effectively and safely assisting people, and under no circumstance cause harm or endanger their physical integrity (De Santis et al., 2008; Riek and Howard, 2014; Vandemeulebroucke et al., 2020). The impact of human-robot interactions has also been studied to a lesser extent in mobile robotics, e.g., providing recommendations on road safety, privacy, fairness, explainability, and responsibility (Bonnefon et al., 2020), or studying fairness in path planning algorithms of robots during emergency situations (Brandão et al., 2020). Similarly, such ethical discussions should be contrived while developing socially-aware robot navigation models. As shown in **Figure 1**, although the number of publications that consider fairness in robot navigation is slowly increasing, it is still over five-times lesser than the overall number of publications that address robot navigation. In this section, we present a series of ethical aspects and social implications that can arise from bias in socially aware-robot navigation algorithms. Additionally, we analyze the impact that these social navigation algorithms can have in human environments.

2.1. Fairness Implications

The cultural and social knowledge in humans is transferred from generations as a cumulative inheritance that allows each member of the society to incorporate moral, political, economic, and social structures that not only have a positive but also a negative value (Castro and Toro, 2004). These inheritance conditions have perpetuated historical discrimination against individuals and groups of people. The data collected in machine learning and AI come from these historical inheritance structures; consequently, social-historical discrimination can also be reflected or even amplified by learning algorithms. In recent years, several unexpected outcomes have been observed in learning algorithms that have caused discrimination and prejudice in society. Numerous examples demonstrate how social prejudices are reflected in machine learning algorithms (Garcia, 2016; Wang et al., 2020). One clear example that was observed in natural language processing was the racial and gender biases while learning language from text (Costa-jussà, 2019; Lu et al., 2020). Another recent example is the automated risk assessments used by U.S. judges to determine bail and sentencing limits. It was shown that it can generate incorrect conclusions, resulting in large cumulative effects on certain groups, such as longer prison sentences or higher bails imposed on darker-skinned users (Benthall and Haynes, 2019). Moreover, another study shows how biased algorithms affect the performance of vision-based object detectors employed in autonomous vehicles. Their

work demonstrates that pedestrians with dark-skinned tones presented higher recognition errors (Wilson et al., 2019). There have also been numerous cases of algorithmic bias that have been observed in algorithms used in healthcare. For example, algorithms trained with gender-imbalanced data have shown higher error at reading chest x-rays for an underrepresented gender (Kaushal et al., 2020).

The numerous cases of discrimination observed in learning algorithms employed in various applications are a source of concern for robotics. In the case of robots that employ learning algorithms to effectively interact, navigate and assist people, it is essential to foresee possible unfair situations. Specifically, as a result of learning socially-aware robot navigation strategies, these trained models can enhance the social impact in terms of human acceptance of mobile robots, daily use, comfort, security, protection, and cooperation (Thrun et al., 2000). Providing robots with a more natural navigation ability also increases their usability. Although incorporating social navigation models in robots improves their usability, comfort, and safety in human spaces, social abilities by themselves do not ensure fair robot decisions, especially while using learning algorithms to imitate or follow human conventions and behaviors. In human social interactions, a series of direct and indirect discrimination behaviors and decisions are often present (Forshaw and Pilgerstorfer, 2008; Zhang et al., 2016; Yu, 2019). Using learning algorithms can negatively affect society, individuals, or groups if unwanted social behavior is replicated and reflected in the actions of the robot. Therefore, this highlights the need to implement fairness considerations and measures. The ability of an agent to dynamically make fair decisions among different people is a fundamental basis for trust in human-robot interaction (Ötting et al., 2017; Claure et al., 2019). If robots after their deployment present an unfair behavior, it will continue to perpetuate discriminatory structures that will be reflected in the way that people are assisted. Moreover, this will cause serious consequences, such as a large population not being benefited by the robots and being reticent to use them. These factors suggest that the robot would only be beneficial for certain groups of people, which would continue to reinforce large social inequalities. Robots should influence society in a positive way by promoting healthier relationships, roles, and dynamics after their deployment in different places with diverse people. This requires the creation of a more reflective, equitable, and inclusive learning methods accompanied by extensive studies from the social perspective.

2.2. Fairness Measures

Fairness is a complex ethical principle that relates to avoiding any form of systematic discrimination against certain individuals or groups of individuals based on the use of particular attributes, such as race, sexual orientation, gender, disability, socioeconomic, and sociodemographic position (Silberg and Manyika, 2019). However, the definition of fairness tends to be dynamic, mobile, and contingent, therefore it should be analyzed from a reflective and ethical perspective. Moreover, fairness highly depends on the context, location, and culture, among other factors. Consequently, defining an accurate fairness measure could be a complex task. With efforts in this direction, bias has been used to represent fairness either in human environments or in technological developments (Howard et al., 2017; Fuchs, 2018; Lee, 2018; Nelson, 2019).

For its part, solutions to algorithmic bias that perpetuate social and historical discrimination against vulnerable and disadvantaged individuals or groups of people tend to be technical rather than moral and ethical (Birhane and Cummins, 2019). Technological solutions to biased decisions making are essential but not solely sufficient. Instead, technical solutions should be accompanied by factors, such as diversity, inclusion, and participation of underrepresented groups during the development of navigation models. Although there is no standard definition of fairness in machine learning and AI, some works state that a prediction is fair when it is not discriminating or when there is no bias (Binns, 2018; Chouldechova and Roth, 2018; Birhane and Cummins, 2019). However, there are two types of biases, positive and negative. Positive bias frequently promotes social good and avoids prejudice through awareness and respect for human differences. Therefore, not all biased outputs are necessarily undesirable and eliminating them can cause unintended outcomes for certain people. For example, consider an algorithm that is used in a bank to perform a credit study of the people who apply for a loan. If the algorithm is trained to guarantee that all the people will have credit, this may be a disadvantage in the long run for those who cannot pay back later. While the algorithm is being equal in this case, it is being unfair in the long term as it negatively affects the low-income people (Silberg and Manyika, 2019).

In socially-aware robot navigation fairness measurements are yet to be studied. As robots interact and assist different groups of people in different settings, creating a unified definition or a metric is impractical due to the complex and diverse cases that robots can encounter after deployment. Accordingly, in order to tackle unfairness, we present a series of fairness considerations for socially-aware robot navigation:

(i) **Value Alignment** refers to the alignment of human values in decision making during navigation. These values include respect, inclusion, empathy, solidarity, recognition, and non-discrimination. In socially-aware robot navigation, it is reflected in cases when the decision-making of the robot reproduces and increases the welfare of vulnerable populations. For example, prioritizing to assist and serve people with physical disabilities in crowded environments.

(ii) **Bias Evaluation** is related to the evaluation of bias in decisions making during navigation. Bias can be considered acceptable if there is adequate reasoning or unacceptable if the bias replicates, promotes, or amplifies discrimination. For example, when robots navigate with a different speed around young people who are faster than around older adults, it is usually accepted because they have important physical differences. Nevertheless, if such decisions are made based on racial differences, it can be considered unacceptable, given that there are no fair reasons for this difference. With this fairness consideration, when biases are presented in navigation models, it can only be accepted if there are fair reasons for doing so.

(iii) **Deterrence** is expressed in preventing and mitigating unwanted bias as well as discrimination during navigation. Since the notion of deterrence is dynamic and can vary depending on the social context, robots should be sensitive to cultures by adapting to people, customs, and their surroundings.

(iv) **Non-maleficence** signifies that the decisions of a robot can never produce damage to people. The damage is primarily interpreted as bodily harm, collisions, interruptions, delay, and obtrusion. However, damage can also refer to the negative effects caused by discrimination, segregation and bias. For example, if a caregiving robot in a hospital becomes an obstacle to the medical personnel responding to an emergency due to biased decisions, then it would be violating this property.

(v) **Shared Benefit** refers to providing equal benefits to diverse people in all scenarios. If a robot is specifically designed for and only tested in a particular geographical area, tailored to the characteristics and behaviors of the people in that region, it can lead to unwanted bias when it is deployed in a new region which may have completely different characteristics. Therefore, the benefits that the robot provides should not be targeted toward people with specific characteristics in a determined geographical area, but should rather be equally beneficial to all users. In this case, adaptability is an important attribute for robots to achieve shared benefit so that the autonomy of the robot is flexible to adapt to characteristics of specific users in the social environment where it is deployed.

2.3. Responsible Innovation

Research in technology studies suggests that the conceptions of responsibility should build upon the understanding that science and technology are not only technically but also socially and politically constituted (Winner, 1978; Grunwald, 2011). Responsible Innovation (RI) was introduced as a concept to address the impact of research and innovation in technology from an ethical and fair perspective. RI states that the technology should be anticipatory, so it should have a foresight guide that provides alternative options for responsible development (Stilgoe et al., 2013; Brandão et al., 2020), and it should account for social, ethical, and environmental issues. Based on RI principles, the framework that we present in this paper aims to identify biased behavior during navigation and promotes fair decision making through the learning and re-learning process to enable flexible and adaptive service. RI articulates and integrates four factors: (i) anticipation of damages, (ii) reflection from an ethical perspective, (iii) protection of sensitive human characteristics, such as age, gender, and race, and (iv) responsiveness (Stilgoe et al., 2013).

With the aforementioned RI factors, responsible robotics aims to ensure that responsible practices are carefully accounted for within each stage of design, development, and deployment. Correspondingly, robot navigation models should address the ethical and legal considerations at the time of development. Given that these considerations are constantly changing depending on the social or cultural factors, these models should be updated accordingly.

3. LEARNING—RELEARNING FRAMEWORK FOR SOCIALLY-AWARE ROBOT NAVIGATION

The goal of our proposed framework is to develop learning models for robot navigation that yield-social and fair behavior. To this end, we define two different stages: learning and relearning. In the first stage, we incorporate social context into learning navigation strategies so that robots can navigate in a socially compliant manner. While, in the second stage, we aim to diminish any bias in the planned paths with the learned navigation model. In this section, we first introduce socially-aware robot navigation. We then describe our proposed framework and present the technical approach that can be used for the implementation. **Figure 2** shows the different stages of our framework. In the learning phase, we learn a navigation policy based on imitation learning with additional social constraints. Whereas, in the relearning phase, we analyze the outputs of the network online and provide the model with updates to reach the navigation target while accounting for and deterring bias to ensure fairness. Science and technology, from the RI perspective, have the ability to provide significant benefit through well-established methodologies that reflect responsibility and ethical principles. This framework tailored exploits the learning and re-learning process as a methodology to achieve responsible robot navigation.

3.1. Socially-Aware Robot Navigation

One of the widely studied requirements for mobile robots to operate in human spaces is the ability to navigate according to social norms and socially compliant behavior. The social navigation models that are employed in robots play an important role in the effect that these automated machines have on society and the perception as well as confidence that humans will have of them. In the case of humans, we develop the ability to navigate while considering numerous variables representing the environment, such as the objects, people, and dynamics of the agents in it. This ability, known as sociability, from an anthropological point of view, is the human capacity to cooperate and engage in joint behavior with others (Simmel, 1949). Further, sociability allows us to navigate while avoiding situations that make us uncomfortable or put us or others in danger.

Different social norms have been developed to provide information about the appropriate behavior, especially in public spaces. Social norms are standards of conduct based on widely shared beliefs of how people should behave in a given situation (Fehr and Fischbacher, 2004). Some of the social norms for navigation are not invading the personal space of people, passing on the right, maintaining a safe velocity, not blocking peoples path, approaching people from the front, among others (Kirby, 2010). Besides social norms, different studies, such as proxemics (Hall et al., 1968), kinesics (Birdwhistell, 2010), and gaze (Argyle et al., 1994) also provide cues to determine the

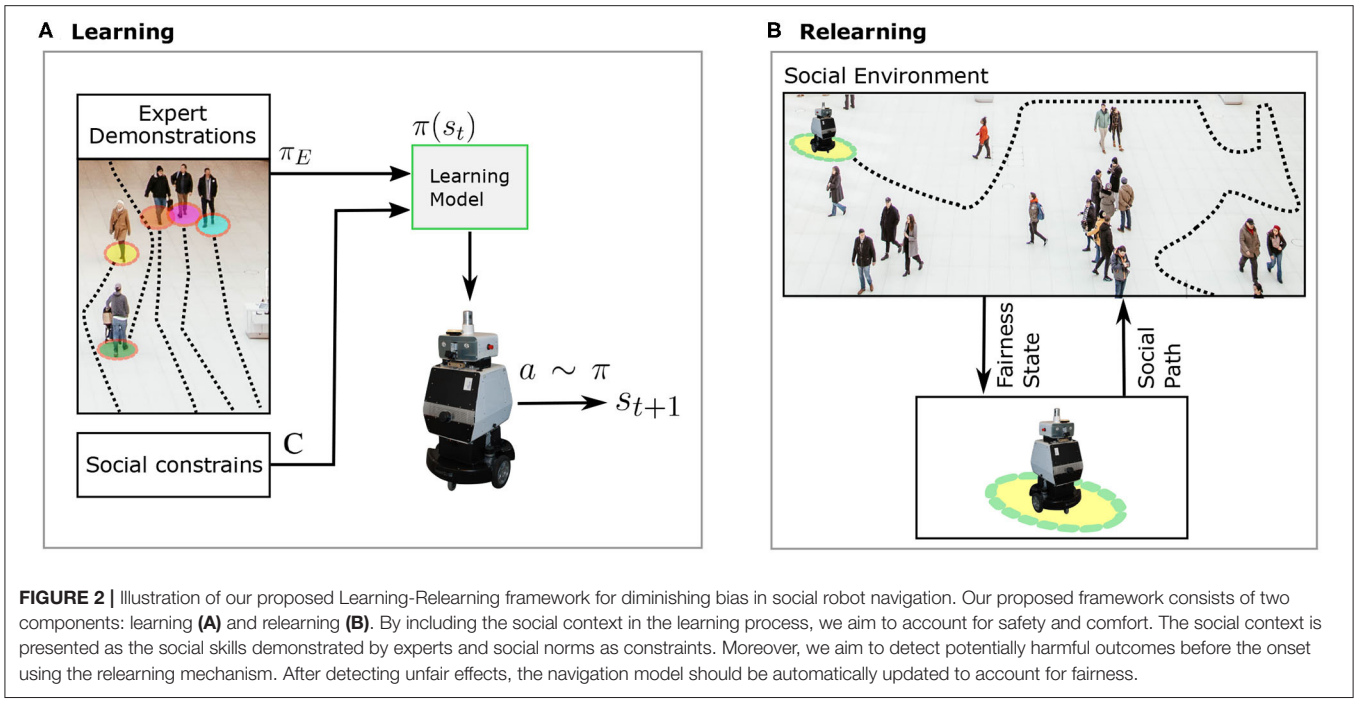

FIGURE 2 | Illustration of our proposed Learning-Relearning framework for diminishing bias in social robot navigation. Our proposed framework consists of two components: learning **(A)** and relearning **(B)**. By including the social context in the learning process, we aim to account for safety and comfort. The social context is presented as the social skills demonstrated by experts and social norms as constraints. Moreover, we aim to detect potentially harmful outcomes before the onset using the relearning mechanism. After detecting unfair effects, the navigation model should be automatically updated to account for fairness.

appropriate manner to approach a person, navigate around, and coordinate in public spaces. Specifically, proxemics is the study of the perception and organization of the personal and interpersonal space. It is associated with the manner of how humans manage their surrounding space when they walk in public environments and how their comfort can be affected by the movement of other pedestrians (Rios-Martinez et al., 2015). Kinesics is related to the actions of the body and positions (Birdwhistell, 1952); and gaze refers to the eye movements and directions during visual interaction (Harrigan, 2005). These studies highlight social skills, such as reading emotions and the prediction of intentions of people. The combination of both social norms and social skills can be considered determinant to sociability. The aforementioned studies and norms are some of the increasingly used factors in learning social robot navigation models. It is long believed that equipping robots with these social skills and social norms will enable them to react socially as humans do.

For instance, we can anticipate that cleaning robots (Fiorini and Prassler, 2000) that are primarily used in houses will be widely used in public spaces in the coming years. Currently, these robots do not conform to any social norms during navigation. Confined to private locations and users who know the device, manufacturers have not made it a priority to include social skills, such as predicting the intention of people and avoiding crashing into them. Nevertheless, sociability is an important skill to deploy cleaning robots in crowded public spaces. In this case, robots must take into account aspects, such as the space that they occupy and the personal space of the people around to determine how close to navigate around them or predict where humans will move so that they do not interfere with their paths. These skills will allow robots to plan a safe route so that their presence is not disturbing, surprising, or scaring the people that share the same space. While planning routes, robots should use social norms, such as not invading the personal space and maintaining a safe speed. Both the use of social skills and social norms change depending on the type of robot and the context in which it is used. We present further discussions of this example in section 4.1.

Socially-aware robot navigation methods can primarily be categorized into two groups. The first category is model-based and consists of handcrafted models that use mathematical formulations to combine a set of effects to determine dynamics of pedestrians, such as reaching the destination, the influence of other pedestrians, keeping a certain distance to another person or the maximal acceptable speed. Helbing and Molnar (1995) introduced the notion that social forces determine human motion and proposed the Social Force Model (SFM) to represent pedestrian dynamics. To navigate in a manner similar to humans, this formulation was later used to provide robots with pedestrian-like behavior for human-robot social interaction (Ferrer et al., 2017). However, SFM requires us to cautiously define and tune the parameters for each specific scenario, which makes it impractical to scale to complex tasks and environments (Tai et al., 2018). The second category consists of learning-based methodologies that use some form of guidance or demonstrations containing the policies that link observations to the corresponding actions. We further discuss learning-based methods in the following section.

3.2. Learning

The rapid progress in machine learning in the past years and the growth of computing power have enhanced the learning capabilities of autonomous mobile robots. Currently, these learning-based methodologies play an essential role in the

development of complex navigation models. These models are primarily trained to achieve the best navigation performance under some given metrics during the learning process. For this purpose, different guidance techniques have gained interest in robot navigation works. The first of which is supervision from labeled data, which uses either data gathered from the real-world or simulations and the corresponding annotations. The data and annotations are then employed to optimize the model so that the output predictions are as close as possible to the labels. Supervised navigation methods can be used directly by learning the mapping from the states in recorded trajectories that contain social policies to their corresponding labels or by learning reactive policies that imitate a planning algorithm (Groshev et al., 2017).

Another extensively explored learning technique is Reinforcement Learning (RL), in which an agent explores the state and actions by itself while a reward function is used to punish or encourage the decisions to obtain an optimal model. RL techniques can be used to provide a robot with the navigation paths that maximize rewards in terms of human safety or comfort (Chen et al., 2017). Moreover, Inverse Reinforcement Learning (IRL) is a technique that has been widely used to capture the navigation behavior of pedestrians. Contrary to supervised learning, IRL is able to recover a cost function that explains an observed behavior (Kuderer et al., 2013). The IRL technique proposed by Hamandi et al. (2019) trains the social navigation model by learning the navigation policy directly from human navigated paths in order to generate actions that conform to human-like trajectories. To include the social context in the learning process, these models aim to clone the navigation behavior of humans. Subsequently, robots are then equipped with these models for socially-compliant navigation.

Specifically, to clone an expert behavior in the RL framework, consider that an agent in an environment reaches a state s_{t+1} after executing an action $a_t \sim \pi$ that follows a policy π. At each transition state, the agent obtains a reward r_t presented as a scalar. The goal is for the agent to adjust the policy π to maximize the expected long-term rewards that it can receive. Q-learning (Watkins and Dayan, 1992) is an approach that enables us to find an optimal policy based on the state transition set. The Q-function represents the value of an action a_t and following a policy π as

$$Q_\pi(s_t, a_t) = \mathbf{E}[R(s_t)|s_t, a_t], \tag{1}$$

where R is the expected long term reward defined as $R = \sum_{t=0}^{\infty} \gamma^t r_t$, being $\gamma \in [0, 1]$ the discount-rate. Given the state s_t and action a_t the Q-function indicates the expected discounted accumulative reward. Using the Q-function, we can estimate an optimal policy π which maximizes the expected return. Particularly, no reward function is given in the IRL framework. Therefore, it is inferred from observed trajectories collected by the expert policy π_E to mimic the observed behavior.

There are numerous works using RL and IRL that generate human-like navigation behavior in controlled conditions. However, we can more elaborately define how we as humans navigate the environment, using a combination of both social skills and social norms as described in section 3.1. Social norms can vary with respect to the context, location, and culture. Extending the social skills of the robot by including social norms is important for social domain adaptation. The social norms that a domestic robot should consider while navigating are substantially different from those that a mobile robot in a hospital should conform to. For example, in order for the robot to navigate in a socially compliant manner in a hospital, it is essential for it to identify emergency situations, understand the priority for interaction, and have fast reaction times, so that the robot can never interfere with the paths of hospital staff and cause accidents or delay the treatment of a patients. Given that the context and priorities differ, the reaction also accordingly changes. We explore these cases in the case study that we describe in section 4.

Recently, a deep inverse Q-learning with constraints technique (Kalweit et al., 2020a) was introduced. This work presents one such model that allows for the combination of imitating human behavior and additional constraints. This is a novel model-free IRL approach that extends learning by imitation with constraints, such as safety or keeping to the right. Using the previous definition of Constrained Q-learning (Kalweit et al., 2020b), it includes a group of constraints C that shapes the possible actions in each state. Besides the Q-function in Inverse Q-learning, it also estimates a constrained Q-function Q_C for which the policy is extracted after Q-learning, considering only the action-values of the actions that satisfy the required constraint. This approach shows promising potential for considering relevant social factors while learning socially-aware robot navigation policies, especially by adding diverse constraints that represent current norms in order to yield socially intelligent and unbiased robot behavior.

3.3. Fairness Considerations

As with most learning approaches, the method described in section 3.2 requires a large number of training examples so that the model learns to yield the desired output. Therefore, it is essential to use either data gathered from the real-world, simulations, or control experiments. With the collected data, developers aim to present representative examples of real-world scenarios or guidance of the desired social behavior during navigation. However, these data collection processes can themselves reproduce biases, and as a consequence, it raises a series of critical concerns. In the specific case of learning socially-aware robot navigation from real-world data, robots can reproduce biased behaviors implicit in human-human interaction. On the other hand, the amount of training data that can be obtained from simulations and control experiments is very limited since only a handful of situations are taken into account. Most data collection processes that do not encompass a balanced set of every possible real-world scenario present a risk for robots trained on them as this could lead to navigation with biased behavior. These circumstances are considered as bias in the data. Accurate generalization of scenarios that highly deviate from the training data is an extremely difficult task. To address this factor, recent methods have been proposed to filter data that is used to train the models. For instance, Hagendorff (2020a) presents a

selection process for training data that improves the data quality in terms of ethical assessments of behavior and influences the training of the model. Nevertheless, methods to reduce bias in the data that is used for learning robot navigation models still remain unstudied.

Apart from the problems in dataset collection, there is still a lack of a deeper understanding of the underlying principles and limitations of modern learning algorithms. Especially, a phenomenon known as shortcut learning which shows how neural networks learn more straightforward predictors that are not necessarily related to the main task or objective (Geirhos et al., 2020). A typical example of this phenomenon can be seen in the hiring tool developed by Amazon which predicts strong candidates based on their curriculum. This tool was later found to be biased toward providing advantages for male applicants. Their model, which was trained on historical human decisions that were made during the hiring process identified that gender was an important feature for prediction (Dastin, 2018). Geirhos et al. (2020) analyses the dependency of outputs to strong predictive attributes found by the model during training.

Data-driven models can contain abstract representations of the data and situations that lead to the prediction. Therefore, it is typically challenging to explain the decisions made by a learned model. To facilitate the fairness analysis, we present an approach that is not solely data-driven and instead, it implicitly incorporates human interpretations of social dynamics using a model that includes high-level and explainable human notions about social conventions, relationships, and interactions to guide a mobile robot. The purpose of analyzing this approach is to demonstrate that biased behaviors can also be learned from biased demonstrations or observations. We analyze the approach proposed by Patompak et al. (2019) to predict personalized proxemics areas that correspond to the characteristics of individual people. This approach generates personalized comfort zones of a specific size and shape by associating the personal area with the activity that a person performs or characteristics of the person. Using these social descriptions, it estimates the proxemic zone that better matches each pedestrian in the scene. Consequently, the approach relies on personalized boundary delineation of two different areas: one area where the human-robot interaction can occur, and another area that is private, which the robot should avoid navigating through. The approach consists of three parts: human-social mode, learning the fuzzy social model, and a path planner. The human social model utilizes proxemics theory and aims to reflect the pedestrians' social factors in the scene. The social factors that are considered include gender, relative distance, and relationship degree. Using these factors, the approach yields the parameters that determine the private zone of comfort for each person in the scene based on the fuzzy logic system. For each social factor that is considered, the approach defines a membership function as follows:

A binary function depending on the gender of the pedestrian, which is given by

$$MF_{gender} = \begin{cases} 0, & \text{if gender is } \textit{Male} \\ 1, & \text{if gender is } \textit{Female}, \end{cases} \tag{2}$$

a sigmoid function with relative distance input r_r, distribution steepness a_r, and inflection point c_r describing *near* or *far* distance defined as

$$MF_{distance} = \frac{1}{1 + \exp(-a_r \times (r_r - c_r))}, \tag{3}$$

and three Gaussian functions representing the degree of relationship as familiar, acquaintance, and stranger, which is given by

$$MF_{relationship} = \begin{cases} \mathcal{N}(\mu_{\text{Fam}}, s^2{}_{\text{Fam}}), & \text{if degree of relationship is } \textit{Familiar} \\ \mathcal{N}(\mu_{\text{Acq}}, s^2{}_{\text{Acq}}), & \text{if degree of relationship is } \textit{Acquaintance} \\ \mathcal{N}(\mu_{\text{Str}}, s^2{}_{\text{Str}}), & \text{if degree of relationship is } \textit{Stranger}. \end{cases} \tag{4}$$

Subsequently, the fuzzy social model is learned from human feedback using an RL approach. The defined membership functions of the social factors can be learned to yield an improved personal area for each pedestrian. This is performed by adjusting the relationship degree in the MF (Equation 4) to update the social map. The reward of the RL model is then obtained from human-robot interaction by means of the emotion or feeling of each corresponding person. Therefore, the approach sets the focus on the degree of the relationship to be learned. Finally, the approach selects a path planner that chooses an optimal navigation path in the social cost map. The consequently designed social interaction area using fuzzy rules presents the output of the model as two separated personal areas: far personal area (FPA) and near personal area (NPA). As part of the rules presented, it is clear that for the input gender female, the near personal area is never an option. Taking into account that the reinforcement learning algorithm updates the model based on the $MF_{relationship}$, the resulting navigation policy would never allow for human-robot interaction close to women. This presents a critical bias of the model due to the inclusion of social dynamics. This is an example where bias appears due to an explicit constrain in the learning algorithm. Not only gender but other factors that may potentially lead to bias as well as other implicit or explicit biases can appear by learning from real-world data. We discuss this technical bias of the aforementioned navigation model with implications and analysis from the social perspective in section 4.

Learning robot navigation policies and models that are unbiased requires analyzing how the input is given, how the data is measured, how the data is labeled, what it means for models to be trained on them, what parameters are used, and how social navigation models are evaluated. If models aim to reflect the features of society, we need to question what behaviors should be replicated and promoted. For example, Kivrak et al. (2020) explicitly exclude women in the real-world experiments of their social navigation framework for assistive robots around humans. Their model that aims to yield human-friendly routes

was only tested in a corridor where women were excluded based on previous analysis (Jones and Healy, 2006), which affirms gender differences in spatial problem solving. This represents bias in the evaluation where the social model of navigation is validated only for a privileged group and can lead to underperformance to the unconsidered after the deployment. This has also been seen before in medical datasets or experiments where women were excluded citing differences in hormonal cycles, which leads to the medicines or medical procedures causing higher side effects for women compared to men. The consequences of these biased experiments or trials have been extensively discussed, which had lead to the inclusion of women in all medical trials (Söderström, 2001).

The technical bias analysis presented in this section shows cases where the high-level representation of social interaction replicates unequal roles and dynamics that already exist in human interaction. It is a significantly larger risk in the case of learning models for social navigation from demonstrations where the assumption is that the best way to teach a robot to navigate is to enable it to learn directly by observing humans.

3.4. Relearning

While learning socially-aware robot navigation models, social biases can be introduced that replicate and even augment the unfair societal dynamics. Most existing socially-aware robot navigation techniques aim to learn social navigation behavior by imitating human navigation. Consequently, it essential to deter biases during the deployment of robots equipped with such models. In this section, we present a mechanism to first detect when the navigation model makes biased decisions, especially against certain groups of people. Subsequently, we use this mechanism to update the model toward yielding more equitable social navigation policies.

There are many situations in the real-world where unequal decisions are desired, such as adapting the speed of the robot near older adults. In this work, we only analyze situations where there is no justifiable reason to yield different actions while interacting with different groups of people. In this case, an unfair or discriminatory system will offer an advantage to a certain group of users or unfavorable interaction to some other groups. Unfair behavior in robot navigation directly affects how users interact with the system. For a mobile robot to amend a discrimination behavior, it is necessary first to detect or measure the biased behavior. An advantage in the case of robots is that the decisions and actions after deployment can be used to measure the degree of biased decisions, for instance, concerning protected characteristics, such as age, gender, and race. Whereas, in the case of bias in deep learning models this task would be significantly harder. For instance, the Microsoft AI Twitter chatbot Tay which learned by interacting with users and presented gender-biased as well as racially offensive tweets (Perez, 2016). In this case, it would be necessary to additionally measure the features behind the posted tweets. Given that most robots are designed to move in the world, this characteristic comes for free in terms of the navigation actions that were made based on distance, speed, among other control variables as well as perception, accuracy, and uncertainty.

The robot can gather a dataset or a log by storing its own experiences and its corresponding actions even after deployment. Subsequently, the first step is to detect bias in the social navigation decisions of the robot. Bias identification is related to detecting disproportionate prejudice or favoritism toward some individuals or groups over others. For example, the paths planned by the robot produces a negative effect more frequently for specific groups of people than they do for another, such as discomfort, lack of interaction, or avoidance. Other situations are related to a disproportionate rate of a favorable or higher quality of attributes prediction for certain groups. This situation can present itself due to a lack of representation and diversity in the data or scenarios that were used in the learning stage. As a result, it can lead to unpredictable or no interaction with individuals of these groups.

One such method to detect if the navigation model exhibits outcomes that differ across subgroups is using clustering. Clustering is the technique for grouping data such that the elements of the same group are assigned closed together, forming assemblies called clusters. Clustering is a well-studied technique that is highly used in unsupervised or exploratory data analytics. Consider that the dataset collected while the robot was navigating contains all the decisions that were taken as well as the sensor data and the actions of other agents that these decisions were based on. Additionally, other navigation and perception attributes can be considered, such as the relative distance of the pedestrians to the robot, collisions, person identification confidence, and intention prediction, as well as additional information, such as rules that were violated and accidents that were caused. The accumulation of actions the robot outputs corresponds to the navigation feature set to be clustered. The resulting clusters can later be correlated to potential protected characteristics.

Having a learned policy π for socially-aware robot navigation, we define $V = \{v_1, v_2, \ldots, v_i\}$ as the set of navigation data that correspond to the experiences that the robot continuously accumulates through certain time steps. Different clustering algorithms can be used depending on the attributes of the selected navigation features (for instance if their nature is categorical or numerical). One promising clustering algorithm is the method proposed in Aljalbout et al. (2018) which consists of a fully convolutional autoencoder trained with two losses, one for reconstruction and the other for cluster hardening. The result of the clustering process is a collection of assemblies $A = A_1, A_2, \ldots, A_K$ consisting of navigation feature combinations. Each A_k represents the navigation experiences that are similar enough to be considered as a cluster of the entire set V. The number of clusters K and the size of each cluster A_k are hyperparameters that can be explored. Additionally, we define $F = \{f_1, f_2, \ldots, f_N\}$ as the set of protected features that we aim to analyze and each f_n has a set of navigation features V. To uncover social-group related bias the next step is to determine the relationship degree $D_{k,n}$ between each protected feature f_n and each generated cluster A_k.

After identifying that the robot actions in the navigation experience set are clustered and correlated to sensitive attributes, the next step is to trigger alarms or corrective actions when protected feature f_n strongly related to each generated cluster A_k,

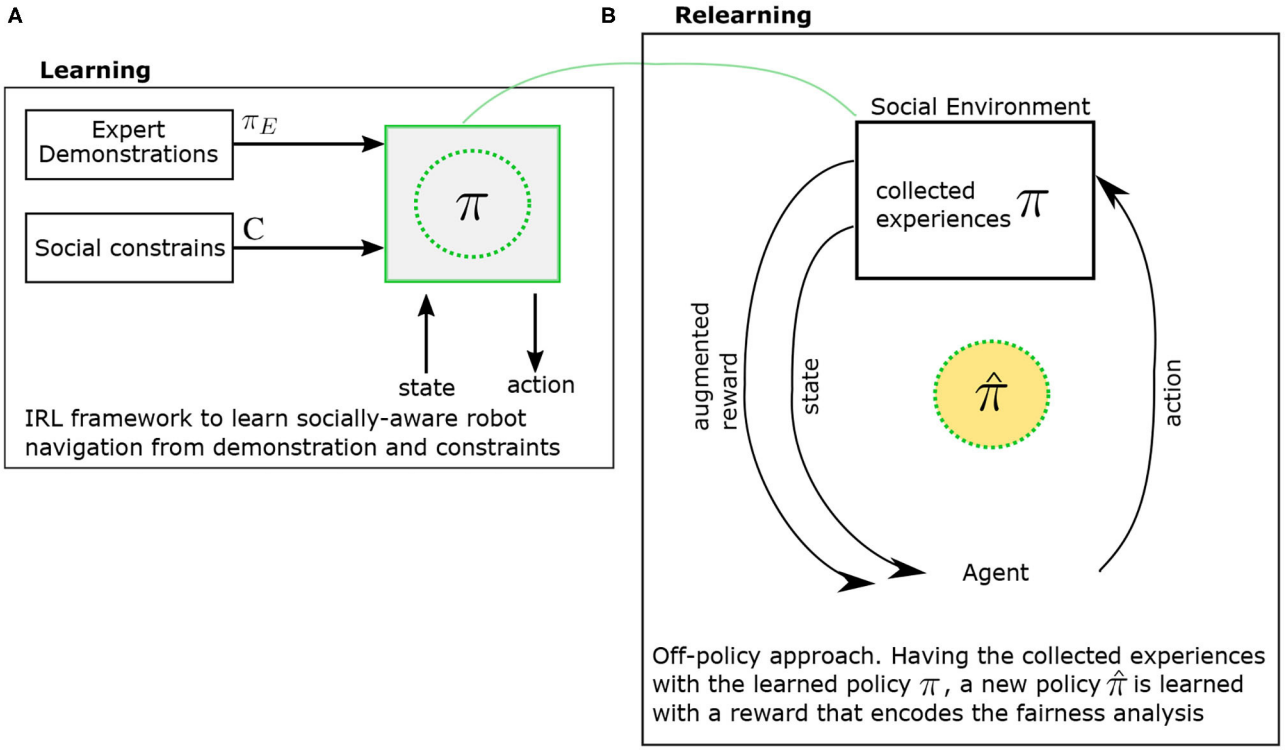

FIGURE 3 | Illustration of the Learning-Relearning framework for diminishing bias in social robot navigation. During the learning phase **(A)**, a policy π is learned for socially-aware robot navigation. During the relearning phase **(B)**, the robot uses the policy π to navigate in the social environment and collects the navigation experiences. An augmented reward that encodes detected biased behavior is used to relearn a new policy $\hat{\pi}$ so that the long term rewards reflect the decreasing unjustified bias related to social-groups.

defined as $D_{k,n} > u_n$ where u_n threshold that can be selected for each protected feature. A system of reward or punishment can be implemented in a off-policy reinforcement learning algorithm that optimizes an augmented reward that encodes the detection of unfair behavior as shown in **Figure 3**. The augmented reward rR_t is penalized when a biased behavior is detected so it does not only comprise the behavior for socially-aware navigation but it is also discounted when we detect bias as $D_{k,n} > u_n$. Therefore, the robot learns the policy π_R so that the long term rewards reflects the decreasing unjustified bias related to social-groups. As a result, it is possible to relearn the navigation model in our framework depending on the information gathered from the social environment.

From a more realistic perspective, demographic information is rarely known. Clustering also allows the reduction of this dependency between predictions and demographic information, when an unsupervised approach is employed. Therefore, when the dataset containing memory experiences of the robot navigating conforms to clusters beyond a given threshold, it can trigger an alarm for further analysis. Other methodologies that can be used to undercover bias in deep learning models are based on visualization of embeddings. Using visualization techniques, we can show how the model groups the data, which is useful to expose the reasons behind the prediction of the model. To do so, different tools can be used, such as T-distributed stochastic neighbor embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) to project the embeddings to reduce the dimensionality of the data. In this work, we focus on the relearning component based on clustering to present a feasible solution to account for fairness while learning socially compliant robot navigation that can be extended to an unsupervised algorithm.

4. CASE STUDIES AND DISCUSSION

In this section, we present extensive discussions that relate the technical analysis of our proposed framework to complex real-world scenarios that we present as three case studies. Each of these case studies contains different levels of human-robot interaction under four specific protected characteristics: gender, disabilities, age, and race. With these scenarios, we analyze the feasibility of model adaptation and the utility of this mechanism to check for fairness as well as to correct the bias. The figures illustrated in this section were generated using Icograms (2020).

4.1. Autonomous Floor Cleaning Robots

One of the most societally accepted robots has been the autonomous floor-cleaning machines (Forlizzi and DiSalvo, 2006; Forlizzi, 2007; Fink et al., 2013) and during the last decade they have been the most sold robots in the world (Research,

FIGURE 4 | Illustration of the autonomous floor cleaning robot scenario. The robot navigates taking the social conventions into account while performing the main task of cleaning the entire area.

2019). These robots have the task of cleaning floors using vacuum systems without any human supervision and recently, they can also mop floors using steam systems. These robots are currently used in households, and their navigation models vary in complexity depending on a wide range of prices. However, these robots are so far not equipped with socially aware navigation models. They do not avoid people or dynamic objects, rather they only change their cleaning route after they collide with an object. This can be attributed to the fact that in household environments, people are typically more tolerant given that they are aware of the task, features, and capacity of the robot.

It can be expected that the use of cleaning robots in the future will spread to different public areas. In this case study, we analyze from both technological and social points of view the functioning, requirements, and implications of the navigation of a cleaning robot that operates in a shopping mall. We illustrate this scenario in **Figure 4**. Consider that the shopping mall consists of multiple and extensive floors, and it is open to the public continually every day of the week. The groups of people visiting the place range from families and groups of friends to individual persons. Additionally, the reasons for the visit can differ, including people making quick shops, taking a walk, eating, etc. Therefore, we also expect varying types of behavior of the visitors, such as walking at a different speeds, talking in groups, and sitting down in different spaces.

The task of the robot in this case is to clean the entire environment effectively. In the following, we examine the effect that a cleaning robot equipped with social context can have. This robot has the ability to plan paths taking into account social conventions in public spaces, such as avoiding interfering with the paths of people, avoiding interrupting the interaction between people, prioritizing safety, avoiding surprising people with movements outside the visual range (or any other movement that might make people uncomfortable), navigating with a safe distance and with a prudent speed, avoiding collisions and predicting the intentions of people. With socially-aware navigation models, robots can fulfill the main task and act socially with predictable actions. The goal of including social context into the navigation model is to ensure that robots are not perceived as dangerous, bothersome, irritating, inconvenient, or obtrusive. The sociability of the cleaning robots can be defined as low or indirect, i.e., humans do not communicate with the robot. However, the interaction is generated by the navigation model in a socially acceptable manner. Social navigation models allow the robot to achieve the main goal without disturbing people sharing the same space. Consequently, the robot can operate in public spaces during the entire opening hours.

Specifically, if we employ the model (Patompak et al., 2019) presented in section 3.3 as the *learning* component in our framework, the personalized size and shape of the personal

zone can in fact improve the social intelligence of the robot. By avoiding crossing the comfort zone of people, these robots can learn to plan paths without disturbing the visitors of the shopping mall while performing the cleaning task. However, the model (Patompak et al., 2019) that takes the gender of a person into account can induce bias in the decisions. Even though women might prefer a larger comfort area during interaction among humans, it does not necessarily imply that they would prefer the same during human-robot interaction. In principle, a robot should never harm or be unfair to people based on their gender. In this work, we consider that the robot is depicted as a gender-neutral machine. Conforming a robot to a specific gender depending on the application could again lead to historical bias, this is an area that requires further research which is out of the scope of this paper. Moreover, according to the *bias evaluation* consideration for fairness described in section 2.2, maintaining different relative distances to people based on their gender is an unacceptable bias. Furthermore, distinguishing the comfort area by gender is not of high relevance to improve the acceptance or beneficial to improve the operation of robots around humans. Instead, there are other essential factors that can be used to improve comfort and confidence, such as safe navigation policies. Given that the bias presented in this case is explicit, it is easier to identify the bias inducing factor influencing the model in the *relearning* component of our framework, for example, by correlating the obtained behavior to the input constraints. After detecting the bias inducing factor, it can be excluded to re-train the model without the gender constraint.

On the other hand, while learning from demonstrations, data-driven models can also reflect negative bias. For instance, if robots learn from data that is not diverse where people with movement impairments are not present, then the robot might not react in a socially acceptable manner when they encounter such people. This can further lead to incorrect prediction of paths of people who walk slower and can make the robot be perceived as obtrusive. Data induced bias represents an implicit bias in the model that is more challenging to detect and correct for. Since the model disproportionately affects a specific group of people, by using our *relearning* component, the recurrent errors in the path prediction can be detected as a cluster that can also be related to the set of protected characteristics (e.g., people with mobility impairment). Consequently, by using a punishment system, the reward value is influenced after the detection of unwanted behavior to adjust the learning policy, allowing model adaptation toward a more fair behavior. This will support the *Value Alignment* consideration presented in section 2.2 in which accepted socially-aware robot navigation also considers inclusion.

4.2. Guidance Robots in a Shopping Mall

Mobile service robots have extensive use in innovative applications, such as for guidance in public spaces where they navigate alongside people and assist them to reach their desired destination. Based on the environment described in section 4.1, in this case study we analyze the effects of a guidance robot that operates in a shopping mall. Unlike the last scenario, the robot not only navigates under social conventions but also guides a person in a social manner. The task of the robot is to provide the requested information about locations in the shopping mall and accompany people to reach their desired location. This scenario is illustrated in **Figure 5**. Apart from guiding to reach a certain destination, the robot should also navigate considering social conventions that are required to provide comfort to all the surrounding people during navigation. Furthermore, the robot should coordinate with the user while navigating by maintaining a desired relative position with respect to the user. This scenario has similar characteristics to the mall in the previous case study where diverse people with different genders, ethnicity, disabilities, age, skin tones, and cultural origins and etc, will be present. In this example, fairness considerations, such as shared benefit, deterrence and value alignment described in the section 2.2 should be considered. Additionally, in the shopping mall scenario, the guidance robot will interact naturally with the user in a socially compliant manner while providing information and route guidance.

The human-robot interaction in this case is direct given that people approach the robot with a specific intention, and they expect a response from the robot that corresponds to the request. The resulting navigation strategy that these robots have next to people and their capacity to react according to the situation is crucial for their acceptance. Some of the important constraints in the navigation behavior of guidance robots are adapting the speed of the robot to the user, and maintaining a relative position and distance. If the robot navigates with a velocity that does not correspond to the user, then the robot risks being too slow or too fast which can cause uncoordinated behavior with the user and can further lead to accidents. On the other hand, relative distance and position are related to how people follow the robot and how the robot guides the user. Ideally, the robot should estimate the position and intention of the user during the execution of the guidance and also be able to interrupt the task if the person does not require any more help. Therefore, robots should adapt their navigation based on speed, intentions, motivations, orientation as well as handle unexpected situations, such as people crossing their path, changes in the speed of the person being guided, unexpected appearance of objects, among others.

Consumers value the unbiased, fast, and error-free behavior that a robot can provide. Therefore, the robot should adapt its behavior according to the current social context. In contrast to the interaction between people and cleaning robots, guidance robots provide personalized interaction, so the degree of sociability of this robot is greater. For example, if a disabled person goes to a shopping mall, the robot should recognize that this person will have different navigation behaviors than others so it should adapt its strategy accordingly. This adaptation will in turn make the person more comfortable using the assistance provided by the robot. In this example, aspects, such as the capability to recognize mobility impairments in a person and navigate accordingly are essential to ensure safe and comfortable guidance. Consider that a person with limited mobility requires guidance from the robot. If the robot is not equipped to react accordingly to mobility difficulties, the interaction can cause

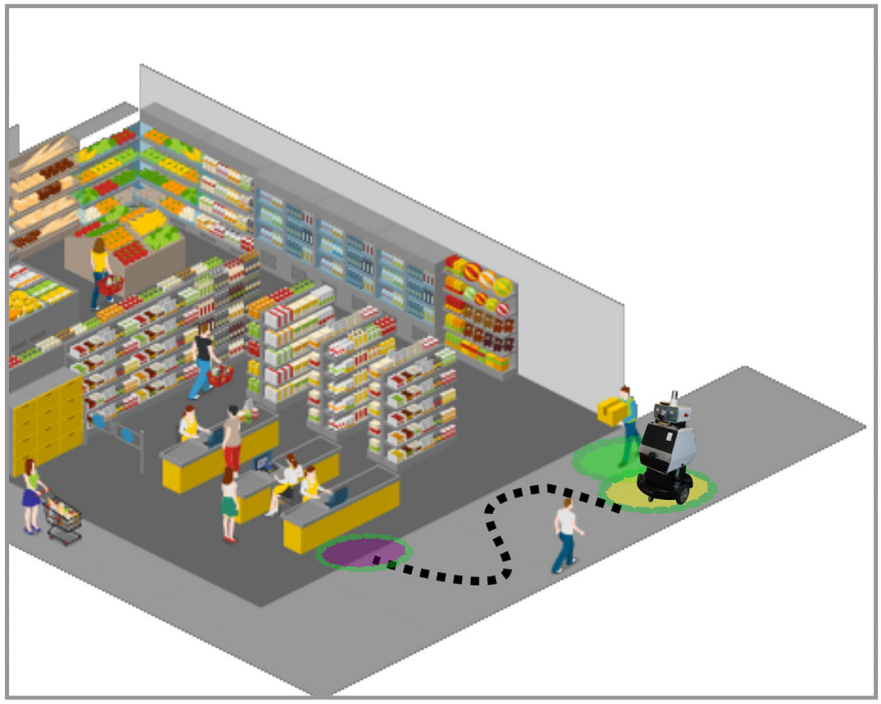

FIGURE 5 | Illustration of the guidance robot in a shopping mall scenario. The robot guides the user (green circle) to reach the destination (purple circle). Additionally, the robot is aware of the people in the surroundings during navigation while maintaining a desired relative position with respect to the user.

distress, physical overexertion, and even accidents. This will eventually make the person to discontinue using the robot in the future. In order to avoid such events, the navigation model in the robot should incorporate social adaptability skills that enable it to detect particular situations that cause discomfort or unintended outcomes for specific individuals.

Assume that a guidance robot is equipped with the navigation model described in section 3.3 and as a consequence it will assist women keeping larger distances with them. This may cause the robot to loose the interaction with them in certain situations and adversely affect the way that women perceive the robot. Similarly, it can reduce the efficiency with this population group representing the systematic disadvantage we aim to avoid toward diminishing bias. The model described in section 3.3 is used to present an example of learning socially-aware robot navigation in which unfair outcomes are associated with a protected characteristic. Other socially-aware navigation models that learn solely from human imitation can cause different types of model-induced biases. In these cases, the navigation model is optimized to yield sociable actions considering different factors, such as the velocity, orientation, priority of interaction, and route selection. The guidance robot will encounter situations where multiple people request for help simultaneously or even situations where people will try to interact with the robot when it is already guiding another person. Deciding which person has the priority is part of the social intelligence. Assume that in the *learning* component of our framework, the navigation model of the robot is trained from demonstrations and as a result, the robot learns the preferred interaction behavior based on those demonstrated interactions. This can lead to unfair outcomes due to human bias that may be existing in the demonstrations, policies reflecting personal bias, unequal society roles, or under-representation of minorities. Specifically, if the learning from demonstration is performed in a shopping mall only from one city, there will be insufficient diversity. Similarly, if the robot is deployed in a different place, or when people belonging to minorities try to use the robot, the robot will maintain its social behavior but it will likely make biased decisions, especially against people who historically have been discriminated, as we observed in other cases (Buolamwini and Gebru, 2018; Brandao, 2019; Wilson et al., 2019; Prabhu and Birhane, 2020). As part of the relearning component, our framework allows to generate clusters related to preferred interaction actions and determine if the generated clusters are strongly related to protected characteristics. Specifically, in case the preferred interaction of the robot is biased favoring or disadvantaging specific visitors of the shopping mall the learning policy is adjusted by a reward value that is penalized when biased behavior is detected. As a consequence, the robot's actions, such as deciding which person has the priority to interact with will follow the fairness requirements.

Since diverse people typically visit shopping malls, the robot should be able to accurately recognize them regardless of factors, such as skin tones. Previous studies (Wilson et al., 2019) have shown that recognition systems based on RGB perception present higher error rates for dark skin tones.

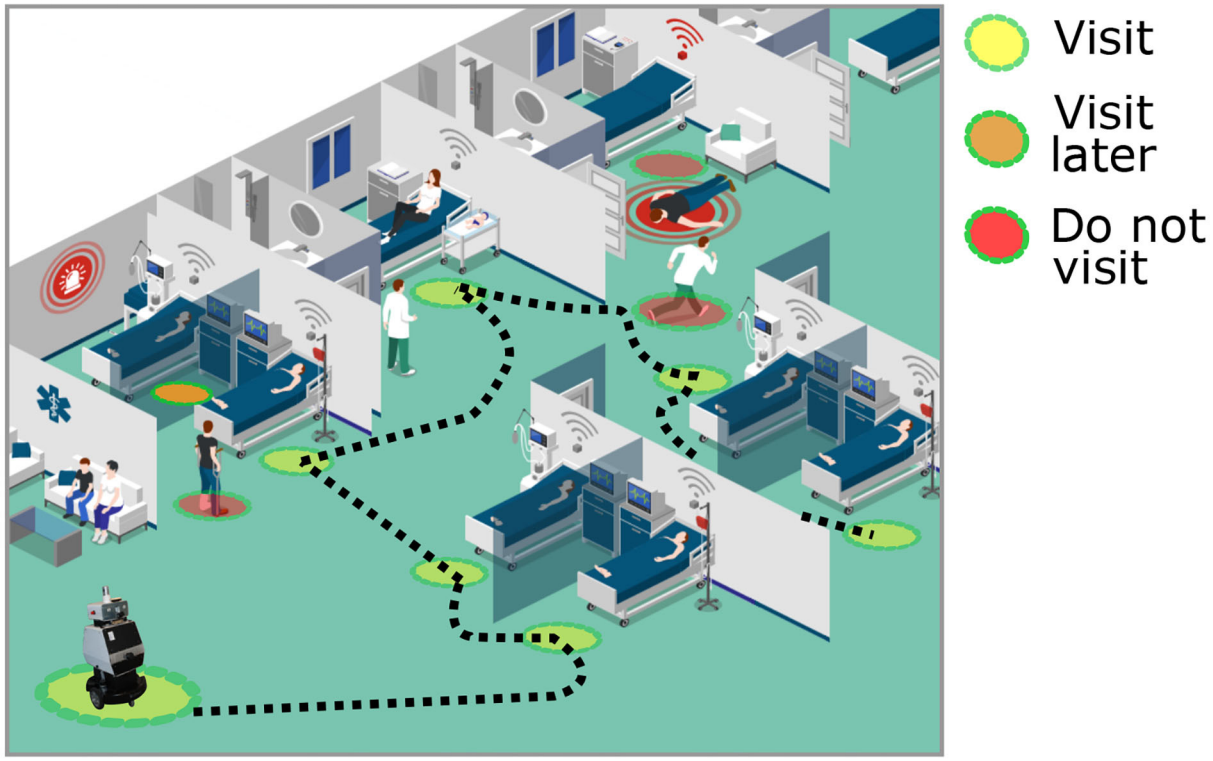

FIGURE 6 | Illustration of the caregiving robot in a hospital scenario. The main task of the robot is to distribute medicines to patients who are admitted in the hospital. The robot takes emergency situations that could happen into account and people requiring special assistance, while navigating.

If similar systems with faulty sensors or algorithms are used to learn social navigation models, the robot will be unable to recognize certain people and adhere to the fairness considerations described in section 2.2. As a consequence, the robot can perpetuate discrimination against groups of people that have historically been segregated, as observed in other learning applications, such as the automated risk assessment used by U.S judges and the biased vision-based object detectors employed in autonomous cars (Benthall and Haynes, 2019; Wilson et al., 2019). Furthermore, discrimination laws prohibit unfair treatment of people based on race. In this case, fairness priority is also important for the legal framework.

4.3. Caregiving Robots in Hospitals

There is significant interest in developing service robots for hospitals due to their ability to provide care for people. The use of robots in hospitals can be especially advantageous in cases where there are patients with contagious diseases, such as in a pandemic situation. In this case study, we analyze the navigation strategy of caregiving robots that operate in hospitals. The main task of robots in this case study is to distribute medicines to patients who are admitted in a hospital. **Figure 6** illustrates this scenario. The human-robot interaction in hospitals requires special caution as the robot will operate around patients who require special assistance. One such example is people with motion impairments who use wheelchairs, crutches, or walking frames. Furthermore, the robot will encounter rapidly changing situations, for example during an emergency where doctors and care staff rush through the hallways. To provide appropriate response, robots should be equipped with algorithms to understand situations and context that enable them to accordingly adapt their behavior. Apart from patients, robots will also interact with other people in the hospital, such as health professionals, secretaries, family members, and visitors. Similar to the shopping mall case study, caregiving robots will be interacting directly with the people. However, the navigation and interaction presents additional complexity, given that they do not assist people individually. Here, the robots aim to assist multiple people who have different medical treatments and deliver medicine to them while maintaining a socially accepted behavior. In this case, not only social conventions and sociability described in the previous case studies are required, but also priority decision making, optimal recognition, faster reaction and adaptability. As a consequence, the navigation models in caregiving robots should have higher requirements of accuracy and adaptability. These robots can particularly encounter unexpected events, such as emergency situations where people will be walking in different directions, speeds, and unpredictable movements. In such situations, there is a higher risk of accidents due to the vulnerability of people and the context in the hospital. Furthermore, the consequences of eventual accidents can be critical for the health of individuals. Caregiving robots should be able to perceive, recognize, and react according to the special requirements of the hospital.

Assume that the robots are going to be used in emergency rooms. Their task there is to deliver a series of necessary supplies to the people who are attending to the emergencies. Therefore, the robots have to interact with several people simultaneously. Based on the proxemics model described in section 3.3, the robot will be perceived as atypical in approaching people in different ways, assisting some people differently than others during urgent situations. Furthermore, taking into account that there are people playing specific roles, namely to care for sick people urgently, their comfort area of interaction is different from that of normal situations. People typically tend to walk fast, to have little personal space, and to quickly perceive what is happening around them. In this scenario, robots that navigate while maintaining different distances to people based on gender have lesser foreseeable utility. Alternatively, other characteristics can be considered that are related to the distribution of medicines depending on the needs of the patients and priorities, such as minimizing delivery time.

The priority of the path planning algorithms in such robots is to deliver medicines to all patients. Assume that in the *learning component* the caregiving robot learns from historical data about the characteristics of the patients. This model may learn that the pain threshold differs between men and women. Consequently, the navigation plan will be biased with negative effects toward men, based on information related to their higher tolerance to pain. Similarly, the robot could learn that women have more tolerance to wait longer for medical treatments and spend more overall time than men in the emergency rooms (Nottingham et al., 2018). In both situations, the behavior of the robot will be biased given that it systematically benefits a specific group of people. In this example, fairness considerations, such as value alignment and non-maleficence described in the section 2.2 can improve the decisions made by the robot. One approach to dealing with difficult cases of priority is to reflect political and commercial neutrality in robot navigation. This signifies that the navigation model in caregiving robots should not favor any particular group of people. Although, advocating for neutrality of assistive robots is a potential solution to bias problems in this case, the concept is substantially complex and requires further research.

Particularly, adapting the model with our *relearning* component to correct for the presented bias will lead the robot to base decisions on other factors. Using the *relearning* component of our framework, we can identify clusters that demonstrate a systematic disadvantage if the time to deliver medicines is higher for men and if women wait for a longer period of time in emergency rooms. Subsequently, to penalize the unfair behavior, we lower the reward value that adjusts the learning policy. As a result, the navigation model is adapted toward more fair behavior. If the model does not rely on the potentially negative bias inducing factors, it can learn better representations that reflect relevant characteristics, such as urgency and needs. While using our relearning technique, this type of bias in navigation will be detected when certain people receive attention more effectively than others. Consequently, if there is no valid reasoning behind such bias, the navigation model should be updated accordingly.

5. CONCLUSIONS

As more and more robots navigate in human spaces, they also require more complex navigation models to accomplish their goals while complying with the high safety and comfort requirements. Toward this direction, different methods incorporate social context into learning models to enable robots to navigate following social conventions. Typically, these methodologies utilize data or experiences from the real world, simulations, or control experiments and social constraints. In this work, we discussed the societal and ethical implications of learned socially-aware robot navigation techniques. We demonstrated that the advances accomplished in social robot navigation are essential for the development of robots that provide well for society. More importantly, we showed how these models that account for socially-aware robot navigation do not guarantee fairness in different real-world scenarios. Research in the direction of fairness in robot learning is of special importance, given that these machines interact with people closely.

To the best of our knowledge, this is the first work that studies the societal implications of bias in learned socially-aware robot navigation models. Our proposed framework that consists of the learning and relearning stages has the ability to effectively diminish bias in social robot navigation models. Additionally, we presented fairness considerations and specific techniques that can be used to implement our framework. We detailed several scenarios that show that the adaptability of the model in terms of fairness enables it to correct for bias. The scenarios demonstrate the potential unwanted outcomes of social navigation models that are described with variables and social conventions which make them easily interpretable. Our framework is especially useful for more complex learning models or models that are trained with imitation or reinforcement learning, given that these models contain more abstract representations of the data and situations. We hope this work contributes toward raising awareness on the importance of fairness in robot learning.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

All authors contributed to the concepts, analysis, and drafting the manuscript.

FUNDING

This work was partly funded by the BrainLinks-Brain Tools center of the University of Freiburg, a scholarship from the Graduate School of Robotics of the University Freiburg (according to the Graduate Funding Law of the Ministry of Science, Research and Arts of the State of Baden-Württemberg), and a grant from the Eva Mayr-Stihl Stiftung.

REFERENCES

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. *arXiv* 1803.02453.

Aljalbout, E., Golkov, V., Siddiqui, Y., Strobel, M., and Cremers, D. (2018). Clustering with deep learning: taxonomy and new methods. *arXiv* 1801.07648.

Anderson, M., and Anderson, S. L. (2010). Robot be good. *Sci. Am.* 303, 72–77. doi: 10.1038/scientificamerican1010-72

Argyle, M., Cook, M., and Cramer, D. (1994). Gaze and mutual gaze. *Br. J. Psychiatry* 165, 848–850. doi: 10.1017/S0007125000073980

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inform. Fusion* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012

Benthall, S., and Haynes, B. D. (2019). "Racial categories in machine learning," in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (New York, NY), 289–298. doi: 10.1145/3287560.3287575

Bicchi, A., and Tamburrini, G. (2015). Social robotics and societies of robots. *Inform. Soc.* 31, 237–243. doi: 10.1080/01972243.2015.1020196

Binns, R. (2018). "Fairness in machine learning: lessons from political philosophy," in *Conference on Fairness, Accountability and Transparency* (New York, NY: PMLR), 149–159.

Birdwhistell, R. L. (1952). *Introduction to Kinesics: An Annotation System for Analysis of Body Motion and Gesture*. Michigan: Department of State, Foreign Service Institute.

Birdwhistell, R. L. (2010). *Kinesics and Context: Essays on Body Motion Communication*. Pennsylvania, PA: University of Pennsylvania Press.

Birhane, A., and Cummins, F. (2019). Algorithmic injustices: towards a relational ethics. *arXiv* 1912.07376.

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., et al. (2017). Principles of robotics: regulating robots in the real world. *Connect. Sci.* 29, 124–129. doi: 10.1080/09540091.2016.1271400

Bogue, R. (2016). Search and rescue and disaster relief robots: has their time finally come? *Ind. Robot* 43, 138–143. doi: 10.1108/IR-12-2015-0228

Boniardi, F., Valada, A., Burgard, W., and Tipaldi, G. D. (2016). "Autonomous indoor robot navigation using sketched maps and routes," in *Workshop on Model Learning for Human-Robot Communication at Robotics: Science and Systems (RSS)* (Michigan). doi: 10.1109/ICRA.2016.7487453

Bonnefon, J. F., Černy, D., Danaher, J., Devillier, N., Johansson, V., Kovacikova, T., et al. (2020). *Ethics of Connected and Automated Vehicles: Recommendations on Road Safety, Privacy, Fairness, Explainability and Responsibility*. Luxembourg: EU Publications.

Brandao, M. (2019). Age and gender bias in pedestrian detection algorithms. *arXiv* 1906.10490.

Brandão, M., Jirtoka, M., Webb, H., and Luff, P. (2020). Fair navigation planning: a resource for characterizing and designing fairness in mobile robots. *Artif. Intell.* 282:103259. doi: 10.1016/j.artint.2020.103259

BSI-2016 (2016). *BS 8611: 2016 Robots and Robotic Devices: Guide to the Ethical Design and Application of Robots and Robotic Systems*. London: British Standards Institution.

Buolamwini, J., and Gebru, T. (2018). "Gender shades: intersectional accuracy disparities in commercial gender classification," in *Conference on Fairness, Accountability and Transparency* (New York, NY), 77–91.

Castro, L., and Toro, M. A. (2004). The evolution of culture: from primate social learning to human culture. *Proc. Natl. Acad. Sci. U.S.A.* 101, 10235–10240. doi: 10.1073/pnas.0400156101

Cath, C. (2018). Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* 376, 1–8. doi: 10.1098/rsta.2018.0080

Chen, Y. F., Everett, M., Liu, M., and How, J. P. (2017). "Socially aware motion planning with deep reinforcement learning," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Vancouver, BC: IEEE), 1343–1350. doi: 10.1109/IROS.2017.8202312

Chouldechova, A., and Roth, A. (2018). The frontiers of fairness in machine learning. *arXiv* 1810.08810.

Claure, H., Chen, Y., Modi, J., Jung, M., and Nikolaidis, S. (2019). Reinforcement learning with fairness constraints for resource distribution in human-robot teams. *arXiv* 1907.00313.

Costa-jussà, M. R. (2019). An analysis of gender bias studies in natural language processing. *Nat. Mach. Intell.* 1, 495–496. doi: 10.1038/s42256-019-0105-5

Dastin, J. (2018). *Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women*. San Fransico, CA: Reuters.

De Santis, A., Siciliano, B., De Luca, A., and Bicchi, A. (2008). An atlas of physical human-robot interaction. *Mech. Mach. Theory* 43, 253–270. doi: 10.1016/j.mechmachtheory.2007.03.003

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (2018). "Measuring and mitigating unintended bias in text classification," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY), 67–73. doi: 10.1145/3278721.3278729

Fehr, E., and Fischbacher, U. (2004). Social norms and human cooperation. *Trends Cogn. Sci.* 8, 185–190. doi: 10.1016/j.tics.2004.02.007

Ferrer, G., Garrell, A., and Sanfeliu, A. (2013). "Social-aware robot navigation in urban environments," in *2013 European Conference on Mobile Robots* (Barcelona: IEEE), 331–336. doi: 10.1109/ECMR.2013.6698863

Ferrer, G., Zulueta, A. G., Cotarelo, F. H., and Sanfeliu, A. (2017). Robot social-aware navigation framework to accompany people walking side-by-side. *Auton. Robots* 41, 775–793. doi: 10.1007/s10514-016-9584-y

Fink, J., Bauwens, V., Kaplan, F., and Dillenbourg, P. (2013). Living with a vacuum cleaning robot. *Int. J. Soc. Robot.* 5, 389–408. doi: 10.1007/s12369-013-0190-2

Fiorini, P., and Prassler, E. (2000). Cleaning and household robots: a technology survey. *Auton. Robots* 9, 227–235. doi: 10.1023/A:1008954632763

Forlizzi, J. (2007). "How robotic products become social products: an ethnographic study of cleaning in the home," in *2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (New York, NY: IEEE), 129–136. doi: 10.1145/1228716.1228734

Forlizzi, J., and DiSalvo, C. (2006). "Service robots in the domestic environment: a study of the roomba vacuum in the home," in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction* (New York, NY), 258–265. doi: 10.1145/1121241.1121286

Forshaw, S., and Pilgerstorfer, M. (2008). Direct and indirect discrimination: is there something in between? *Ind. Law J.* 37, 347–364. doi: 10.1093/indlaw/dwn019

Fuchs, D. J. (2018). The dangers of human-like bias in machine-learning algorithms. *Missouri S&Ts Peer Peer* 2:1.

Garcia, M. (2016). Racist in the machine: the disturbing implications of algorithmic bias. *World Policy J.* 33, 111–117. doi: 10.1215/07402775-3813015

Gaydashenko, A., Kudenko, D., and Shpilman, A. (2018). "A comparative evaluation of machine learning methods for robot navigation through human crowds," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (Orlando, FL: IEEE), 553–557. doi: 10.1109/ICMLA.2018.00089

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., et al. (2020). Shortcut learning in deep neural networks. *arXiv* 2004.07780. doi: 10.1038/s42256-020-00257-z

Goodman, B., and Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI Mag.* 38, 50–57. doi: 10.1609/aimag.v38i3.2741

Goodwin, C. (2000). Action and embodiment within situated human interaction. *J. Pragmat.* 32, 1489–1522. doi: 10.1016/S0378-2166(99)00096-X

Groshev, E., Goldstein, M., Tamar, A., Srivastava, S., and Abbeel, P. (2017). Learning generalized reactive policies using deep neural networks. *arXiv* 1708.07280.

Grunwald, A. (2011). Responsible innovation: bringing together technology assessment, applied ethics, and STS research. *Enterpr. Work Innov. Stud.* 31, 10–19.

Hagendorff, T. (2020a). Ethical behavior in humans and machines-evaluating training data quality for beneficial machine learning. *arXiv* 2008.11463.

Hagendorff, T. (2020b). The ethics of AI ethics: an evaluation of guidelines. *Minds Mach.* 30, 99–120. doi: 10.1007/s11023-020-09517-8

Hall, E. T., Birdwhistell, R. L., Bock, B., Bohannan, P., Diebold, A. R. Jr., Durbin, M., et al. (1968). Proxemics [and comments and replies]. *Curr. Anthropol.* 9, 83–108. doi: 10.1086/200975

Hamandi, M., D'Arcy, M., and Fazli, P. (2019). "Deepmotion: learning to navigate like humans," in *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (New Delhi: IEEE), 1–7. doi: 10.1109/RO-MAN46459.2019.8956408

Harrigan, J. A. (2005). *Proxemics, Kinesics, and Gaze*. Oxford: Oxford University Press.

Hasan, K. M., Abdullah-Al-Nahid, and Reza, K. J. (2014). "Path planning algorithm development for autonomous vacuum cleaner robots," in *2014 International Conference on Informatics, Electronics & Vision (ICIEV)* (Dhaka: IEEE), 1–6. doi: 10.1109/ICIEV.2014.6850799

Helbing, D., and Molnar, P. (1995). Social force model for pedestrian dynamics. *Phys. Rev. E* 51:4282. doi: 10.1103/PhysRevE.51.4282

Howard, A., Zhang, C., and Horvitz, E. (2017). "Addressing bias in machine learning algorithms: a pilot study on emotion recognition for intelligent systems," in *2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)* (Austin, TX: IEEE), 1–7. doi: 10.1109/ARSO.2017.8025197

Hurtado, J. V., Mohan, R., and Valada, A. (2020). MOPT: multi-object panoptic tracking. *arXiv* 2004.08189.

Hutchins, E. (2006). The distributed cognition perspective on human interaction. *Roots Hum. Soc. Cult. Cogn. Interact.* 1:375. doi: 10.4324/9781003135517-19

Icograms (2020). *Illustrations*. Avaialble online at: https://icograms.com/ (accessed December 18, 2020).

Jamshidi, P., Cámara, J., Schmerl, B., Käestner, C., and Garlan, D. (2019). "Machine learning meets quantitative planning: enabling self-adaptation in autonomous robots," in *2019 IEEE/ACM 14th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS)* (Montreal, QC: IEEE), 39–50. doi: 10.1109/SEAMS.2019.00015

Jarvis, P. (2006). *Towards a Comprehensive Theory of Human Learning*, Vol. 1. New York, NY: Psychology Press.

Johnson, C., and Kuipers, B. (2018). "Socially-aware navigation using topological maps and social norm learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (Tulane), 151–157. doi: 10.1145/3278721.3278772

Johnson, K., Pasquale, F., and Chapman, J. (2019). Artificial intelligence, machine learning, and bias in finance: toward responsible innovation. *Fordham L. Rev.* 88:499.

Jones, C. M., and Healy, S. D. (2006). Differences in cue use and spatial memory in men and women. *Proc. R. Soc. B Biol. Sci.* 273, 2241–2247. doi: 10.1098/rspb.2006.3572

Kalweit, G., Huegle, M., Werling, M., and Boedecker, J. (2020a). "Deep inverse Q-learning with constraints," in *Advances in Neural Information Processing Systems*, 33.

Kalweit, G., Huegle, M., Werling, M., and Boedecker, J. (2020b). Interpretable multi time-scale constraints in model-free deep reinforcement learning for autonomous driving. *arXiv* 2003.09398.

Kaushal, A., Altman, R., and Langlotz, C. (2020). *Health Care AI Systems Are Biased*. Scientific American.

Khambhaita, H., and Alami, R. (2020). "Viewing robot navigation in human environment as a cooperative activity," in *Robotics Research* (Springer), 285–300. doi: 10.1007/978-3-030-28619-4_25

Kirby, R. (2010). *Social robot navigation* (Ph.D. thesis), Carnegie Mellon University, Pittsburgh, PA, United States.

Kivrak, H., Cakmak, F., Kose, H., and Yavuz, S. (2020). Social navigation framework for assistive robots in human inhabited unknown environments. *Eng. Sci. Technol. Int. J.* 24, 284–298. doi: 10.1016/j.jestch.2020.08.008

Kretzschmar, H., Spies, M., Sprunk, C., and Burgard, W. (2016). Socially compliant mobile robot navigation via inverse reinforcement learning. *Int. J. Robot. Res.* 35, 1289–1307. doi: 10.1177/0278364915619772

Kruse, T., Pandey, A. K., Alami, R., and Kirsch, A. (2013). Human-aware robot navigation: a survey. *Robot. Auton. Syst.* 61, 1726–1743. doi: 10.1016/j.robot.2013.05.007

Kuderer, M., Kretzschmar, H., and Burgard, W. (2013). "Teaching mobile robots to cooperatively navigate in populated environments," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems* (Tokyo: IEEE), 3138–3143. doi: 10.1109/IROS.2013.6696802

Lee, N. T. (2018). Detecting racial bias in algorithms and machine learning. *J. Inform. Commun. Ethics Soc.* 16, 252–260. doi: 10.1108/JICES-06-2018-0056

Lin, P., Abney, K., and Bekey, G. A. (2012). *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge: Intelligent Robotics and Autonomous Agents Series.

Liu, H. Y., and Zawieska, K. (2017). From responsible robotics towards a human rights regime oriented to the challenges of robotics and artificial intelligence. *Ethics Inform. Technol.* 22, 321–333. doi: 10.1007/s10676-017-9443-3

Lu, K., Mardziel, P., Wu, F., Amancharla, P., and Datta, A. (2020). "Gender bias in neural natural language processing," in *Logic, Language, and Security* (Cham: Springer), 189–202. doi: 10.1007/978-3-030-62077-6_14

Luber, M., Spinello, L., Silva, J., and Arras, K. O. (2012). "Socially-aware robot navigation: a learning approach," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems* (Vilamoura-Algarve: IEEE), 902–907. doi: 10.1109/IROS.2012.6385716

McDonald, D. W., McCarthy, J. F., Soroczak, S., Nguyen, D. H., and Rashid, A. M. (2008). Proactive displays: supporting awareness in fluid social environments. *ACM Trans. Comput. Hum. Interact.* 14, 1–31. doi: 10.1145/1314683.1314684

Mittal, M., Mohan, R., Burgard, W., and Valada, A. (2019). Vision-based autonomous UAV navigation and landing for urban search and rescue. *arXiv* 1906.01304.

Nelson, G. S. (2019). Bias in artificial intelligence. *North Carolina Med. J.* 80, 220–222. doi: 10.18043/ncm.80.4.220

Nolfi, S., and Floreano, D. (2002). Synthesis of autonomous robots through evolution. *Trends Cogn. Sci.* 6, 31–37. doi: 10.1016/S1364-6613(00)01812-X

Nottingham, Q. J., Johnson, D. M., and Russell, R. S. (2018). The effect of waiting time on patient perceptions of care quality. *Qual. Manage. J.* 25, 32–45. doi: 10.1080/10686967.2018.1404368

Ötting, S. K., Gopinathan, S., Maier, G. W., and Steil, J. J. (2017). "Why criteria of decision fairness should be considered in robot design," in *20th ACM Conference on Computer-Supported Cooperative Work and Social Computing* (Portland, OR).

Patompak, P., Jeong, S., Nilkhamhang, I., and Chong, N. Y. (2019). Learning proxemics for personalized human-robot social interaction. *Int. J. Soc. Robot.* 12, 267–280. doi: 10.1007/s12369-019-00560-9

Perez, S. (2016). *Microsoft Silences Its New AI Bot Tay, After Twitter Users Teach It Racism*. Tech Crunch.

Piano, S. L. (2020). Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Human. Soc. Sci. Commun.* 7, 1–7. doi: 10.1057/s41599-020-0501-9

Poudel, D. B. (2013). Coordinating hundreds of cooperative, autonomous robots in a warehouse. *AI Mag.* 27, 1–13.

Prabhu, V. U., and Birhane, A. (2020). Large image datasets: a pyrrhic win for computer vision? *arXiv* 2006.16923.

Reed, C., Kennedy, E., and Silva, S. (2016). *Responsibility, Autonomy and Accountability: Legal Liability for Machine Learning*. London: Queen Mary School of Law Legal Studies Research Paper.

Research, I. (2019). *Floor Cleaning Robot Market by Robot Type, by Sales Channel, by Region–Global Forecast Up to 2025*. Research and Markets.

Riek, L., and Howard, D. (2014). "A code of ethics for the human-robot interaction profession," in *Proceedings of We Robot*.

Rios-Martinez, J., Spalanzani, A., and Laugier, C. (2015). From proxemics theory to socially-aware navigation: a survey. *Int. J. Soc. Robot.* 7, 137–153. doi: 10.1007/s12369-014-0251-1

Silberg, J., and Manyika, J. (2019). *Notes From the AI Frontier: Tackling Bias in AI (and in Humans)*. Mckinsey Global Institute.

Silver, D., Bagnell, J. A., and Stentz, A. (2010). Learning from demonstration for autonomous navigation in complex unstructured terrain. *Int. J. Robot. Res.* 29, 1565–1592. doi: 10.1177/0278364910369715

Simmel, G. (1949). The sociology of sociability. *Am. J. Sociol.* 55, 254–261. doi: 10.1086/220534

Söderström, M. (2001). Why researchers excluded women from their trial populations. *Lakartidningen* 98, 1524–1528.

Stilgoe, J., Owen, R., and Macnaghten, P. (2013). Developing a framework for responsible innovation. *Res. Policy* 42, 1568–1580. doi: 10.1016/j.respol.2013.05.008

Tai, L., Zhang, J., Liu, M., and Burgard, W. (2018). "Socially compliant navigation through raw depth inputs with generative adversarial imitation learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)* (Brisbane: IEEE), 1111–1117. doi: 10.1109/ICRA.2018.8460968

Tewari, A., Peabody, J., Sarle, R., Balakrishnan, G., Hemal, A., Shrivastava, A., et al. (2002). Technique of Da Vinci robot-assisted anatomic radical prostatectomy. *Urology* 60, 569–572. doi: 10.1016/S0090-4295(02)01852-6

Thrun, S. (1995). An approach to learning mobile robot navigation. *Robot. Auton. Syst.* 15, 301–319. doi: 10.1016/0921-8890(95)00022-8

Thrun, S., Schulte, J., and Rosenberg, C. (2000). "Interaction with mobile robots in public places," in *IEEE Intelligent Systems*, 7–11.

Torresen, J. (2018). A review of future and ethical perspectives of robotics and AI. *Front. Robot. AI* 4:75. doi: 10.3389/frobt.2017.00075

Toupet, O., Biesiadecki, J., Rankin, A., Steffy, A., Meirion-Griffith, G., Levine, D., et al. (2020). Terrain-adaptive wheel speed control on the curiosity mars rover: algorithm and flight results. *J. Field Robot.* 37, 699–728. doi: 10.1002/rob.21903

Ulrich, I., and Borenstein, J. (2001). The guidecane-applying mobile robot technologies to assist the visually impaired. *IEEE Trans. Syst. Man Cybern. A Syst. Hum.* 31, 131–136. doi: 10.1109/3468.911370

Valada, A., Tomaszewski, C., Kannan, B., Velagapudi, P., Kantor, G., and Scerri, P. (2012). "An intelligent approach to hysteresis compensation while sampling using a fleet of autonomous watercraft," in *International Conference on Intelligent Robotics and Applications* (Montreal, QC: Springer), 472–485. doi: 10.1007/978-3-642-33515-0_47

Vandemeulebroucke, T., de Casterlé, B. D., and Gastmans, C. (2020). Ethics of socially assistive robots in aged-care settings: a socio-historical contextualisation. *J. Med. Ethics* 46, 128–136. doi: 10.1136/medethics-2019-105615

Vayena, E., Blasimme, A., and Cohen, I. G. (2018). Machine learning in medicine: addressing ethical challenges. *PLoS Med.* 15:e1002689. doi: 10.1371/journal.pmed.1002689

Verbeek, P. P. (2008). "Morality in design: design ethics and the morality of technological artifacts," in *Philosophy and Design* (Dordrecht: Springer), 91–103. doi: 10.1007/978-1-4020-6591-0_7

Wang, Q., Xu, Z., Chen, Z., Wang, Y., Liu, S., and Qu, H. (2020). Visual analysis of discrimination in machine learning. *IEEE Trans. Vis. Comput. Graph.* 27, 1470–1480. doi: 10.1109/TVCG.2020.3030471

Watkins, C. J., and Dayan, P. (1992). Q-learning. *Mach. Learn.* 8, 279–292. doi: 10.1007/BF00992698

Wilson, B., Hoffman, J., and Morgenstern, J. (2019). Predictive inequity in object detection. *arXiv* 1902.11097.

Winner, L. (1978). *Autonomous Technology: Technics-Out-of-Control as a Theme in Political Thought*. Cambridge: MIT Press.

Wittrock, M. C. (2010). Learning as a generative process. *Educ. Psychol.* 45, 40–45. doi: 10.1080/00461520903433554

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. (2017). Learning non-discriminatory predictors. *arXiv* 1702.06081.

Yu, A. (2019). Direct discrimination and indirect discrimination: a distinction with a difference. *WJ Legal Stud.* 9:1. doi: 10.5206/uwojls.v9i2.8072

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. (2017). "Fairness constraints: mechanisms for fair classification," in *Artificial Intelligence and Statistics* (Fort Lauderdale, FL: PMLR), 962–970.

Zhang, L., Wu, Y., and Wu, X. (2016). A causal framework for discovering and removing direct and indirect discrimination. *arXiv* 1611.07509. doi: 10.24963/ijcai.2017/549

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Robotics and AI

ORIGINAL RESEARCH
published: 26 April 2021
doi: 10.3389/frobt.2021.627958

Do Privacy Concerns About Social Robots Affect Use Intentions? Evidence From an Experimental Vignette Study

Christoph Lutz[1]* *and Aurelia Tamò-Larrieux*[2]

[1] Nordic Centre for Internet and Society, BI Norwegian Business School, Oslo, Norway, [2] Institute for Work and Employment Research, University of St. Gallen, St. Gallen, Switzerland

While the privacy implications of social robots have been increasingly discussed and privacy-sensitive robotics is becoming a research field within human–robot interaction, little empirical research has investigated privacy concerns about robots and the effect they have on behavioral intentions. To address this gap, we present the results of an experimental vignette study that includes antecedents from the privacy, robotics, technology adoption, and trust literature. Using linear regression analysis, with the privacy-invasiveness of a fictional but realistic robot as the key manipulation, we show that privacy concerns affect use intention significantly and negatively. Compared with earlier work done through a survey, where we found a robot privacy paradox, the experimental vignette approach allows for a more realistic and tangible assessment of respondents' concerns and behavioral intentions, showing how potential robot users take into account privacy as consideration for future behavior. We contextualize our findings within broader debates on privacy and data protection with smart technologies.

Keywords: social robots, privacy, trust, social influence, privacy paradox, survey

OPEN ACCESS

Edited by:
Martim Brandão, King's College London, United Kingdom

Reviewed by:
Meg Tonkin, University of Technology Sydney, Australia
Ina Schiering, Ostfalia University of Applied Sciences, Germany
Trenton Schulz, Norwegian Computing Center, Norway

***Correspondence:**
Christoph Lutz
christoph.lutz@bi.no

Specialty section:
This article was submitted to Ethics in Robotics and Artificial Intelligence, a section of the journal Frontiers in Robotics and AI

Received: *10 November 2020*
Accepted: *25 February 2021*
Published: *26 April 2021*

Citation:
Lutz C and Tamò-Larrieux A (2021) Do Privacy Concerns About Social Robots Affect Use Intentions? Evidence From an Experimental Vignette Study. Front. Robot. AI 8:627958. doi: 10.3389/frobt.2021.627958

INTRODUCTION

With the increasing interaction among humans and social robots (Fong et al., 2003; Gupta, 2015; Van den Berg, 2016), research on the benefits and concerns of close human–machine interaction has emerged. A field of research that has gained traction in recent years describes the privacy implications of social robots (cf. for an overview Lutz et al., 2019). This topic is particularly pressing because social robots tend to exhibit greater mobility, social presence, and autonomy than static devices (Calo, 2012; Kaminski, 2015; Lutz and Tamò, 2015, 2018; Sedenberg et al., 2016; Kaminski et al., 2017; Rueben et al., 2017a, 2018; Fosch-Villaronga et al., 2020). While research on privacy and social robotics has largely remained conceptual and has taken a critical approach to the data processing and privacy implications of social robots, a few studies provide quantitative evidence on the privacy concerns and implications of social robots (Lutz et al., 2019). However, initial survey-based studies have analyzed the existence of a robot privacy paradox (Lutz and Tamò-Larrieux, 2020), the trust implications of social robots (Alaiad and Zhou, 2014), as well as general attitudes toward them (Eurobarometer, 2012; Liang and Lee, 2017).

In this article, we aim to deepen our understanding of privacy in the context of social robots. We therefore present the results of an experimental vignette survey that assessed non-experts' privacy concerns about social robots and how these concerns affect use intention. The findings indicate

that privacy matters. Individuals who are exposed to a more privacy-friendly robot, with the same functionality as a non-privacy-friendly robot, have significantly higher use intentions, even after controlling for relevant variables such as demographics, trusting beliefs, social influence, and general opinions about robots. Thus, our study furthers knowledge in the area of privacy-sensitive robotics (Rueben et al., 2018).

We start by describing the term "privacy" and point to a rich literature on the topic of privacy in the context of social robots. The literature review calls for a holistic understanding of the concept of privacy and embeds the topic in the human–robot interaction literature. We then describe the research model for the empirical study. An overview of the research method, including the sample, data analysis approach, and measurement, is followed by a description of the results. Subsequently, we discuss the findings, address the limitations of our approach, and contextualize the results.

LITERATURE REVIEW

Social Robots and Privacy Concerns

The introduction of new technologies has, throughout history, triggered a response in privacy scholarship (Warren and Brandeis, 1890; Calo, 2012; Finn et al., 2013). We can thus rely on a rich academic tradition of privacy scholarship when analyzing the privacy implications and concerns of social robots (Warren and Brandeis, 1890; Westin, 1967; Altman, 1975; Bygrave, 2002; Solove, 2008; Finn et al., 2013; Kaminski, 2015; Koops et al., 2016; Kaminski et al., 2017). While these discussions have had strong roots in the legal field, privacy research has become a multidisciplinary topic with various disciplines—from communication, computer science, psychology, sociology, and economy—collaborating with each other (Pavlou, 2011). This multitude of perspectives is very much welcome, yet also shows that defining a common notion of what privacy is remains difficult if not impossible (Solove, 2008). The difficulty arises not only out of the multitude of perspectives but also due to subjective and cultural differences and perceptions on privacy (Krasnova et al., 2012; Trepte et al., 2017). The cultural differences result also in different legal approaches of protecting informational privacy, with international agreements shaping their material and territorial scopes (Greenleaf, 2014; Greenleaf and Cottier, 2020).

Nonetheless, useful privacy categorizations and classifications exist. It is interesting to note that the literature conceptualizing privacy has often looked backward, describing how new technologies impact private and social life and finding remedies to address them (e.g., Warren and Brandeis, 1890). Newer scholarship (notably: Finn et al., 2013; Koops et al., 2016) provides more forward-looking frameworks by elaborating on the impact of newer technologies. These frameworks or taxonomies build upon the rich Western privacy literature. One important dimension here is the idea of "zones," i.e., differentiating between more personal zones and more public ones (Koops et al., 2016). While the dichotomy between private and public spheres has been criticized in light of the increased pervasiveness of technology (Nissenbaum, 2004; Rouvroy, 2008; Acharya, 2015), different dimensions of privacy have been proposed (Rueben et al., 2017a; Lutz et al., 2019). One dimension deals with physical privacy concerns as the concerns relating to an individual's personal space (Finn et al., 2013). Such an understanding of privacy was already propagated by Warren and Brandeis (1890) and revolves around "physical access to an individual and/or the individual's surroundings and private space" (Smith et al., 2011, p. 990). Physical privacy concerns become especially apparent with the use of social robots at home due to the robot's ability to enter (uninvited) into private spaces (e.g., bathrooms, bedrooms) (Calo, 2012). However, new technologies, such as genetic codes and smart health tracking technologies (e.g., pills), have resulted in stronger demands for physical privacy. Proposals include the privacy of the person, which includes "the right to keep body functions and body characteristics private" (Finn et al., 2013, p. 8).

A second key dimension revolves around informational privacy concerns (Smith et al., 2011). At its core, informational privacy should enable individuals to have control about their information (Westin, 1967), thereby reducing institutional privacy threats by data-processing institutions (e.g., robot manufacturers, government agencies, and third parties such as data brokers or cloud providers) as well as social threats occurring by the processing of information by private individuals (e.g., familiar users, hackers) (Raynes-Goldie, 2010; Young and Quan-Haase, 2013). These aspects point to a core concern, namely, surveillance enabled by social robots that are equipped with innovative sensors and processors, enabling greater observation and profiling of individuals (Calo, 2012). In light of these technological changes, Koops et al. (2016) call for intellectual, decisional, associational, and behavioral privacy to ensure the self-development of individuals. Similarly, Finn et al. (2013) include in their seven types of privacy at least three types that are linked to informational privacy concerns, such as the privacy of personal behaviors and actions (including the revelation of sensitive habits and sexual orientation), the privacy of communication, and the privacy of data and images. All these types of information can be collected or disseminated through social robots. Similarly, and the reason why informational privacy concerns are closely tied to the ones mentioned below under boundary management, emerging technologies such as social robots will likely impact a user's privacy of thought and feelings (Finn et al., 2013). In addition, the way automated decision-making systems classify information about individuals and reach decisions (by correlations and pattern finding) affects a new class of privacy, namely, privacy of associations (Finn et al., 2013).

Closely tied are boundary management approaches, understanding privacy as a "selective control of access to the self or to one's group" (Altman, 1975, p. 18). This understanding of privacy as boundary management (Petronio, 2002) links back the discussion to the physical privacy concerns mentioned. However, boundary management approaches must be understood more broadly than pure "freedom from" and physical protection claims (Koops et al., 2016), as they put individuals and their agency to make own life choices at the center about when their privacy is unreasonably constrained (Carnevale, 2016).

Agency requires understanding how information within social robots and various stakeholders is shared. In addition, research building on the boundary management literature indicates that the design of smart environments (e.g., setting of sensors and cameras), including ones with acting social robots in homes, impacts how these data-processing devices are perceived and privacy boundaries are negotiated (Schulz and Herstad, 2018; Schulz et al., 2018). The boundary management negotiations are highly dependent on the affordances of technologies (e.g., ability to turn sensors on and off) and the visibility of certain functionalities (e.g., surveillance through camera). Moreover, the anthropomorphic or zoomorphic effect of social robots (Fong et al., 2003; Weiss et al., 2009; Darling, 2016) may increase the pervasiveness of social robots (Turkle, 2011) and the bonding between individual and robot may inhibit rational and privacy-oriented considerations by individuals (Syrdal et al., 2007; Calo, 2012).

Previous Research on Privacy and Social Robots

While there is a rich literature on human–robot interaction across disciplines (for an overview, see Baxter et al., 2016), research on privacy and social robots is still a comparatively nascent field. Early empirical studies on privacy concerns in the context of social robots have explored by means of qualitative interviews how individuals perceive the use of social robots in the work environment (e.g., Snackbot, see Lee et al., 2011). The study of Lee et al. (2011) revealed that most participants did not understand what data categories Snackbot collected and failed to differentiate between sensed data ("what the robot sees/hears") and inferred information ("what the robot knows", p. 182). Moreover, the anthropomorphic shape of Snackbot sometimes misled the participants' notion of the capabilities of the robot to record information (e.g., participants did not consider the ability of the robot to sense objects behind it).

Other empirical research has focused on concern related to information disclosure in human–robot interactions. For instance, in one study, participants stated that they overcame their fear of robots storing and accessing sensitive information about them because such processing activities were necessary (and thus tolerated) in order to benefit from the social robot's functions (Syrdal et al., 2007).

Other studies analyzed the privacy-utility tradeoff further, for instance, in the domain of teleoperated robots (Butler et al., 2015; Krupp et al., 2017). Butler et al. (2015) explored how, by means of visual filters, the privacy concerns of individuals can be reduced, and the benefits of teleoperated robots can still be reaped. Krupp et al. (2017) used focus groups to identify salient privacy concerns about telepresence robot. They found that informational concerns were most strongly discussed (106 occurrences of the theme in coding). However, physical concerns also received high attention with 60 occurrences. Social and psychological privacy, by contrast, received far less attention (both 16 occurrences). In addition, the study found important emerging categories that were sometimes understood in privacy terms, for example, marketing and theft. Other studies on home telepresence robots have studied how the framing of human–robot interaction and presentation of robot actions within a home by means of short video excerpts affects individual's privacy responses toward the robot (Rueben et al., 2017b). Rueben et al. (2017b) demonstrate the impact of what the authors call "contextual frames" on individuals' privacy judgments.

Furthermore, we see a growing, interdisciplinary interest in research about the privacy implications of social robots, with an uptick in publications across disciplines since 2015 (Lutz et al., 2019). To bridge the disciplinary gaps, expert workshop insights on currently under-addressed topics have led to the identification of interdisciplinary research needs (Rueben et al., 2018; Fosch-Villaronga et al., 2020; Kapeller et al., 2020) and have stipulated the emergence of new research fields, such as the field of privacy-sensitive robotics (Rueben et al., 2018). These workshops with experts across disciplines provide qualitative insights into the ethical, social, and legal implications of social (Rueben et al., 2018; Fosch-Villaronga et al., 2020) and wearable robots (Kapeller et al., 2020), pointing to the privacy-relevant issues to be tackled in the future. Privacy-related aspects include data privacy and transparency, deception and manipulation, agency and control, accountability, as well as trust, and recommendations on how to address them have been developed, for example, increased control and transparency requirements and the prohibition of data collection in certain instances.

Larger-scale quantitative studies, such as general population survey assessing citizens' attitudes and concerns toward robots, exist as well (e.g., Eurobarometer, 2012, 2015; Madden and Rainie, 2015). In the European Union (EU), the general attitudes toward robots are positive (64%) even though many fear that robots will take away jobs and alter the current labor market (70%). Interestingly, citizens in the EU did express some uneasiness with the idea of robotic companionship for elderly and surgical robots; yet the Eurobarometer (2015) did not link these feelings/responses to potential privacy concerns.

Overall, though, there seems to be a lack of empirical studies that assess the privacy concerns of social robots, especially with a quantitative approach (Lutz et al., 2019). Empirical research would prove effective to better understand the validity of theoretical knowledge on privacy. Moreover, empirical research can add a non-expert view on commonly theorized issues and, thus, take into account a more thorough perspective, potentially helping to shape responsible adoption in the future.

Our current study builds on earlier research that used a survey to test the privacy paradox among non-experts (Lutz and Tamò-Larrieux, 2020). This study found evidence for a robot privacy paradox, where users revealed privacy concerns (different levels, depending on the privacy type), but these concerns were not significantly correlated to robot use intentions, even after controlling for salient control variables such as expected benefits, social influence, scientific knowledge, and trust. Following up on this work, we aimed at a test that allowed to identify the role of privacy concerns less generally and more specifically. Thus, in contrast to the aforementioned study, our work here asks for privacy concerns about a fictional but concrete social robot, rather than social robots more broadly. The chosen method of an experimental vignette survey thus provides a more realistic

test of the relationship between privacy concerns and robot use intentions.

Privacy and Trust

The intricacies between privacy and trust is a complex phenomenon (Richards and Hartzog, 2016; Waldman, 2018). The abovementioned control and boundary-management functions of privacy enable interpersonal relationships that are built upon trust and trusting beliefs (Westin, 1967). At the same time, from an institutional perspective, companies including manufacturers of social robots might be incentivized to promote consumer trust by means of enhanced privacy features, linking privacy and trust via an economic element (Hartzog, 2018; Tamò-Larrieux, 2018). The importance of privacy for trust has also been acknowledged in more recent policy papers and ethical guidelines (European Commission, 2018, 2020; Delcker, 2019). While these papers and guidelines focus on artificial intelligence (AI) and ways to promote trustworthy AI, many operations of social robots already today employ such technology. These strategic objectives for AI will thus influence the development of social robots.

The relationship between trust and automation is complex, and literature on the subject has emerged (Lee and See, 2004; Cheshire, 2011; Hoffman et al., 2013; Schaefer et al., 2016; Botsman, 2018). While the relational perspective on trust among humans often defines trust as "a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behaviors of another" (Rousseau et al., 1998, p. 395), Botsman (2018) understands trust as a "confident relationship with the unknown" (p. 20). Similarly, Möllering (2001) identifies trust as a three-step mental process of expectation, interpretation, and suspension. Interaction with a social robot requires trust because private information is disclosed to the machine. Such a disclosure requires a favorable expectation of an outcome that is uncertain. Whether or not an individual interprets an outcome to be favorable relies on a mix of different elements, including rational and emotional ones, and finally, the individual must take what Möllering (2001, p. 414) calls a "leap of trust," meaning that what an individual interpreted becomes accepted and the unknownable momentarily certain.

The literature on interpersonal trust can help to understand trusting beliefs among humans and social robots (or more broadly: automation). However, "trust in automation involves other factors that relate specifically to technology's limitations and foibles" such as its "reliability, validity, utility, robustness, and false-alarm rate" (Hoffman et al., 2013, p. 85). The capabilities of technologies (such as social robots) as well as their affordances (e.g., ability to control certain features, communication with a device) impact trust in automation overall (Schaefer et al., 2016). What complicates the relationship further is that these technical features, which are continuously evolving with technological progress, are only one aspect in the broader calculus that impacts trust in the robot: Human factors (e.g., their personality, trust propensity, attitudes, etc.) and environmental ones (e.g., ways in which technologies are used within an environment, cultural notions) contribute to the full picture (Lee and See, 2004; Schaefer et al., 2016). Moreover, neither trust nor automation is static, but constantly evolving with new experiences. Capturing trust and trusting believes can thus only be done via proxy and illustrates a specific point in time and interaction with one technology.

With respect to trust and privacy in the context of social robots, an interesting field has emerged that analyzed situations in which users trust (Aroyo et al., 2018; Kundinger et al., 2019) and overtrust social robots (Booth et al., 2017; Borenstein et al., 2018; Wagner et al., 2018). Overtrust has an implication for privacy, as it indicates a tendency to allow physical, social, and informational constraints to be crossed. Overtrust is defined as "a situation in which a person misunderstands the risk associated with an action because the person either underestimates the loss associated with a trust violation; underestimates the chance the robot will make such a mistake; or both" (Wagner et al., 2018, p. 22). Thus, the topic of overtrust is closely linked to the one of deception, which—as mentioned above—multiple expert workshops have pointed to with a call for future, interdisciplinary research in the field. These aspects tie back the discussion on a political level, where the promotion and development of trustworthy technology is at the forefront of European policymakers' agenda (European Commission, 2018, 2020).

MODEL AND THEORETICAL DEVELOPMENT

Behavioral intentions to use a social robot is the key dependent construct in this study. We used behavioral intentions, rather than actual or reported behavior, because of the topic of the study and nature of the data collection. We expected that few respondents had themselves interacted with a social robot. Thus, behavioral assessments would be unevenly distributed, less reliable, and less appropriate for the statistical analysis. Naturally, the reliance on behavioral intentions as the dependent construct, rather than actual behavior, prevents the test of the privacy paradox in the narrow sense. The privacy paradox is generally understood as the divergence between attitudes and behavior when it comes to privacy (Dienlin and Trepte, 2015; Kokolakis, 2017). It has evoked substantial research interest, especially in the context of digital and social media, as shown in the meta-analysis by Baruh et al. (2017), which included 166 studies with more than 75,000 participants across 34 countries in total. Their meta-analysis also discussed the distinction between behavioral intentions and actual behavior as the outcome variable in research on the privacy paradox. A considerable number of studies (about half of all the effects) in the meta-analysis relied on behavioral intentions, rather than actual behavior, as the outcome variable and the authors found similar effects for intentions and behavior: "[T]here were no significant differences between studies investigating behavioral intentions vs. behavior regarding use of online services [. . .] and use of SNSs" (p. 39). The same was true for privacy protection intentions vs. privacy protection behavior as the outcome. Only for sharing information as the dependent variable, behavior and intentions behaved differently, with the effects for intentions being stronger than for behavior. Thus, based on the meta-analysis, two conclusions can be drawn.

First, intentions is a frequently used dependent construct in privacy paradox, although it does not align with the commonly accepted definition. Second, intentions and behavior are similarly affected by privacy concerns. Taken together, and in conjunction with the empirical constraints of studying behavior in the context of social robots through a survey-based study, we deem it justifiable to look at intentions, rather than behavior.

In our research model, there are several attitudinal constructs to predict behavioral intentions: trust, privacy concerns, perceived benefits of robots, and scientific interest. We will discuss each of these factors in turn.

Trust, and specifically trusting beliefs, should be associated with robot use intentions. Trusting beliefs can be differentiated into specific dimensions such as integrity, benevolence, and ability (McKnight et al., 2002). Thus, the trustor must assess the trustee as honest, benevolent, and competent in order to form trust. If this is the case, individuals are more likely to develop trusting intentions, which will, in turn, lead to trusting behavior, for example, the use of a new technology. Trusting beliefs, rather than, for example, Möllering's (2001) leap of trust approach, are used because they are easier to measure. Based on the trust literature and trust theory, we propose the following hypothesis.

H1: *Trusting beliefs in a robot have a positive effect on social robot use intentions.*

Citizens need to overcome certain concerns to start using social robots voluntarily, privacy concerns being an important type. If the privacy risks of a social robot are thought to be high, we expect lower levels of adoption intention. However, ample research on self-disclosure and privacy in online contexts has shown that privacy attitudes—including concerns—often do not match privacy behavior (Kokolakis, 2017). Although they are concerned about their privacy, many users of digital services disclose sensitive information and do not protect their privacy adequately, for example, by choosing restrictive privacy settings. This divergence between attitudes and behavior is captured by the privacy paradox (Barnes, 2006). As discussed above, empirical studies that look at intentions, rather than behavior, are often also framed within a privacy paradox framework. In a way, we can interpret this understanding as a widened and broad take on the privacy paradox. To date, the empirical evidence on the privacy paradox—both in a strict and broad sense—is mixed. Many studies have identified a privacy paradox, but a considerable number of studies, especially newer ones, found significant effects between privacy attitudes and behavior or intentions, thus rejecting the privacy paradox. Kokolakis (2017) provides a systematic review of this literature, showing how the empirical evidence is inconclusive. Baruh et al. (2017) noted the absence of the paradox (i.e., there are small but significant effects between privacy attitudes and privacy-related behavior or intentions). However, their study also suggested that contexts matters because for social network sites, the privacy paradox seems to hold. In light of the emerging nature of social robots and low adoption rates, we expect that privacy concerns have a significant and negative effect on robot use intentions.

H2: *Privacy concerns about a robot have a negative effect on social robot use intentions.*

In the literature on the privacy paradox, different theoretical explanations for the paradox can be differentiated (Hoffmann et al., 2016). However, the privacy calculus has emerged as the dominant explanation (Dinev and Hart, 2006). Within this approach, users weigh the benefits and risks of a technology against each other and if the former outweigh the later, they will start or keep using the technology. A rich literature exists that analyzes the perceived risks and benefits of social network sites and elaborates on the privacy calculus in this context (Dienlin and Metzger, 2016; Trepte et al., 2017). This literature highlights the influence of cultural norms on privacy calculations (Krasnova et al., 2012; Trepte et al., 2017), showing how the privacy calculus is not a purely rational process but heavily influenced by cultural and psychological default positions. Moreover, the framing of privacy concerns and sharing benefits will likely affect use intentions. Coopamootoo and Groß (2017) found that privacy attitudes and sharing attitudes differed significantly in terms of emotional connotation. Privacy attitudes related to fear, bringing up actors with a negative connotation such as hackers and data collectors (e.g., Google). By contrast, sharing attitudes related to joy, bringing up actors with a positive connotation such as family and friends. Thus, whether individuals are in a privacy mindset or a sharing mindset might result in different behaviors. Applied to social robots, depending on the framing of the discussion and if this technology is seen as very useful and benefitting their personal lives (i.e., sharing attitudes are prioritized over privacy attitudes), individuals will have higher use intentions. On the other hand, if a social robot is framed more in privacy terms, individuals will have lower use intentions. Given theories such as the theory of planned behavior (TPB; Ajzen, 1991) and previous research (Alaiad and Zhou, 2014), we expect that perceived benefits exert a positive influence on robot use intentions.

H3: *Perceived benefits of social robots have a positive effect on social robot use intentions.*

In TPB, social influence is an important antecedent of behavioral intention (McEachan et al., 2011). Likewise, technology adoption approaches, for example, the technology acceptance model and the unified theory of acceptance and use of technology (UTAUT) highlight the key role of social factors people's adoption decisions (Venkatesh and Morris, 2000; Venkatesh et al., 2003). In these theories, social influence increases behavioral intentions to adopt a new technology. As a not yet widely adopted technology, social robots should drive use intentions when someone's social environment encourages or expects their use. Citizens that have more social robot-friendly networks should therefore have higher intentions to use them.

H4: *Social influence has a positive effect on social robot use intentions.*

Scientific interest was included as a control variable. Citizens who are more scientifically interested tend to be more up-to-date with recent technological developments, including those that pertain to social robots. Since social robots are still not widely adopted, we consider scientific interest as a proxy for knowledge and awareness of social robots—and technology skills with social robots. Extant research has shown that (digital) technology skills vary by education (Van Deursen and Van Dijk, 2011; De Boer et al., 2020). Based on De Boer et al. (2020) study about internet-of-things technologies, which share similarities with social robots, we expect technology skills with social robots

to vary by education level as well. Given that scientific interest and technology skills are both shaped by someone's education, we think it is justifiable to use scientific interest as a proxy for technology skills with social robots, particularly in a situation where individuals do not have experience with the technology itself (i.e., they do not own a robot). Scientifically interested citizens should be able to assess the benefits and risks of the technology more closely, including the privacy risks. They might also be more technologically open minded and curious. Using diffusion of innovation theory as a conceptual basis (Rogers, 2003), citizens interested in scientific development should have higher behavioral intentions to use novel technologies, including social robots. By including scientific interest, we also follow existing survey-based studies (Eurobarometer, 2012).

H5: *Scientific interest has a positive effect on social robot use intentions.*

METHODS

Sample

The experimental vignette study was conducted in December 2018, in the form of a randomized survey with two conditions: high privacy risks and low privacy risks. We programmed the survey in Qualtrics and relied on MTurk for the participant recruitment, surveying respondents located in the United States (see Aguinis et al., 2020 for more information on MTurk as a data source). The average completion time was 8 min and participants were compensated with 1.25 US dollars, leading to an average hourly wage of 9.5 US dollars. We aimed for a sample of 300 participants–150 per condition—and in the end, 298 respondents completed the study. Because they failed at least one of two attention checks, six individuals were eliminated from further analysis, leaving us with a final sample of 292. The average age in the final sample is 35 years old (median = 33 years; SD = 9.5 years). One hundred fourteen respondents identify as female (39%), 177 as male (60.5%), and one person prefers not to say (0.5%). The sample is relatively educated, with 16% having high school as their highest degree, 36% some college, 41% a Bachelor, 6.5% a Master, and 0.5% a Doctorate or Other.

Measurement and Data Analysis

To test our hypotheses, we used a people paper study (Aguinis and Bradley, 2014) with a between-subjects design and with a manipulation of privacy risks into a high and low condition. Participants were randomly assigned into either the high or low privacy risk scenario. The vignette described a fictional social robot called MIMO. MIMO is portrayed as an affordable companion robot that offers useful functionality. In both conditions, the respondents saw the same introductory paragraph describing MIMO's general functionality and purpose. However, the next two paragraphs differed between the two conditions. In the first and low privacy risk scenario, MIMO is e-privacy certified and fully complies with current US and European privacy laws. MIMO tends to have privacy-by-default settings in this scenario and fewer privacy-invasive capabilities than in the high privacy risk scenario. Moreover, the data MIMO collects is stored more securely and locally. By contrast, MIMO in the high privacy risk scenario is not e-privacy certified and does not comply with US and European privacy law (compliance with privacy laws is not a condition for market entrance but suppliers that violate privacy laws risk facing steep fines, Newlands et al., 2020). In this scenario, MIMO is more privacy-invasive, for example, by being switched on by default and performing additional analyses on the users' voice. Moreover, MIMO has worse security in this scenario. The two privacy risk scenarios are shown in **Supplementary Figures 1** and **2** in **Supplementary Material**. We focused strongly on informational privacy for these scenarios but included elements of other privacy types as well. For the formulation of the scenarios, we followed established privacy conceptualizations and measurements (e.g., Malhotra et al., 2004; Stutzman et al., 2011), intending that the low-risk scenario would result in lower privacy concern scores on these scales and that the high privacy risk scenario would result in higher scores. As a manipulation check, participants responded to 16 privacy concern questions/items that can be grouped into four privacy concern types [Lutz and Tamò-Larrieux (2020) for more information on these four dimensions]. The manipulation checks indicated that the conditions clearly differentiated privacy concerns (**Table 1**).

The privacy risk manipulations were entered into a regression as dummy variables (0—low privacy risk, 1—high privacy risk), and we used robot use intentions as the dependent variable. Principal component analysis was used to bundle all constructs with more than one item (i.e., robot use intentions, overall privacy concerns, trusting beliefs, social influence). All four constructs loaded neatly on one component and had high internal consistency. Cronbach's α was 0.96 for robot use intentions, 0.94 for overall privacy concerns, 0.92 for trusting beliefs, and 0.91 for social influence. No significant demographic differences in age ($t = 0.76$, $p = 0.45$), education ($t = 0.11$, $p = 0.92$), and gender (Chi-Square = 2.15, $p = 0.34$) exist between the respondents in the low and high privacy risk scenarios.

We used the measures from Lutz and Tamò-Larrieux (2020) to assess social robot use intention, social influence, trust, and scientific interest but slightly adjusted them to make the connection to the vignettes and MIMO. More specifically, the prompts at times reminded the respondents to think of the social robot described in the scenario and the items referred to this specific social robot rather than robots in general (e.g., for trusting beliefs, two sample items were *"I believe that such a robot acts in my best interest."* and *"Overall, such a robot is a capable and proficient service provider."*). For perceived benefits, we used a more succinct measurement with only one item based on the Eurobarometer (2012) survey. The item assessed respondents' opinion about social robots in general terms and had four response options: very negative, fairly negative, fairly positive, and very positive. The full questionnaire used is shown in **Supplementary Material**.

To test the hypotheses, we conducted a linear regression analysis in Stata (v.15), using the "robust" option for heteroscedasticity-corrected standard errors. We also tested for colinearity, and the largest variance inflation factors were 2.17 and 2.14 for the education categories "some college" and "Bachelor's degree," respectively, indicating the absence of severe

TABLE 1 | Manipulation check.

	Low privacy risk condition	High privacy risk condition	*t*-value	Sig.	Mean difference [confidence interval]
Physical privacy concerns	2.02	2.61	5.01	0.00	0.59 [0.36, 0.83]
Institutional informational privacy concerns	3.14	4.22	8.15	0.00	1.08 [0.82, 1.34]
Social informational privacy concerns	2.66	3.89	9.18	0.00	1.23 [0.97, 1.49]
Overall privacy concerns	2.31	3.67	9.82	0.00	1.36 [1.09, 1.63]

Arithmetic means are displayed for columns 2 and 3; 1–5 scales; N = 143 for low(er) privacy risk scenario and 149 for high(er) privacy risk scenario; Levene's test for equality of variances yields p-values > 0.05 for social, physical, and global privacy concerns, indicating equal variances assumed, but < 0.05 for institutional privacy concerns; measurement of privacy concerns dimensions based on Lutz and Tamò (2015).

TABLE 2 | Regression of robot use intentions on demographics, privacy, trusting beliefs, general opinion/beliefs, social influence, and scientific interest.

	Unstandardized coefficient (robust standard errors)	Beta
Age	0.01 (0.01)	0.04
Gender (reference: female)		
Male	−0.03 (0.10)	−0.01
Other	−0.77*** (0.13)	−0.04
Education (reference: high School)		
Some college	0.2 (0.14)	0.07
Bachelor	0.19 (0.14)	0.07
Master	0.43* (0.21)	0.08*
Doctor	0.78*** (0.25)	0.04***
Other	−0.02 (0.16)	0.00
Privacy risk condition (reference: low risk)	−0.65*** (0.11)	−0.25***
Trusting beliefs	0.29*** (0.07)	0.22***
General opinion/benefits	0.22** (0.08)	0.12**
Social influence	0.54*** (0.06)	0.47***
Scientific interest	0.07 (0.10)	0.03
Constant	−0.47 (0.42)	

*N = 292; R^2 = 0.62; ***p < 0.001, **p < 0.01, *p < 0.05, no star, not statistically significant. A Bonferroni correction that assumes a p-value threshold of 0.05 would result in a corrected statistical significance threshold of 0.00625 (0.05/8), since there are eight predictor variables, five from the hypotheses, and three control variables. Education: Master is the only effect that becomes insignificant after such a correction. All other significant predictors have p-values below 0.00625.*

colinearity. Privacy concerns were entered as an independent dummy variable based on the condition (high privacy risk scenario vs. low privacy risk scenario).

RESULTS

Table 2 displays the results of the linear regression analysis. Trusting beliefs in the robot have a significant and positive effect on robot use intentions. The null hypothesis is therefore rejected, offering some support for H1. Privacy concerns have a significant and negative effect on social robot use intention, rejecting the null hypothesis and offering some support for H2. Controlling for demographic characteristics, trust, benefits/general opinion toward social robots, social influence, and scientific interest, respondents in the high-risk scenario score two thirds of a point lower (on a five-point scale) in their intention to adopt the social robot. Perceived benefits, in the form of general opinions about robots, have a significant and positive effect on robot use intentions, refuting the null hypothesis and offering some support for H3. Social influence affects robot use intentions significantly and positively. Thus, the null hypothesis is rejected for H4, and some support is found for this hypothesis. The more supportive someone's social environment toward robots, the higher the intentions to use such a robot. Finally, H5 is rejected as scientific interest does not influence robot use intentions significantly. The demographic predictors exert a limited influence on robot use intentions, but more educated users have somewhat higher use intentions.

Overall, four out of the five hypotheses tend to be supported, and one is rejected. Importantly, the main hypothesis about the privacy paradox (H2) was not rejected. We are able to explain 62 percent of the variance in intention to use the fictional social robot with our independent variables.

DISCUSSION AND CONCLUSION

Previous research on privacy in digital contexts has detected a privacy paradox between users' privacy attitudes and their behaviors as well as between privacy attitudes and intentions. Users report high levels of privacy concerns but exhibit behavior that could be interpreted divergently, such as high levels of disclosure of personal information and low levels of privacy protection (Chen and Rea, 2004; Milne et al., 2009). While the individualistic notion of the privacy paradox and privacy self-management is increasingly contested (e.g., Obar, 2015; Lutz et al., 2020), we, nevertheless, took the privacy paradox as a useful starting for investigating social robots as an emerging but not yet widely adopted technology. Following up on earlier work, where we had tested the privacy paradox for social robots more generally and indeed found evidence for a privacy paradox (Lutz and Tamò-Larrieux, 2020), we wanted to check whether the privacy paradox between privacy concerns and intentions holds when confronted with a concrete robot. We, thus, opted for an experimental vignette study as a middle ground between a lab study, which is costly and time intensive, and a more generic survey. The experimental vignette study allows testing the relationship between privacy concerns and robot use intentions in a causal sense and is more tangible than a general survey.

We found that the privacy manipulation had a relatively pronounced effect on robot use intentions. Respondents in the more invasive scenario were significantly less likely to be willing to use such a robot, controlling for a range of predictors. Trusting beliefs, social influence, and general opinion about robots also influenced robot use intentions significantly—and positively.

Several theoretical and practical *implications* come with our findings. Importantly, we did not find a privacy paradox and instead identified a strong role of privacy invasiveness in affecting use intentions. This is in line with overview articles that looked at the privacy paradox more generally. For example, Baruh et al. (2017), in their meta-analysis of research on the privacy paradox, identified that, on aggregate, the privacy paradox does not hold, and there is, in fact, an association between privacy concerns and privacy-related behavior as well as intentions. Similarly, Kokolakis (2017), in a systematic literature view on the privacy paradox, discussed a temporal trajectory with newer studies increasingly refuting the privacy paradox. Our research indicates that when individuals are confronted with concrete privacy-invasive technologies, they do take privacy into consideration. Thus, privacy matters—and will matter—for social robots (Rueben et al., 2018). However, a limitation of our study is that we used intentions rather than actual behavior as the dependent variable, due to the practical constraints of recruiting social robot owners with a general survey and constraints in doing a lab study. As discussed, the focus on intentions aligns with other research on the privacy paradox (see Baruh et al., 2017) but does not follow the original conceptualization of the privacy paradox as a divergence between attitudes and behavior. Thus, future research could confront individuals with actual robots that vary in privacy friendliness and test whether individuals use them differently in a controlled setting (of course making sure that no ethical boundaries are crossed and that users' privacy is not actually violated within the study). Research could also investigate the privacy paradox for adjacent technologies such as smart speakers and smart toys, which share similarities with social robots but are more widely adopted and therefore easier to sample for (Peter et al., 2019; Lutz and Newlands, 2021). Moreover, the privacy aspects were quite prominent in our vignettes. When deciding about purchasing a social robot in real life, potential users will probably not have the same concise privacy information available as in the study. With this in mind—and taking the literature on the privacy calculus and cultural differences into account (Krasnova et al., 2012; Trepte et al., 2017) as well as the one highlighting the limitations of rational decisions with respect to privacy (e.g., Acquisti and Grossklags, 2005)—it remains to be seen how privacy-friendly design of social robots impacts the willingness of users to buy and engage with them. Overall, as indicated by the literature on cultural privacy preferences, we assume that our results with respect to the use intention provided with substantial information on the privacy risks of devices would change depending on the dominant culture of a test group.

Another important finding is that trusting beliefs affect social robot use intention positively. Thus, individuals take the trustworthiness of a social robot into consideration when considering using it. We have discussed in the literature review how this can have ambivalent consequences, especially if users trust social robots too much (Booth et al., 2017; Borenstein et al., 2018; Wagner et al., 2018). Future research should explore the dynamics of trust and overtrust in social robots, and their connection to privacy. Such research is needed as overtrusting social robots might have serious privacy implications as overtrust leads to a tendency to allow physical, social, and informational boundaries to be crossed. Interdisciplinary research in this field should furthermore examine how deception by social robots influences privacy perceptions, use intentions, and trusting beliefs. Findings in those areas would further promote the policy objectives of the European Union, which aims at developing trustworthy technology (European Commission, 2018, 2020).

The support for H3 about a positive influence of perceived benefits/general opinion about robots points to the partly utilitarian nature of the technology. More positive opinions about robots will translate into higher use intentions. Future research could disentangle these opinions somewhat and investigate how positive or negative opinions are formed based on different factors such as media consumption, education, and technology attitudes more broadly. A limitation of our study was the single-item measurement of this construct. Future research should use more robust, multi-item scales to assess perceived benefits and general opinions about robots. Uses and gratifications would be a helpful theory to systematically develop perceived benefits (De Jong et al., 2019).

The finding that social influence has a positive effect on use intentions suggests that the use of social robots, as an emerging technology, depends heavily on someone's social environment. Thus, social robots have to be understood in context and their situatedness within certain social milieus (e.g., educated and tech-affine people) begs for further study, especially through observational and qualitative approaches. Our findings show that social norms are of crucial importance in the context of social robots. Robotics firms might want to invest in community management and word-of-mouth promotion to leverage this social influence.

Overall, our study suggests that privacy matters. Robotics firms should therefore take privacy sensitivity into consideration as an important design factor. If privacy is neglected and privacy invasions occur, the media are quick to highlight these issues, as it happened when privacy norm violations with related technologies, such as smart speakers, occurred (Estes, 2018; Day et al., 2019a,b). Thus, robotics firms should construe privacy as a key part of their development philosophy and not as an afterthought. In Europe, this is legally mandated by the privacy-by-design and privacy-by-default principle established within the General Data Protection Regulation (GDPR). How the principle of privacy-by-design and privacy-by-default will impact the concrete design of social robots is still to be seen.

Robotics firms should be aware of the fact that consumers value privacy and consider it in their purchasing decisions when faced with tangible risks. In that sense, manufacturers might want to increase investments into privacy-sensitive robotics (Rueben et al., 2018). Not only should manufacturers develop social robots that are privacy friendly, but they should also communicate their privacy-protection efforts to potential customers in concise and

transparent ways (Felzmann et al., 2019). Here too, the GDPR paves the way in Europe with a list of specific information duties that data controllers (i.e., entities determining what data are being processed for what purpose) must provide to the data subjects (i.e., the person affected by a data processing of a social robot and to whom the personal data being processed belongs to).

Aside from government strategy positions (e.g., European Commission, 2018, 2020), newer industry standards on "trustworthiness in artificial intelligence" (ISO/IEC TR 24028:2020) elaborate on approaches toward security and privacy in AI. Such self-regulatory standards show that also the industry has realized the need for a holistic and standardized manner to ensure trust in AI as well as AI-based products (e.g., social robots). It will be interesting to follow how the strategy positions of governments will shape the approaches of the industry through standardization efforts as well as upcoming legislation.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors upon request, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

CL and AT-L were jointly responsible for coming up with the paper idea and the research design, carrying out the data collection, and writing the Introduction and passages within Discussion and Conclusion. CL was mostly responsible for the data analysis and writing of the sections Model and Theoretical Development, Methods, and Results. AT-L was mostly responsible for writing the Literature Review section and parts of the Discussion and Conclusion section. Both authors contributed to the article and approved the submitted version.

FUNDING

CL received funding from the Research Council of Norway under Grant Agreement 275347 Future Ways of Working in the Digital Economy. AT-L received funding from the Digital Society Initiative fellowship (University of Zurich, Switzerland) and the International Postdoctoral Fellowship Grant (University of St. Gallen, Switzerland; Project Number 1031564).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frobt.2021.627958/full#supplementary-material

Supplementary Figure 1 | Experimental vignette for low privacy risk scenario (privacy-friendly robot).

Supplementary Figure 2 | Experimental vignette for high privacy risk scenario (privacy-unfriendly robot).

REFERENCES

Acharya, A. (2015). *Are we ready for driver-less vehicles? Security vs. privacy – A social perspective*. Available online at: https://arxiv.org/abs/1412.5207

Acquisti, A., and Grossklags, J. (2005). Privacy and rationality in individual decision making. *IEEE Secur. Priv.* 3, 26–33. doi: 10.1109/MSP.2005.22

Aguinis, H., and Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organ. Res. Methods* 17, 351–371. doi: 10.1177/1094428114547952

Aguinis, H., Villamor, I., and Ramani, R. S. (2020). MTurk research: review and recommendations. *J. Manage.* 47, 823–837. doi: 10.1177/0149206320969787

Ajzen, I. (1991). The theory of planned behavior. *Organ. Behav. Hum. Decis. Process.* 50, 179–211. doi: 10.1016/0749-5978(91)90020-T

Alaiad, A., and Zhou, L. (2014). The determinants of home healthcare robots adoption: an empirical investigation. *Int. J. Med. Inform.* 83, 825–840. doi: 10.1016/j.ijmedinf.2014.07.003

Altman, I. (1975). *The Environment and Social Behavior: Privacy, Personal Space, Territory, and Crowding*. Monterey, CA: Wadsworth Publishing Company.

Aroyo, A. M., Rea, F., Sandini, G., and Sciutti, A. (2018). Trust and social engineering in human robot interaction: will a robot make you disclose sensitive information, conform to its recommendations or gamble? *IEEE Robot. Automat. Lett.* 3, 3701–3708. doi: 10.1109/LRA.2018.2856272

Barnes, S. B. (2006). A privacy paradox: social networking in the United States. *First* Monday 11:1394. doi: 10.5210/fm.v11i9.1394

Baruh, L., Secinti, E., and Cemalcilar, Z. (2017). Online privacy concerns and privacy management: a meta-analytical review. *J. Commun.* 67, 26–53. doi: 10.1111/jcom.12276

Baxter, P., Kennedy, J., Senft, E., Lemaignan, S., and Belpaeme, T. (2016). "From characterising three years of HRI to methodology and reporting recommendations," in *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (Christchurch), 391–398. doi: 10.1109/HRI.2016.7451777

Booth, S., Tompkin, J., Pfister, H., Waldo, J., Gajos, K., and Nagpal, R. (2017). "Piggybacking robots: Human-robot overtrust in university dormitory security," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY: ACM), 426–434.

Borenstein, J., Wagner, A. R., and Howard, A. (2018). Overtrust of pediatric health-care robots: a preliminary survey of parent perspectives. *IEEE Robot. Automat. Magazine* 25, 46–54. doi: 10.1109/MRA.2017.2778743

Botsman, R. (2018). Who Can you *T*rust? *How Technology Brought us Together - and Why it Could Drive us Apart*. London: Penguin Books.

Butler, D. J., Huang, J., Roesner, F., and Cakmak, M. (2015). "The privacy-utility tradeoff for remotely teleoperated robots," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (Portland OR: ACM), 27–34. doi: 10.1145/2696454.2696484

Bygrave, L. A. (2002). Data *P*rotection *L*aw: Approaching its *Rationale, Logic and Limits*. Alphen aan den Rijn: Wolters Kluwer.

Calo, R. (2012). "Robots and privacy," in *Robot Ethics: The Ethical and Social Implications of Robotics,* eds P. Lin, G. Bekey, and K. Abney (Cambridge: MIT Press), 187–202.

Carnevale, A. (2016). Will robots know us better than we know ourselves? *Rob. Auton. Syst.* 86, 144–151. doi: 10.1016/j.robot.2016.08.027

Chen, K., and Rea, A. I. (2004). Protecting personal information online: a survey of user privacy concerns and control techniques. *J. Comput. Inform. Syst.* 44, 85–92. doi: 10.1080/08874417.2004.11647599

Cheshire, C. (2011). Online trust, trustworthiness, or assurance? *Daedalus* 140, 49–58. doi: 10.1162/DAED_a_00114

Coopamootoo, K. P., and Groß, T. (2017). Why privacy is all but forgotten. *Proc. Privacy Enhanc. Technol.* 4, 97–118.

Darling, K. (2016). "Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects," in *Robot Law*, eds R. Calo, M. Froomkin, and I. Kerr. (Cheltenham: Edward Elgar), 213–234. doi: 10.4337/9781783476732.00017

Day, M., Turner, G., and Drozdiak, N. (2019a, April 11). Amazon workers are listening to what you tell Alexa. *Bloomberg*. Available online at: https://www.bloomberg.com/news/articles/2019-04-10/is-anyone-listening-to-you-on-alexa-a-global-team-reviews-audio

Day, M., Turner, G., and Drozdiak, N. (2019b, April 24). Amazon's Alexa team can access users' home addresses. *Bloomberg*. Available online at: https://www.bloomberg.com/news/articles/2019-04-24/amazon-s-alexa-reviewers-can-access-customers-home-addresses

De Boer, P. S., van Deursen, A. J., and van Rompay, T. J. (2020). Internet-of-things skills among the general population: task-based performance test using activity trackers. *JMIR Human Factors* 7:e22532. doi: 10.2196/22532

De Jong, C., Kühne, R., Peter, J., Van Straten, C. L., and Barco, A. (2019). "What do children want from a social robot? Toward gratifications measures for child-robot interaction," in *2019 28th IEEE International Conference on Robot and Human Interactive Communication* (RO-MAN) (New Delhi: IEEE), 1–8. doi: 10.1109/RO-MAN46459.2019.8956319

Delcker, J. (2019, May 19). US to endorse new OECD principles on artificial intelligence. *Politico*. Available online at: https://www.politico.eu/article/u-s-to-endorse-new-oecd-principles-on-artificial-intelligence/

Dienlin, T., and Metzger, M. J. (2016). An extended privacy calculus model for SNSs—Analyzing self-disclosure and self-withdrawal in a U.S. representative sample. *J. Comput.-Mediat. Commun.* 21, 368–383. doi: 10.1111/jcc4.12163

Dienlin, T., and Trepte, S. (2015). Is the privacy paradox a relic of the past? An in-depth analysis of privacy attitudes and privacy behaviors. *Eur. J. Soc. Psychol.* 45, 285–297. doi: 10.1002/ejsp.2049

Dinev, T., and Hart, P. (2006). An extended privacy calculus model for e-commerce transactions. *Inform. Syst. Res.* 17, 61–80. doi: 10.1287/isre.1060.0080

Estes, A. C. (2018, May 25). Your worst Alexa nightmares are coming true. *Gizmodo*. Available online at: https://gizmodo.com/your-worst-alexa-nightmares-are-coming-true-1826327301

Eurobarometer (2012). *Special Eurobarometer 382: Public attitudes towards robots*. Available online at: http://ec.europa.eu/public_opinion/archives/ebs/ebs_382_en.pdf

Eurobarometer (2015). *Special Eurobarometer 427: Autonomous systems*. Available online at: http://ec.europa.eu/public_opinion/archives/ebs/ebs_427_en.pdf

European Commission (2018). *Ethics guidelines for trustworthy AI: High-Level Expert Group on Artificial Intelligence*. Available online at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

European Commission (2020). *On Artificial Intelligence - A European approach to excellence and trust*. Available online at: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

Felzmann, H., Fosch-Villaronga, E., Lutz, C., and Tamò-Larrieux, A. (2019). Robots and transparency: the multiple dimensions of transparency in the context of robot technologies. *IEEE Robot. Automat. Magazine* 26, 71–78. doi: 10.1109/MRA.2019.2904644

Finn, R. L., Wright, D., and Friedewald, M. (2013). "Seven types of privacy," in *European Data Protection: Coming of Age*, eds S. Gutwirth, R. Leenes, P. De Hert, and Y. Poullet (New York: Springer), 3–32. doi: 10.1007/978-94-007-5170-5_1

Fong, T., Nourbakhsh, I., and Dautenhahn, K. (2003). A survey of socially interactive robots. *Rob. Auton. Syst.* 42, 143–166. doi: 10.1016/S0921-8890(02)00372-X

Fosch-Villaronga, E., Lutz, C., and Tamò-Larrieux, A. (2020). Gathering expert opinions for social robots' ethical, legal, and societal concerns: findings from four international workshops. *Int. J. Soc. Robot.* 12, 441–458. doi: 10.1007/s12369-019-00605-z

Greenleaf, G. (2014). Sheherezade and the 101 data privacy laws: origins, significance and global trajectories. *J. Law Inform. Sci.* 23, 4–49. doi: 10.2139/ssrn.2280877

Greenleaf, G., and Cottier, B. (2020). 2020 ends a decade of 62 new data privacy laws. *Priv. Laws Bus. Int. Report* 163, 24–26. Available online at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3572611

Gupta, S. K. (2015). Six recent trends in robotics and their implications. *IEEE Spectrum*. Available online at: https://spectrum.ieee.org/automaton/robotics/home-robots/six-recent-trends-in-robotics-and-their-implications

Hartzog, W. (2018). *Privacy's Blueprint: The Battle to Control the Design of New Technologies*. Cambridge, MA: Harvard University Press. doi: 10.4159/9780674985124

Hoffman, R. R., Johnson, M., Bradshaw, J. M., and Underbrink, A. (2013). Trust in automation. *IEEE Intell. Syst.* 28, 84–88. doi: 10.1109/MIS.2013.24

Hoffmann, C. P., Lutz, C., and Ranzini, G. (2016). Privacy cynicism: a new approach to the privacy paradox. *Cyberpsychology* 10:7. doi: 10.5817/CP2016-4-7

Kaminski, M. E. (2015). Robots in the home: what will we have agreed to? *Idaho Law Rev.* 51, 661–677. Available online at: https://www.uidaho.edu/-/media/UIdaho-Responsive/Files/law/law-review/articles/volume-51/51-3-kaminski-margot-e.pdf

Kaminski, M. E., Rueben, M., Grimm, C., and Smart, W. D. (2017). Averting robot eyes. *Maryland Law Rev.* 76, 983–1023. Available online at: https://ssrn.com/abstract=3002576

Kapeller, A., Felzmann, H., Fosch-Villaronga, E., and Hughes, A. M. (2020). A taxonomy of ethical, legal and social implications of wearable robots: an expert perspective. *Sci. Eng. Ethics* 26, 3229–3247. doi: 10.1007/s11948-020-00268-4

Kokolakis, S. (2017). Privacy attitudes and privacy behaviour: a review of current research on the privacy paradox phenomenon. *Comput. Secur.* 64, 122–134. doi: 10.1016/j.cose.2015.07.002

Koops, B. J., Newell, B. C., Timan, T., Škorvánek, I., Chokrevski, T., and Galič, M. (2016). A typology of privacy. *J. Int. Law* 38, 483–575. Available online at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2754043

Krasnova, H., Veltri, N. F., and Günther, O. (2012). Self-disclosure and privacy calculus on social networking sites: the role of culture. *Bus. Inform. Syst. Eng.* 4, 127–135. doi: 10.1007/s12599-012-0216-6

Krupp, M. M., Rueben, M., Grimm, C. M., and Smart, W. D. (2017). "Privacy and telepresence robotics: what do non-scientists think?," in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna: ACM), 175–176. doi: 10.1145/3029798.3038384

Kundinger, T., Wintersberger, P., and Riener, A. (2019). "(Over) Trust in automated driving: the sleeping pill of tomorrow?," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow: ACM), 1–6. doi: 10.1145/3290607.3312869

Lee, J. D., and See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Hum. Factors* 46, 50–80. doi: 10.1518/hfes.46.1.50.30392

Lee, M. K., Tang, K. P., Forlizzi, J., and Kiesler, S. (2011). "Understanding users' perception of privacy in human-robot interaction," in *Proceedings of the 6th International Conference on Human-Robot Interaction* (Lausanne: ACM), 181–182. doi: 10.1145/1957656.1957721

Liang, Y., and Lee, S. A. (2017). Fear of autonomous robots and artificial intelligence: evidence from national representative data with probability sampling. *Int. J. Soc. Robot.* 9, 379–384. doi: 10.1007/s12369-017-0401-3

Lutz, C., Hoffmann, C. P., and Ranzini, G. (2020). Data capitalism and the user: an exploration of privacy cynicism in Germany. *New Media Society* 22, 1168–1187. doi: 10.1177/1461444820912544

Lutz, C., and Newlands, G. (2021). Privacy and smart speakers: a multi-dimensional approach. *Inform. Soc.* 37:1897914. doi: 10.1080/01972243.2021

Lutz, C., Schöttler, M., and Hoffmann, C. P. (2019). The privacy implications of social robots: scoping review and expert interviews. *Mobile Media Commun.* 7, 412–434. doi: 10.1177/2050157919843961

Lutz, C., and Tamò, A. (2015). "RoboCode-Ethicists: privacy-friendly robots, an ethical responsibility of engineers?," in *Proceedings of the 2015 ACM Web Science Conference* (Oxford: ACM). doi: 10.1145/2793013.2793022

Lutz, C., and Tamò, A. (2018). "Communicating with robots: ANTalyzing the interaction between healthcare robots and humans with regards to privacy," in *Human-Machine Communication: Rethinking Communication, Technology, and Ourselves*, ed A. Guzman (Bern: Peter Lang), 145–165.

Lutz, C., and Tamò-Larrieux, A. (2020). The robot privacy paradox: understanding how privacy concerns shape intentions to use social robots. *Hum.-Mach. Commun.* 1, 87–111. doi: 10.30658/hmc.1.6

Madden, M., and Rainie, L. (2015). Americans' attitudes about privacy, security and surveillance. *Pew Internet, Science and Tech Report*. Available

online at: http://www.pewinternet.org/2015/05/20/americans-attitudes-about-privacy-security-and-surveillance

Malhotra, N. K., Kim, S. S., and Agarwal, J. (2004). Internet users' information privacy concerns (IUIPC): the construct, the scale, and a causal model. *Inform. Syst. Res.* 15, 336–355. doi: 10.1287/isre.1040.0032

McEachan, R. R. C., Conner, M., Taylor, N. J., and Lawton, R. J. (2011). Prospective prediction of health-related behaviours with the theory of planned behaviour: a meta-analysis. *Health Psychol. Rev.* 5, 97–144. doi: 10.1080/17437199.2010.521684

McKnight, D. H., Choudhury, V., and Kacmar, C. (2002). Developing and validating trust measures for e-commerce: an integrative typology. *Inform. Syst. Res.* 13, 334–359. doi: 10.1287/isre.13.3.334.81

Milne, G. R., Labrecque, L. I., and Cromer, C. (2009). Toward an understanding of the online consumer's risky behavior and protection practices. *J. Consum. Affairs* 43, 449–473. doi: 10.1111/j.1745-6606.2009.01148.x

Möllering, G. (2001). The nature of trust: from Georg Simmel to a theory of expectation, interpretation and suspension. *Sociology* 35, 403–420. doi: 10.1177/S0038038501000190

Newlands, G., Lutz, C., Tamò-Larrieux, A., Villaronga, E. F., Harasgama, R., and Scheitlin, G. (2020). Innovation under pressure: implications for data privacy during the Covid-19 pandemic. *Big Data Society* 7, 1–14. doi: 10.1177/2053951720976680

Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Rev.* 79, 101–139. Available online at: https://core.ac.uk/download/pdf/267979739.pdf

Obar, J. A. (2015). Big data and the phantom public: Walter Lippmann and the fallacy of data privacy self-management. *Big Data Society,* 2, 1–16. doi: 10.1177/2053951715608876

Pavlou, P. A. (2011). State of the information privacy literature: where are we now and where should we go? *MIS Q.* 35, 977–988. doi: 10.2307/41409969

Peter, J., Kühne, R., Barco, A., de Jong, C., and van Straten, C. L. (2019). "Asking today the crucial questions of tomorrow: social robots and the Internet of Toys," in *The Internet of Toys: Practices, Affordances and the Political Economy of Children's Smart Play,* eds G. Mascheroni and D. Holloway (Cham: Palgrave Macmillan), 25–46. doi: 10.1007/978-3-030-10898-4_2

Petronio, S. (2002). Boundaries of *Privacy: Dialectics of Disclosure*. New York, NY: State University of New York Press.

Raynes-Goldie, K. (2010). Aliases, creeping, and wall cleaning: understanding privacy in the age of Facebook. *First Monday* 15:2775. doi: 10.5210/fm.v15i1.2775

Richards, N., and Hartzog, W. (2016). Taking trust seriously in privacy law. *Stanford Technol. Law Rev.* 19, 431–472. doi: 10.2139/ssrn.2655719

Rogers, E. (2003). *Diffusion of Innovations, 4th Edn*. New York, NY: Free Press.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., and Camerer, C. (1998). Not so different after all: a cross-discipline view of trust. *Acad. Manage. Rev.* 23, 393–404. doi: 10.5465/amr.1998.926617

Rouvroy, A. (2008). Privacy, data protection, and the unprecedented challenges of ambient intelligence. *Stud. Ethics Law Technol.* 2:1. doi: 10.2202/1941-6008.1001

Rueben, M., Aroyo, A. M., Lutz, C., Schmölz, J., Van Cleynenbreugel, P., Corti, A., et al. (2018). "Themes and research directions in privacy sensitive robotics," in *2018 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)* (Genova: IEEE), 1–8. doi: 10.1109/ARSO.2018.8625758

Rueben, M., Bernieri, F. J., Grimm, C. M., and Smart, W. D. (2017b). "Framing effects on privacy concerns about a home telepresence robot," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna: ACM), 435–444. doi: 10.1145/2909824.3020218

Rueben, M., Grimm, C. M., Bernieri, F. J., and Smart, W. D. (2017a). A taxonomy of privacy constructs for privacy-sensitive robotics. *arXiv*. preprint arXiv:1701.00841. Available online at: https://arxiv.org/pdf/1701.00841.pdf

Schaefer, K. E., Chen, J. Y., Szalma, J. L., and Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: implications for understanding autonomy in future systems. *Hum. Factors* 58, 377–400. doi: 10.1177/0018720816634228

Schulz, T., and Herstad, J. (2018). "Walking away from the robot: negotiating privacy with a robot," in *Proceedings of the 31th International BCS Human Computer Interaction Conference* (eWiC) (Swindon: British Computer Society). doi: 10.14236/ewic/HCI2017.83

Schulz, T., Herstad, J., and Holone, H. (2018). "Privacy at home: an inquiry into sensors and robots for the stay at home elderly," in *International Conference on Human Aspects of IT for the Aged Population* (Las Vegas: Springer), 377–394. doi: 10.1007/978-3-319-92037-5_28

Sedenberg, E., Chuang, J., and Mulligan, D. (2016). Designing commercial therapeutic robots for privacy preserving systems and ethical research practices within the home. *Int. J. Soc. Robot.* 8, 575–587. doi: 10.1007/s12369-016-0362-y

Smith, H. J., Dinev, T., and Xu, H. (2011). Information privacy research: an interdisciplinary review. *MIS Q.* 35, 989–1016. doi: 10.2307/41409970

Solove, D. J. (2008). *Understanding Privacy*. Cambridge, MA; London:Harvard University Press.

Stutzman, F., Capra, R., and Thompson, J. (2011). Factors mediating disclosure in social network sites. *Comput. Human Behav.* 27, 590–598. doi: 10.1016/j.chb.2010.10.017

Syrdal, D. S., Walters, M. L., Otero, N., Koay, K. L., and Dautenhahn, K. (2007). ""He knows when you are sleeping" – Privacy and the personal robot companion," in *Proceedings of the 2007 AAAI Workshop Human Implications of Human–Robot Interaction* (Washington DC: AAAI), 28–33. Available online at: https://www.aaai.org/Papers/Workshops/2007/WS-07-07/WS07-07-006.pdf

Tamò-Larrieux, A. (2018). *Designing for Privacy and its Legal Framework*. Cham: Springer. doi: 10.1007/978-3-319-98624-1

Trepte, S., Reinecke, L., Ellison, N. B., Quiring, O., Yao, M. Z., and Ziegele, M. (2017). A cross-cultural perspective on the privacy calculus. *Social Media + Society* 3, 1–13. doi: 10.1177/2056305116688035

Turkle, S. (2011). "Authenticity in the age of digital companions," in *Machine Ethics*, eds M. Anderson and S. L. Anderson (Cambridge: Cambridge University Press), 62–76. doi: 10.1017/CBO9780511978036.008

Van den Berg, B. (2016). "Mind the air gap," in *Data Protection on the Move: Current Developments in ICT and Privacy/Data Protection*, eds S. Gutwirth, R. Leenes, and P. De Hert (Dordrecht: Springer), 1–24. doi: 10.1007/978-94-017-7376-8_1

Van Deursen, A., and Van Dijk, J. (2011). Internet skills and the digital divide. *New Media Society* 13, 893–911. doi: 10.1177/1461444810386774

Venkatesh, V., and Morris, M. G. (2000). Why don't men ever stop to ask for directions? Gender, social influence, and their role in technology acceptance and usage behavior. *MIS Q.* 24, 115–139. doi: 10.2307/3250981

Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. (2003). User acceptance of information technology: toward a unified view. *MIS Q.* 27, 425–447. doi: 10.2307/30036540

Wagner, A. R., Borenstein, J., and Howard, A. (2018). Overtrust in the robotic age. *Commun. ACM* 61, 22–24. doi: 10.1145/3241365

Waldman, A. E. (2018). *Privacy as Trust: Information Privacy for an Information Age*. Cambridge: Cambridge University Press. doi: 10.1017/9781316888667

Warren, S. D., and Brandeis, L. D. (1890). The Right to privacy. *Harward Law Rev.* 4, 193–220. doi: 10.2307/1321160

Weiss, A., Wurhofer, D., and Tscheligi, M. (2009). "I love this dog"—children's emotional attachment to the robotic dog AIBO. *Int. J. Soc. Robot.* 1, 243–248. doi: 10.1007/s12369-009-0024-4

Westin, A. (1967). *Privacy and Freedom*. Cambridge, MA: Atheneum Press.

Young, A. L., and Quan-Haase, A. (2013). Privacy protection strategies on Facebook: the Internet privacy paradox revisited. *Inform. Commun. Soc.* 16, 479–500. doi: 10.1080/1369118X.2013.777757

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Robotics and AI

HYPOTHESIS AND THEORY
published: 04 May 2021
doi: 10.3389/frobt.2021.632394

A Deeper Look at Autonomous Vehicle Ethics: An Integrative Ethical Decision-Making Framework to Explain Moral Pluralism

Jimin Rhim[1,2], *Ji-Hyun Lee*[3], *Mo Chen*[2] *and Angelica Lim*[1]*

[1]*Robots with Social Intelligence and Empathy (ROSIE) Lab, School of Computing Science, Simon Fraser University, Burnaby, BC, Canada,* [2]*Multi-Agent Robotic Systems (MARS) Lab, School of Computing Science, Simon Fraser University, Burnaby, BC, Canada,* [3]*Information-based Design Research Group, Korea Advanced Institute of Science and Technology (KAIST), Graduate School of Culture Technology, Daejeon, South Korea*

The autonomous vehicle (AV) is one of the first commercialized AI-embedded robots to make autonomous decisions. Despite technological advancements, unavoidable AV accidents that result in life-and-death consequences cannot be completely eliminated. The emerging social concern of how an AV should make ethical decisions during unavoidable accidents is referred to as the moral dilemma of AV, which has promoted heated discussions among various stakeholders. However, there are research gaps in explainable AV ethical decision-making processes that predict how AVs' moral behaviors are made that are acceptable from the AV users' perspectives. This study addresses the key question: What factors affect ethical behavioral intentions in the AV moral dilemma? To answer this question, this study draws theories from multidisciplinary research fields to propose the "Integrative ethical decision-making framework for the AV moral dilemma." The framework includes four interdependent ethical decision-making stages: AV moral dilemma issue framing, intuitive moral reasoning, rational moral reasoning, and ethical behavioral intention making. Further, the framework includes variables (e.g., perceived moral intensity, individual factors, and personal moral philosophies) that influence the ethical decision-making process. For instance, the framework explains that AV users from Eastern cultures will tend to endorse a situationist ethics position (high idealism and high relativism), which views that ethical decisions are relative to context, compared to AV users from Western cultures. This proposition is derived from the link between individual factors and personal moral philosophy. Moreover, the framework proposes a dual-process theory, which explains that both intuitive and rational moral reasoning are integral processes of ethical decision-making during the AV moral dilemma. Further, this framework describes that ethical behavioral intentions that lead to decisions in the AV moral dilemma are not fixed, but are based on how an individual perceives the seriousness of the situation, which is shaped by their personal moral philosophy. This framework provides a step-by-step explanation of how pluralistic ethical decision-making occurs, reducing the abstractness of AV moral reasoning processes.

Keywords: autonomous vehicle, AI ethics, ethical decision-making, moral pluralism, explainability, transparency

OPEN ACCESS

Edited by:
Martin Magnusson, Örebro University, Sweden

Reviewed by:
Veljko Dubljevic, North Carolina State University, United States
Christoph Lütge, Technical University of Munich, Germany

***Correspondence:**
Angelica Lim
angelica@sfu.ca

Specialty section:
This article was submitted to Ethics in Robotics and Artificial Intelligence, a section of the journal Frontiers in Robotics and AI

Received: *23 November 2020*
Accepted: *07 April 2021*
Published: *04 May 2021*

Citation:
Rhim J, Lee J-H, Chen M and Lim A (2021) A Deeper Look at Autonomous Vehicle Ethics: An Integrative Ethical Decision-Making Framework to Explain Moral Pluralism. Front. Robot. AI 8:632394. doi: 10.3389/frobt.2021.632394

INTRODUCTION

With recent artificial intelligence (AI) advancements, robots are expanding from conducting predefined tasks in confined environments to becoming autonomous agents in real-world contexts. Autonomous vehicles (AVs) are among the most significant commercialized AI-embedded autonomous agents that reflect this technological transition. A report of Americans' long-term adoption of AVs forecasts mass production of AVs with high automation by 2024 (Bansal and Kockelman, 2017). The adoption of AV promises many benefits that improve transportation experiences such as reduced costs, more rest time for vehicle users, mobility to nondrivers, and minimized pollutions (Schoettle and Sivak, 2014; Fagnant and Kockelman, 2015). Most importantly, AVs are expected to increase road safety by reducing the number of accidents and severity of crash consequences by making more rational decisions (Anderson et al., 2014; Kumfer and Burgess, 2015; Nyholm and Smids, 2016; Gogoll and Müller, 2017; Hulse et al., 2018).

Despite these technological advancements, AV accidents cannot be entirely eliminated (Goodall, 2014b; Bonnefon et al., 2016; Guo et al., 2018; Nyholm and Smids, 2018). In this regard, AVs are among the first autonomous agents that make decisions with potential life-and-death consequences (Awad et al., 2020). While vehicle accidents have existed, the introduction of AVs has shifted ethical implications during accidents Danielson (2015), Shariff et al. (2017), Awad et al. (2018a), Taddeo & Floridi (2018) because humans and AVs make intrinsically different ethical decisions. In conventional accidents, human drivers tend to show crash avoidance behaviors Lerner (1993), Yan et al. (2008) within 2 s of reaction time Lin (2015), resulting in reflexive and instinctive decisions (Goodall, 2014a). Thus, human decisions or driving behaviors cannot be held morally accountable (Goodall, 2014b; Lin, 2015; Shariff et al., 2017). In contrast, AVs are equipped with advanced sensors and preprogrammed algorithms that can anticipate and react to accidents better than human drivers. Therefore, AV decisions that impact human lives are preprogrammed (Goodall, 2014a; Carsten et al., 2015; Karnouskos, 2020b). The decision of an AV to protect whom or what during an emergency falls into distributing harm, a universally agreed-upon moral domain (Haidt, 2001). As an AV is an artificial moral agent capable of making decisions with ethical consequences (Allen et al., 2005; Wallach et al., 2010), an in-depth understanding of AV ethics is necessary.

The emerging social concern of how AVs should behave ethically in unavoidable crashes started a heated discussion in AV ethics, which is referred to as the moral dilemma of AVs (Bonnefon et al., 2016; Gogoll and Müller, 2017; Goodall, 2014a, Goodall, 2014b; J. Greene, 2016; Hevelke and Nida-Rümelin, 2015; Lin, 2015; Nyholm and Smids, 2016). The most dominantly discussed AV ethical issue is based on an extension of the trolley problem Goodall (2014a), J. Greene (2016), Lin (2015), Shariff et al. (2017), which asks whether people prefer deontology (determining good or bad based on a set of rules) or utilitarianism (determining good or bad based on outcomes) (Gawronski and Beer, 2017). However, many researchers are dismissive of AV ethics based on the trolley problem for the following reasons. First, the hypothetical scenarios adopted in the thought experiment are too simplified and ambiguous (Gawronski and Beer, 2017; De Freitas et al., 2020b). In fact, most scenarios in AV moral dilemmas tend to focus mainly on the consequences made from predefined binary choices, e.g., the number or characteristics of people who are impacted. This approach disregards other important AV crash-related factors such as regulations, responsibilities, or moral norms. Second, the results are highly likely to be biased. Trolley problem-based scenarios often begin by favoring a specific moral theory, resulting in a biased interpretation of the results (Dubljević and Racine, 2014). Studies have shown a discrepancy between people's preference and acceptance of utilitarian AVs due to this bias. For instance, people answered that they prefer utilitarian AVs that save more lives but would not purchase such AVs, as they might sacrifice themselves (Bonnefon et al., 2016; Shariff et al., 2017; Awad et al., 2018a). Third, ethical decisions based on the trolley problem tend to be unfair (Goodall, 2014a; J. Greene, 2016; Taddeo and Floridi, 2018). The results reveal people's preferences to determine who to kill based on personal features (e.g., save women and kill men) Bigman & Gray (2020), which disregards the equal right to human, an integral ethical concern (Kochupillai et al., 2020). Further, such unfair preferences violate the Rule 9 of German Ethics Code for Automated and Connected Driving, which strictly prohibits discrimination based on personal features (Luetge, 2017). As a result, people are angered at AVs that make prejudiced decisions (De Freitas et al., 2021). Consequently, public fear and outrage could delay the adoption of AVs (Shariff et al., 2017). Finally, trolley problem-based AV ethics tends to rely on a single moral doctrine (e.g., utilitarian). Relying only on one specific moral principle cannot explain complex real-world values. Indeed, human morality is pluralistic (Graham et al., 2013; Schoettle and Sivak, 2014; Fagnant and Kockelman, 2015). Therefore, providing AV ethical perspectives other than utilitarianism needs to be considered (Dubljević, 2020). To overcome the limitations of the trolley problem-based AV ethics, an alternative approach that incorporates varying human values and crash contexts should be considered.

Providing explainable AV moral behaviors is essential to ensuring the transparency of AV systems (J. Greene, 2016). One way to achieve this goal is to develop an AV framework that explains and predicts the full ethical decision-making process Winfield et al. (2019), Karnouskos (2020a) matching end-users' values (Bonnemains et al., 2018). AV ethics requires a collaborative and interdisciplinary effort from technical, regulatory, and social spheres (Borenstein et al., 2019; De Freitas et al., 2020a; De Freitas et al., 2020b; Mordue et al., 2020). Therefore, it is integral for various stakeholders (e.g., AV developers, engineers, regulators, ethicists, and social scientists) to have an open discussion about forming value-aligned moral behaviors of AV Goodall (2014b), De Freitas et al. (2020b). As AI-based reasoning is a blackbox Castelvecchi (2016), AV moral reasoning will be challenging to fully understand, even for those who programmed them. Furthermore, AVs are mostly elaborated by engineers, transportation experts, policy makers Bansal and Kockelman,

(2017) and AI ethicists Vrščaj et al. (2020) lacking prospective AV users' values or expectations. Further, experiment results show that moral judgments on human drivers and AVs were similar (Kallioinen et al., 2019). Consequently, many researchers emphasize the importance of including public morality and preference in AV ethics (Awad et al., 2018b; De Freitas et al., 2020a; Savulescu et al., 2019; De Freitas et al., 2020a; De Freitas et al., 2020a). It is important to note that the focus of this study is limited to understanding acceptable AV moral behaviors for the public, which has been underexplored. Thus, technical approaches to implement the system are beyond the scope of this research.

The study that observed lay drivers' moral reasoning showed that moral emotions are an important part of moral judgment during the AV moral dilemma (Rhim et al., 2020). Accordingly, a comprehensive ethical decision-making framework that explains both intuitive and rational aspects of AV ethical behaviors that answers the following research questions is required: What factors affect ethical behavioral intentions in the AV moral dilemma? How do these variables shape ethical behavioral intentions? To answer these questions, this study aims to synthesize a framework that uses the dual-process theory of moral reasoning Greene et al. (2001) to explain and predict pluralistic moral reasoning in the AV moral dilemma.

This study attempts to provide descriptive ethics to enhance understanding of the broad ethical phenomena of the AV moral dilemma by providing a conceptual framework with propositions. The assumption that acceptable or understandable AV behaviors can be learned from the existing data should be avoided De Freitas et al. (2020b), because there are not enough AV crash cases and the discussion of acceptable AV moral behaviors is not finalized. As a result, it is neither possible nor realistic to provide normative guidance that lists how AVs "ought to" behave. Moreover, the established normative AV ethics may not be adequate as AV technology would advance in unexpected ways, or user values may evolve while using the technology. Also, once AVs are embedded in daily lives, it would be difficult to modify AV decisions and policies (Vrščaj et al., 2020). Thus, making normative ethical rules should be done with caution (Dubljević, 2020). In summary, the purpose of this research is to propose a comprehensive conceptual framework called the "Integrative ethical decision-making framework for the AV moral dilemma," which theorizes that individual characteristics and perceived seriousness of the AV moral dilemma are antecedents of intuitive and rational moral judgments. The contributions of this study are as follows. First, this study provides explanations for the dual-process theory of ethical decision-making during the AV moral dilemma by including both the cognitive and affective mechanisms as integral aspects of AV ethics. Second, this study emphasizes the importance of how the issue is framed instead of focusing only on the impact of a specific moral doctrine to explain flexible and versatile moral judgment during the AV moral dilemma. Last, this study provides a holistic view of how ethical decision-making occurs in the unknown and vague context of the AV moral dilemma, by providing definitions of moderating variables with explanations and propositions.

BACKGROUND

Review of Theoretical Ethical Decision-Making Approaches

Extending human morality literature into artificial agents may facilitate the articulation of computational models (Wallach, 2010; Malle, 2016; Cervantes et al., 2020). Therefore, having a comprehensive understanding of the existing moral judgment theories is crucial to building realistic and accountable AV ethical behaviors. The definition of ethical decision-making is "a process by which individuals use their moral base to determine whether a certain issue is right or wrong" (Carlson et al., 2009, p. 536). Researchers from multiple disciplines have proposed a number of theoretical and conceptual frameworks to explain, predict, and learn about human moral reasoning. Although moral judgment models are not specifically devised to explain AV ethics, some of the representative models have evolved over several decades to provide comprehensiveness to explain complex moral dilemma scenarios, which offers general applicability to other fields (S. D. Hunt and Vitell, 2006). Therefore, understanding the human moral reasoning will provide possible explanations of how moral judgment will occur in the AV moral dilemma.

Traditional moral reasoning approaches are based on rationalist approaches, which posit that people make conscious and intentional ethical decisions (Vitell, 2003). Recently, social psychologists began to focus on the nonrational or intuitionist approaches in moral reasoning by emphasizing the importance of intuition and emotions in moral reasoning (Haidt, 2001; Sonenshein, 2007; Dubljević et al., 2018). Therefore, this study attempts to gain significant insights from a theoretical investigation of the dual-process theory Haidt (2001), Kahneman (2003), Evans (2008), Zollo (2020) by understanding both the rationalist and intuitive approaches to explain socially acceptable AV ethical behaviors.

The Rationalist Approach

Rest's model has inspired rationalist ethical decision-making frameworks in the literature across many disciplines (Beu and Buckley, 2001; Dubinsky and Loken, 1989; Ferrell and Gresham, 1985; S. D. Hunt and Vitell, 1986; Jones, 1991; Trevino, 1986). The rationalist approach of ethical decision-making can be summarized as representing a cognitive perspective of an individual, which is rational, controlled, deliberate, intentional, and conscious. The most widely acknowledged ethical decision-making framework is the four-component model by Rest (1986), which is the foundation of most models (Groves et al., 2008). Rest's model, as well as the majority of ethical decision-making frameworks, begins when a person recognizes that there is an ethical issue, which is called the *Recognize Moral Issue* phase. If an ethical issue has been recognized, an individual's reasoning moves on to the next step of *Make Moral Judgment*, which is an individual's cognitive process to "judge which course of action is morally right" (Trevino, 1992, p.445), then the third step called *Establish Moral Intent* follows. This is a cognitive moral development phase that occurs after making a moral judgment Kohlberg (1969), Rest (1986), in which people prioritize their moral values to determine appropriate ethical behaviors. The last

TABLE 1 | Characteristics of AV moral dilemma vignette (source: Rhim et al., 2020, p. 44).

Vignette description	Crash option and result	Moral conflict description
The participant is a driver (V1) who is driving a truck at a two-lane road in a rural area. There are three small passenger cars (V2, V3, and V4) and a truck (V5) on the road. Suddenly, (V2) is changing lane, and a head-on collision with (V1) is expected. There are two crash options with known consequences, and the participant has to choose an option from the perspective of the driver (V1)	The truck (V1) brakes and turns right. This will lead to a collision between the truck (V1) and a small passenger car (V2). As a result, the driver of (V1) will get a minor injury, while the driver of (V2) has died	Whether to make a self-protecting decision that results in the death of the negligent driver
	The truck (V1) turns right. This will lead (V1) to deviate off the road and collide into a utility pole. As a result, the driver of (V1) will be seriously injured while all the other drivers are intact	Whether to make a utilitarian decision to save a life on behalf of sacrificing oneself when one is not at fault

step is *Engage in Moral Behavior*, in which an individual makes actions based on his or her moral intentions. These four phases describe the moral reasoning of individuals to be "intentionally rationalize, re-evaluate, and justify, moral standards, rules of conduct, and moral life" (Zollo et al., 2018, *p*.694).

The following are the examples of rationalist ethical decision-making frameworks from the multidisciplinary literature that are based on Rest's (1986) Model. The contingency framework by Ferrell and Gresham (1985) describes that an individual's moral reasoning begins when he or she faces an ethical salient context. This model synthesizes multiple variables to explain whether an individual's behavior is ethical or unethical. An individual's moral reasoning is influenced by the following factors: individual (i.e., knowledge, values, attitudes, intentions), significant others (i.e., differential association, role set configuration), and opportunity (i.e., professional codes, corporate policy, rewards/punishment). This model also includes social and cultural environmental factors that shape an individual's ethical intentions. The Person-Situation Interactionist model by Trevino (1986) implements the stage of Kohlberg's cognitive moral development (Kohlberg, 1969) as an integral predictor of ethical behavior. Moral judgment in Trevino's model is moderated by both an individual moderator (i.e., ego strength, field of dependence, and locus of control) and a situational moderator (i.e., immediate job context, organizational culture, characteristic of the work). The general theory of marketing ethics of Hunt and Vitell (1986) was developed to reduce the ethics gap between the marketers and the society by providing a general ethical decision-making theory with a visible process model (S. D. Hunt and Vitell, 2006). This model is similar to the Contingency framework by Ferrell and Gresham, (1985) as both acknowledge the impact of external factors (i.e., cultural, industry, and organizational environment) and individual factors in moral judgment. However, the Hunt and Vitell (1986) model explains that individuals use specific moral doctrines (deontological or teleological) to evaluate and determine ethical consequences during perceived ethical problem stages. That is, this model puts emphasis on the micro aspects of an individual's cognitive decision-making process. Jones' (1991) issue-contingent model includes the four moral reasoning phases like other models and proposes that environmental factors and individual factors positively impact the ethical decision-making phases. On top of this, Jones (1991) emphasizes the moral intensity of a particular context (see **Table 1**) for further definitions and application for AV ethics). A comprehensive rationalist ethical decision-making framework is illustrated in (**Figure 1**). While these models provide variables and their relations that explain how individuals perform moral reasoning, they focus on a rationalist approach. Thus, the rationalist approach does not consider the role of emotions or intuitions, which are integral components of moral value codes derived in the AV moral dilemma (Rhim et al., 2020).

The Intuitionist Approach

Researchers have realized that the dominant rational perspective fails to convey the full spectrum of the ethical decision-making processes (Chatzidakis et al., 2018; Cherry and Caldwell, 2013; Yacout and Vitell, 2018). The premise that moral agents are rational decision makers disregards the impact of nonrational or intuitive elements such as emotions and intuition in moral judgment (Sonenshein, 2007; Ruedy et al., 2013; Chowdhury, 2017). Consequently, researchers began to acknowledge the significance of intuitive approaches in ethical decision-making, which include consideration of moral values, emotions, and intuitions (Cherry and Caldwell, 2013; Dedeke, 2015; Haidt, 2001; Haidt and Joseph, 2004; Zollo, forthcoming; Zollo et al., 2017). Haidt (2001) defines moral intuition as "the sudden appearance in consciousness of a moral judgment, including an affective valence (good-bad, like-dislike), without any conscious awareness of having gone through steps of searching, weight evidence, or inferring a conclusion" (p.818).

The dual-process theory of human cognition Kahneman (2003), Evans (2008) explains that moral intuition is an automatic response antecedent to rational moral reasoning (Haidt, 2001; Sonenshein, 2007). The social intuitionist model Haidt (2001), among the most well-known intuitionist models, adopts the dual-process theory and accentuates the role of moral intuition as the initial stage in moral reasoning (Greene et al., 2001; Cushman et al., 2006; Zollo et al., 2017). The theory explains that when the decision maker experiences a morally salient context, he or she makes moral judgments based on intuitions, followed by the post hoc rationalization of moral reasoning. In summary, Haidt (2001) explains that emotive intuition occurs quickly and effortlessly, whereas cognitive reasoning occurs slowly and requires efforts.

Another well-known dual process theory includes the notion of *System 1* and *System 2* (Evans, 2008; Kahneman, 2003). Under this theory, human cognition comprises two information processing

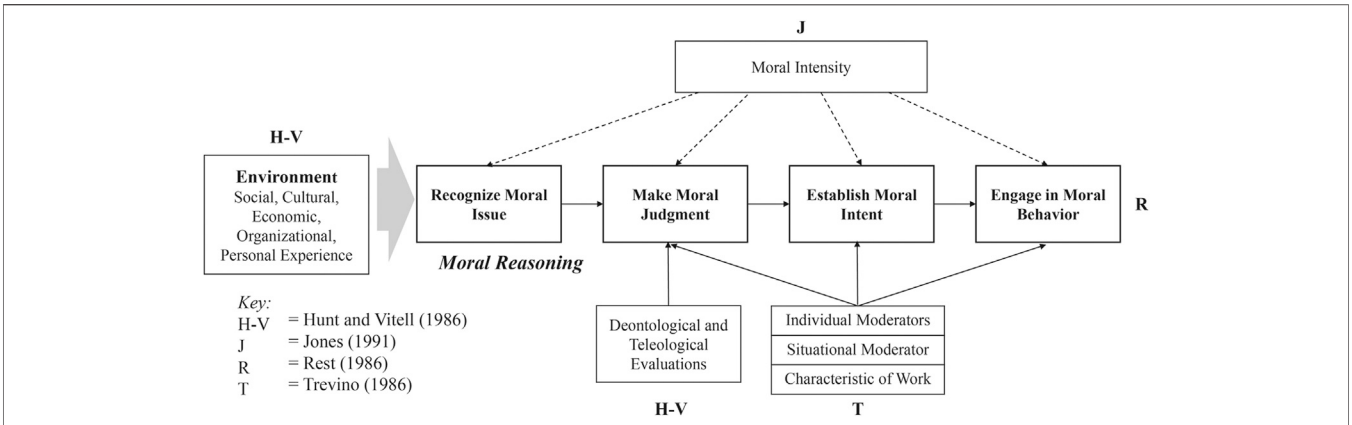

FIGURE 1 | Rationalist approach of ethical decision-making.

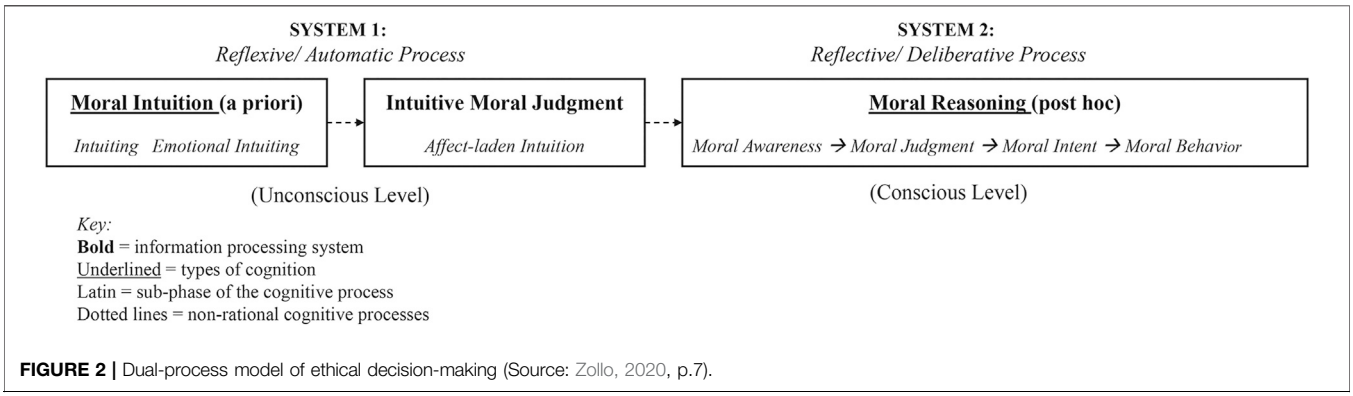

FIGURE 2 | Dual-process model of ethical decision-making (Source: Zollo, 2020, p.7).

systems, which also apply to the ethical decision-making process (Zollo, 2020). *System 1* is the intuitive, effortless, fast, reflexive, and nonconscious cognitive process (Dane and Pratt, 2007). "Intuiting" can be interpreted as *System 1*, which allows a moral agent to make a holistic and intuitive moral judgment during dynamic and uncertain situations (Dane and Pratt, 2007). The next phase, *System 2*, is the controlled, reflective, and analytical cognitive moral reasoning process (Zollo et al., 2017). Basic emotions that arise effortlessly and unconsciously are part of *System 1* (i.e., fear, surprise, and sadness), whereas *System 2* includes more complex emotions that are derived from deliberate, and rational cognition (i.e., disgust, anguish, relief, and embarrassment) (Metcalfe and Mischel, 1999; Zollo et al., 2017). Adopted from Zollo (2020), **Figure 2** shows the dual process of ethical decision-making, which includes both moral intuition (System 1) and cognitive moral reasoning (System 2). A more recent study in neuroethics introduced the Agent–Deed–Consequence (ADC) model of moral judgment, which follows an integrative approach to explain moral intuitions (Dubljević and Racine, 2014). More specifically, the ADC model posits that "moral judgment relies on positive and negative evaluations of three different components of moral intuitions: the character of a person; their actions; and the consequences brought about by the given situation" (Dubljević et al., 2018, p.2). The ADC model is simple yet effective in verifying and explaining whether a behavior is ethical or not. Overall, the moral intuitionists Cushman et al.(2006), Greene et al. (2001), Haidt (2001), Sonenshein (2007), Tenbrunsel and Smith-Crowe (2008), Zollo (2020), Zollo et al. (2017) agree that "moral judgments arise as intuitions generated by automatic cognitive processes, and that the primary role of conscious reasoning is not to generate moral judgments, but to provide a post hoc basis for a moral justification" (Cushman et al., 2006, *p.* 1982). Recent literature on the ethics indicates that considering both the rationalist and intuitive approaches provides a complete understanding of human moral reasoning. Moreover, as AV accidents impose hazards for both individual AV user and the traffic users around the AV user, consideration of intuitive moral judgment along with rational judgment to consider overall impact for the society is important. Consequently, the AV ethics should be in line with the dual-process theory and consider both the rational and intuitive moral judgment phases to discuss socially acceptable AV morality.

Linking Ethical Decision-Making and AV Ethics

The various ethical decision-making frameworks listed in the previous sections are effective at providing explanations for how moral reasoning variables shape an individual's ethical intentions.

Many researchers agree with the necessity of formulating AV ethics frameworks for varying reasons. First, providing a formal specification of AV moral behaviors will aid other traffic users (e.g., cyclists and pedestrians) to have a better understanding of AVs (Dogan et al., 2016; Mermet and Simon, 2016). Second, an appropriate AV ethics framework helps decision-makers advance responsible AVs that align with societal values, Stilgoe et al. (2013), which can mitigate conflicts between potential harms when adopting AVs (Leikas et al., 2019; Vrščaj et al., 2020). Third, a comprehensive AV ethics model will facilitate translating vague real-world moral theories into machine operationalizable codes by reducing abstractness (Bonnemains et al., 2018).

Several AV ethics frameworks were developed in an attempt to fulfill these goals. Karnouskos (2020b) has utilized the utilitarian principle to explain the acceptance of AVs. Although this model is based on empirical findings, it relies only on a single ethical approach, which can lead to biased decisions. To overcome this limitation, Karnouskos (2020a) has verified that multiple moral frameworks (e.g., utilitarianism, deontology, relativism, absolutism, and pluralism) impact the acceptance of AVs. However, these models do not take into consideration situational or individual factors that impact ethical decision-making. While Smith (2019) has concluded that personality (Honest-Humility vs. Conscientiousness) and ethics positions (Idealism vs. Relativism) impact moral judgment during AV accidents, the model has a gap in explaining the procedural relationships among the variables. The "Generalized Framework for moral dilemmas Involving AV" categorizes layers of factors (cast of characters, vehicle assemblage, and perspective) and suggests four research agendas (Novak, 2020). However, Novak's model does not have clear definitions of concepts and their interrelations that explain the moral judgment process. While the aforementioned AV frameworks aim to provide accountable and transparent AV ethics, these models do not consider intuitive moral reasoning phases. Furthermore, these models cannot explain the pluralistic ethical decision-making of AV ethics required in complex and dynamic real-world crash contexts. To provide holistic explanations of ethical decision-making during the AV moral dilemma, this study aims to develop a comprehensive AV ethics framework by integrating both the intuitionist and rationalist moral reasoning approaches and understanding how individual and situational characteristics affect ethical decision-making phases.

METHODOLOGY

The theorization of explainable pluralistic AV ethical decision-making is based on the conceptual analysis method to "generate, identify, and trace a phenomenon's major concepts, which together constitute its theoretical framework" by linking together knowledge from multidisciplinary backgrounds (Jabareen, 2009, p.53). A conceptual framework is the end result of this method, which provides a broader understanding of the phenomenon of interest by providing explanations of possible relationships between concepts (Imenda, 2014; Liehr and Smith, 1999). Moreover, a conceptual framework lays a foundation for research questions and hypotheses for further investigation (McGaghie et al., 2001). This study follows the research stages of Eizenberg and Jabareen (2017), Jabareen (2009) to develop "Integrative ethical decision-making framework for the AV moral dilemma" depicted in **Figure 3**. First, multidisciplinary literature was reviewed in search of relevant concepts for the AV moral dilemma (e.g., ethics, psychology, sociology, traffic, law, machine ethics, and AI ethics). Second, the reviewed literature was categorized. As the moral reasoning process occurs when an individual perceives a morally salient context, the literature was classified to identify three initial categories: moral reasoning phases, individual factors, and situational factors impacting ethical decisions during the AV moral dilemma. Third, specific concepts were identified. For the moral reasoning category, four interdependent ethical decision-making stages were defined. Both intuitive and rational moral judgment stages were included to describe the dual-process and pluralistic nature of human moral reasoning. Concepts and propositions for both intuitive and rational moral judgment stages include moral value codes that were derived from the AV moral dilemma ethical decision-making process (Rhim et al., 2020). For the individual factors categories, concepts that describe the characteristics and ethical stance of an individual were identified. For situational factors impacting the moral reasoning phases, a variable called perceived moral intensity (PMI) was selected, which evaluates multiple aspects of the AV moral dilemma. PMI includes the perception of risk and uncertainty, important features to consider during AV accidents Kruegel and Uhl (2020); therefore, these two latter concepts were not included separately. Last, the selected concepts were synthesized to provide a comprehensive explanation of how ethical behavioral intentions are shaped during AV moral dilemmas. Further descriptions of the "Integrative ethical decision-making framework for the AV moral dilemma" will be provided in the next section.

THE PROPOSED MODEL: INTEGRATED AV ETHICAL DECISION-MAKING FRAMEWORK

No matter how complicated AVs are, they are products that can be represented as an extension of their users, owners, or occupants, as the driving task of AV is becoming a comanaged task with humans (Smith, 2019; Bellet et al., 2011). Therefore, the authors posit that AV users will better understand, accept, and trust AVs that make moral judgments similar to oneself. To explain the AV ethical decision-making process during the AV moral dilemma, we have reformulated an integrative ethical decision-making model that includes both the rationalist and intuitive approaches based on previous models (Singhapakdi et al., 1999; Haidt, 2001; Dedeke, 2015; Schwartz, 2016; Zollo, 2020). Aligned with Haidt (2001), Zollo (2020), our framework is descriptive, which describes how people are likely to make ethical intentions during the AV moral dilemma. This study defines the AV moral dilemma as an

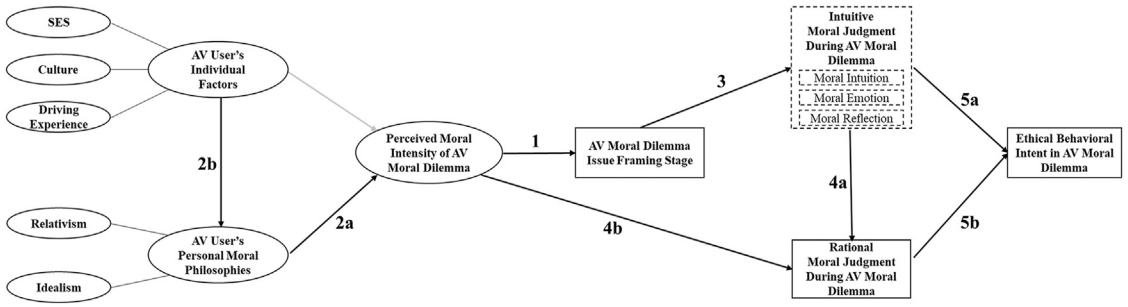

FIGURE 3 | Integrative ethical decision-making framework for the AV moral dilemma.

unavoidable crash situation in which an AV user must reflect upon competing moral standards and determine the appropriate moral behavior of an AV. Moreover, this model posits that moral judgment will vary depending on the individual (e.g., different individuals may perceive varied levels of moral saliency when faced with the same AV moral dilemmas) and situational characteristics (e.g., the same individual may behave differently depending on the characteristic of AV moral dilemma one is facing). The ADC model (Dubljević and Racine, 2014) is one of the most up-to-date and effective models to explain the flexible moral judgment of AVs and overcome the limitation of relying only on utilitarian AV ethics (Dubljević, 2020). The framework developed in this study is complementary to the ADC model. As the components of the ADC model indicate, the model assesses ethical consequences based on deeds of agents. While the Theory of Planned Behavior (Ajzen, 1991) links intention and behavior, studies in ethics demonstrate that how an individual intends to act may not necessarily lead to actual ethical behaviors during the moral dilemma (Weber and Gillespie, 1998). Consequently, understanding ethical intentions will provide further insights into why a certain ethical behavior or deed occurs. The "Integrated AV ethical decision-making framework" (**Figure 3**) in this study describes how ethical behavioral intentions are shaped with specific variables that need to be considered during the AV moral dilemma.

The "Integrated AV ethical decision-making framework" consists of two major components:1) the ethical decision-making process (intuitive and rational) and 2) variables (or factors) that influence the ethical decision-making process. The ethical decision-making process is composed of four stages: AV moral dilemma issue framing, intuitive moral reasoning, rational moral reasoning, and ethical behavioral intention making stages which reflect Rest's (1986) basic process framework. The ethical decision-making variables include 1) individual factors and 2) personal moral philosophy (PMP), and 3) perceived moral intensity (PMI). The model consists of 9 links, which are shown in arrows in **Figure 3**. The solid boxes represent mental state, and the dotted boxes represent mental processes. The current model assumes that accountable ethical behavior of an AV is contingent on the particular AV moral dilemma context that an individual faces. In summary, the "Integrated AV ethical decision-making framework" explains pluralistic nature of AV ethics by investigating how context-specific ethical intentions are shaped during the AV moral dilemmas.

AV Moral Dilemma Issue Framing Stage

It is widely accepted that moral judgment is based on how an individual perceives the moral issue rather than the actual characteristics of the issues (Jones, 1991; Robin et al., 1996; A. E. Tenbrunsel and Messick, 1999; Trevino, 1986). That is, the situational context impacts an individual's unique moral frame, which is a key component in the ethical decision-making process. It is highly likely that each AV crash's characteristics will be unique (e.g., number of passengers in the car, severity of the injury, damage done to one's vehicle, liability, relationship to the injured victims), and understanding how an individual frames the specific AV moral dilemma is important. According to Rhim et al. (2020), how participants framed the moral issue impacted their AV moral dilemma decisions. For instance, in the AV moral dilemma vignette three (see **Table 2**; **Figure 4**) that involved the conflict between making a self-protecting decision or following a utilitarian doctrine to minimize the overall harm, the individual's moral value code (e.g., Harm Mitigation vs. Self-Preservation) determined their decisions. Furthermore, locus of control is known to impact the moral issue framing stage (Forte, 2005; Dedeke, 2015). In the case of AV moral dilemma, when the locus of control was perceived as internal (making decisions in the first-person perspective), participants' ethical decisions varied (e.g., moral values: kin-preservation, pedestrian-preservation, physical harm avoidance, and responsibility distribution) (Rhim et al., 2020).

All these findings support the inclusion of the moral issue framing stage from the first-person perspective as the initial stage of ethical decision-making in the AV moral dilemma, in compliance with the Cognitive-Intuitionist Model (Dedeke, 2015). Moral issue framing in this framework posits that individuals organize the characteristics of moral issues based on the perceived seriousness of the AV moral dilemma, which is impacted by individual characteristics. Hence, the following proposition can be made:

TABLE 2 | Definition of moral intensity factors.

Factor	Definition [source: Jones (1991)]	Example in unavoidable AV crashes
Magnitude of consequences	"Sum of harms (or benefits) done to victims (or beneficiaries) of the moral act in question" (p. 374)	The AV's decision that causes the death of a person is more consequential than the one that causes a minor injury
Social consensus	"The degree of social agreement that a proposed act is evil (or good)" (p. 375)	The AV's decision to protect law-abiding pedestrians has a greater social consensus than a decision to protect the AV driver who has caused the accident
Probability of effect	"A joint function of the probability that the act in question will actually take place and the act in question will actually cause the harm (benefit) predicted" (p. 375)	The AV's decision that has the 10% probability of causing a serious injury to one passenger has a lower probability of effect than the decision that causes minor injury to all passengers with 100% probability
Temporal immediacy	"The length of time between the present and the onset of consequences of the moral act in question (shorter length of time implies greater immediacy)" (p. 376)	AV that causes harm to 1% of traffic users within 5 years has higher temporal immediacy than AV that harms 1% of traffic users within 20 years
Proximity	"The feeling of nearness (social, cultural, psychological, or physical) that the moral agent has for victims (beneficiaries) of the evil (beneficial) act in question" (p. 376)	The AV's decision that harms a passenger who is a family member has a higher proximity effect than when the effect will be experienced by a stranger in a different vehicle
Concentration of effect	"The moral act is an inverse function of the number of people affected by an act of given magnitude" (p. 377)	The AV's decision that leads to 10 fatalities has a higher concentrated effect than causing fatalities to 5 people

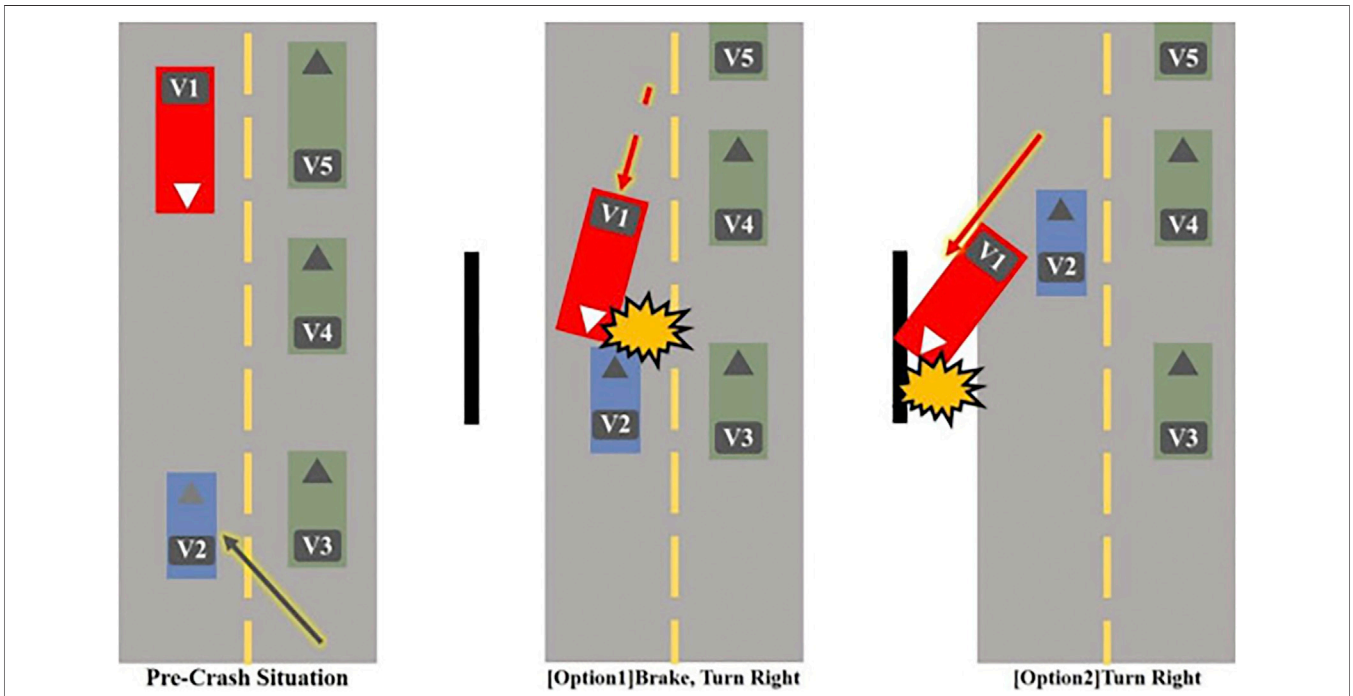

FIGURE 4 | AV moral dilemma: a rural two-lane road (source: Rhim et al., 2020, p. 45).

Link 1: The AV user frames the characteristics of moral issues based on his or her perceived seriousness of the AV moral dilemma.

The Consequence of Perceived Moral Intensity of AV Moral Dilemma

Extensive studies show that the characteristics of a moral issue will impact the ethical decision-making process. Characteristics of moral issues can be measured or described by moral intensity, which is defined as "a construct that captures the extent of issues-related moral imperative in a situation" (Jones, 1991, p.372). Moral intensity is composed of six components. See **Table 2** for a definition of each component with examples in the AV moral dilemma. This framework focuses on perceived moral intensity (PMI) because it is effective for describing moral perceptions that vary across situations and individuals. For instance, while an individual perceives the moral issue to be of high moral intensity, another individual might perceive the identical issue as being of low moral intensity (Robin et al. (1996) depending on his or her individual characteristics and perceptions of the context (further

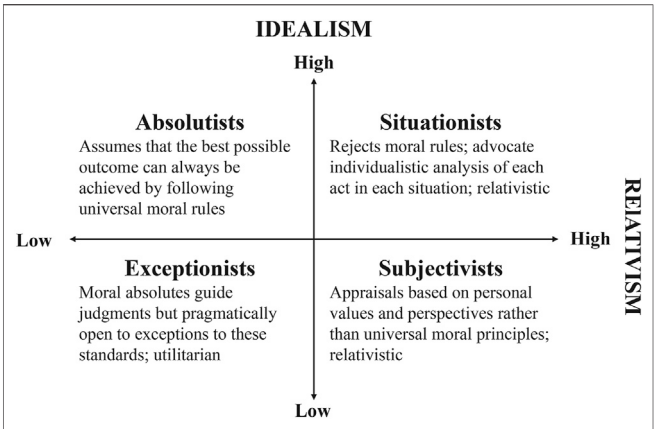

FIGURE 5 | Taxonomy of ethical ideologies (Source: Forsyth, 1980, p. 176).

explained in upcoming sections). Specifically, we posit that an AV moral dilemma that triggers high PMI will cause more extensive moral judgment cycles, while situations that prompt low PMI will lead to less in-depth moral judgment. Furthermore, empirical studies have shown a significant correlation between PMI and moral intents (Dubinsky and Loken, 1989; Ferrell et al., 1998; May and Pauli, 2002; Singhapakdi et al., 1999). Hence, this model expects that PMI will impact the ethical behavioral intent stage. In summary, this model specifies PMI as an integral variable that shapes ethical decision-making in the AV moral dilemma. Specifically, the characteristics of an AV accident will impact how an AV occupant frames the moral issue, which in turn will impact moral judgment and ethical behavioral intentions.

The Antecedents of Perceived Moral Intensity

PMI focuses on the exogenous characteristics of the moral situation, excluding traits of the moral decision-maker such as values, knowledge, or moral development (Ferrell and Gresham, 1985; Jones, 1991; Kohlberg, 1969). Therefore, AV users' innate variables impacting PMP will be explored in the following section.

AV User's Personal Moral Philosophy

Many researchers agree that a decision-maker will utilize ethical guidelines based on their personal moral philosophy (PMP) during ethically salient situations (Ferrell and Gresham, 1985; D. R. Forsyth, 1980; D. R. Forsyth et al., 1988, 2008; A. Singhapakdi et al., 1999; Vitell et al., 1993). Based on the established study results, this model presupposes that AV users will apply ethical guidelines based on their PMP when making ethical evaluations in AV moral dilemmas.

Forsyth (1980) explains that the predictors of an individual's moral judgments can be described by two nomothetic dimensions of PMP: relativism and idealism. Relativism indicates "the extent to which the individual rejects universal rules" when making ethical decisions. That is, relativists base their moral judgments on skepticism and "generally feel that moral actions depend upon the nature of the situation and individuals involved . . . more than the ethical principle that was violated" (Forsyth, 1992, p.462). On the other hand, idealists have "concern for the welfare of others . . . feel that harming others is always avoidable, and they would rather not choose between the lesser of two evils which will lead to negative consequences for other people" (Forsyth, 1992, p.462). Moreover, idealists feel that "desirable consequences can, with the 'right' action, always be obtained" (Forsyth, 1980, p.176). That is, idealists are moral optimists who value altruism.

Forsyth (1980) has classified four dichotomized ethical perspectives based on both dimensions rather than classifying individuals as either relativistic or idealistic, which is called the Ethics Position (see **Figure 5**). An individual's Ethics Position (Forsyth, 1980) is formed over a lifetime of experiences and has a strong impact on an individual's decision-making in a morally salient situation (D. R. Forsyth, 1980; D. R. Forsyth et al., 2008). Research results over the past 2 decades show relatively consistent findings. Idealism had an overall positive relation to moral judgment, whereas relativism had an overall negative relation to ethical decision-making (O'Fallon and Butterfield, 2013). Moreover, PMP has been empirically tested to operate through PMI (D. Forsyth, 1985; D. Forsyth and Pope, 1984; A. Singhapakdi et al., 1999). Based on the previous studies, this model explains that PMP will impact PMI in the AV moral dilemma. Further, this study proposes that AV occupants who score higher in idealism (e.g., who aim to secure the overall welfare of road sharers) and lower in relativism (e.g., who prioritizes protecting oneself more over others) will be more sensitive to ethical issues than their counterparts. Hence, we propose the following propositions:

Link 2a. PMP of AV user impacts PMI of AV moral dilemma.

-A more idealistic AV user will have a higher PMI than a less idealistic AV user
-A more relativistic AV user will have a lower PMI than a less relativistic AV user

AV User's Individual Factors as Antecedent of Personal Moral Philosophy

Singhapakdi et al. (1999) emphasized the role of individual characteristics in shaping PMP, which impact PMI and ethical decision-making processes. This study will explore the following individual and cultural factors that are likely to impact PMI in the AV moral dilemma: 1) Socioeconomic status (SES): income and education, 2) culture (or nationality), and 3) driving experience. Moreover, this model posits that individual characteristics impact the moral issue framing stage, which aligns with the model of (Sonenshein, 2007). Thus, the following proposition is developed:

Link 2b: Individual factors impact the moral issue framing stage of the AV moral dilemma.

Socioeconomic status.Income Despite the scarcity of previous studies, it is essential to explore the impact of SES on the perception of AV ethics for the following reasons. As SES affects AV users' acceptance of adopting AVs (wealthier people tend to favor and anticipate the adoption of AVs more) (Webb et al., 2019), it is more likely that users with higher SES will adopt AVs first. When a new product or

service is deployed, it is natural that feedback from the initial users will be incorporated to modify the product or service. In general, income tends to rise with the advancement of education levels. Relatively few studies explored the impact of income on ethical decision-making. Among a few empirical results, Pratt (1991) found a consistent tendency for higher salary individuals to be more sensitive to unethical actions than those with lower salaries. Moreover, Singhapakdi et al. (1999) found that salary was negatively related to relativism. Similarly, the ethical perceptions of AV users who are higher in SES are highly likely to be referenced more for modifying the ethical behaviors of AVs. Therefore, before an actual system is implemented, it is imperative to explore the PMP of a wide range of SES, which in turn would impact the overall perception of the ethical decision-making process of AVs.

Link 2b-1: Income will have an impact on PMP.

-AV users with higher income will be more idealistic than AV users with lower income

-AV users with higher income will be less relativistic than AV users with lower incomeEducation Studies in ethics have included education (types and number of years) as a variable that impacts ethical decision-making because education is linked to an individual's cognitive moral development stages (Rest, 1986). Some study results showed significant differences in moral reasoning among individuals with different education levels (Wimalasiri et al., 1996; Latif, 2001; Kracher et al., 2002). For instance, Sparks and Hunt (1998) found that individuals with more domain knowledge were more ethically sensitive than novices. Cole and Smith (1996) found that less educated individuals were more accepting of ethically questionable statements than more educated people. Moreover, people showed a significant difference in recognition of ethical scenarios after receiving education (Wu, 2003).

Singhapakdi et al. (1999, p.23) explain that education shows a noticeable impact on PMP, because "with education may come greater sensitivity to alternative points of view, skepticism regarding moral absolutes, and pessimism that moral dilemmas can always have desirable outcomes." Moreover, ethical decision-makers in higher education are conventionally at higher moral development levels, thus becoming more aware of people holding varying values or rules that can be relative to one's norm (Kholberg, 1969). Likewise, AV users with higher education levels are likely situated at higher stages of moral development, which enables consideration of the overall impact of crash consequences. For these reasons, the following propositions are developed:

Link 2b-2: The education level of an AV user will have an impact on PMP.

-More educated AV users are less idealistic than less educated AV users

-More educated AV users will be more relativistic than less educated AV usersCulture It is widely accepted that culture influences an individual's perception of moral dilemmas and the ethical decision-making process (Ferrell and Gresham, 1985; Graham et al., 2013; Hunt and Vitell, 1986; Hunt and Vitell, 2006). Further, it would be neither feasible nor acceptable to develop universally agreed upon AV ethics, as preferred moral decisions vary depending on cultures or countries (Awad et al., 2018b; De Freitas et al., 2020b; Dubljević, 2020). There are various definitions for culture, but one of the most accepted definitions is by Hofstede, which defines culture as "the collective programming of the mind that distinguishes the members of one group or category of people from another" (Hofstede et al., 2005, p.516). As culture includes values, shared beliefs, norms, and ideals Reidenbach and Robin (1991), moral obligations that are socially acceptable in one culture are rejected in other societies, despite the existence of universal moral principles (Mikhail, 2007). Moreover, cross-cultural studies in AV ethics indicated that people from different cultural backgrounds favored different AV moralities (Awad et al., 2018a; Ranasinghe et al., 2020; Rhim et al., 2020).

Forsyth et al. (2008) conducted a meta-analysis to investigate cultural differences by measuring the level of PMP. The review of 139 studies (29 nations, total n = 30,230) revealed that idealism and relativism levels vary across cultures in predictable ways and dominant ethics positions existed in each culture: Western culture (subjectivism), Eastern cultures (situationism), and Middle Eastern cultures (absolutism and situationism). The variations of idealism and relativism tend to be uniform with cultural characteristics (e.g., Hofstede and McCrae, 2004; Inglehart and Baker, 2000). Forsyth et al. (2008) explain that regarding idealism, it is predicted that Western cultures adopt less idealistic moral philosophies compared to the Eastern cultures, which can be explained by individualism (a defining characteristic of Western culture). Individualism focuses on the independence of each individual and allows the pursuit of autonomy and free will among groups, whereas collectivism (a defining characteristic of Eastern Culture) prioritizes the goal or well-being of a group before an individual. Thus, Eastern cultures that accentuate a sense of collectivism imply higher idealism than Western cultures. In terms of relativism, it is expected that Eastern cultures will be more relativistic than Western cultures. Eastern cultures tend to be more contextual and relational in comparison with Western cultures (Forsyth et al., 2008). In terms of ethics position, situationism (high idealism and high relativism, see **Figure 3**) is dominant in Eastern cultures. Situationists posit that an individual should act to secure the most beneficial consequences for all the group members, even if such a consequence is the result of violating moral rules. The situationists' moral outlook can be described by ethical skepticism or value pluralism, which suggests that the consequences of an action can determine the situation's moral values (D. R. Forsyth, 1992). On the other hand, Western cultures' dominant ethics position classification is exceptionist (low idealism and low relativism, see **Figure 3**), which posits that an individual fundamentally seeks to follow moral rules but is open to pragmatic results. The exceptionist moral outlook highly corresponds to "rule-utilitarianism," which indicates that "moral principles are useful because they provide a framework for making choices and acting in a way that will tend to produce

the best consequences for all concerned" (Forsyth, 1992, p. 463). Cross-cultural studies in AV ethics showed similar patterns. Eastern cultures showed a higher tendency to make context-dependent decisions during AV moral dilemmas (Rhim et al., 2020). On the other hand, the Westerns culture showed a stronger tendency to spare a greater number of people during the AV moral dilemma Awad et al. (2018b), Rhim et al. (2020), which corresponds to the exceptionist moral outlook. In summary, it is expected that cultural background can have a general impact on PMP. Hence, the following propositions are provided:

Link 2b-3: The cultural background of an AV user will have an impact on their PMP.

-AV users from Eastern cultures will tend to be more idealistic than AV users from Western cultures
-AV users from Eastern cultures will tend to be more relativistic than AV users from Western cultures
-AV users from Western cultures will generally endorse an exceptionist ethics position (Low idealism, Low relativism)
-AV users from Eastern cultures will generally endorse a situationist ethics position (High idealism, High relativism)Driving Experience Crashes caused by teen drivers comprise a major part of conventional vehicle collisions. The causes of teen crashes include inexperience in driving and underestimation of perilous driving behaviors (Williams, 2003; Rhodes and Pivik, 2011). Conversely, older drivers are likely to have more experience and have driven longer distances, thus are likely to have experienced situations with a greater variety of ethical problems. As studies that investigate the correlation between ethical decision-making and driving experiences are underexplored, the current study will refer to ethics studies that explored age as a predictor of ethical decision-making, as age and driving experience have a possible association. According to a meta-analysis, more than twenty studies have observed a positive relationship between age and ethical decision-making (O'Fallon and Butterfield, 2013). Study results show that older individuals tend to be more ethically sensitive than younger individuals (Karcher, 1996; Deshpande, 1997; Peterson et al., 2001). Furthermore, older generations made more ethical decisions than younger generations (Hunt and Jennings, 1997; Lund, 2000; Kim and Chun, 2003). In terms of PMP, the literature reveals that a negative association between age and relativism exists, whereas the findings for idealism are inconsistent (D. R. Forsyth, 1980; Ho et al., 1997; Vitell et al., 1991). In summary, it is expected that AV users with more driving experience (both direct and indirect) would be more sensitive to ethical transgressions and provide more suitable moral solutions to novel AV moral dilemma scenarios. Another expectation is that older drivers are more likely to be married and have children of their own than younger drivers, which would impact their commitment to producing outcomes that are more desirable for the overall society (e.g., protect adults who might be parents of children, protect children).

Link 2b-4: Driving experience will have an impact on PMP.

-More experienced AV users will be more idealistic than less experienced AV users
-More experienced AV users will be less relativistic than less experienced AV users

The Intuitive Moral Judgment During AV Moral Dilemma

More researchers emphasize the nonrationalist approach by including intuition and/or emotion in the moral reasoning process (Haidt, 2001; Saltzstein and Kasachkoff, 2004; Cushman et al., 2006; Sonenshein, 2007; Ruedy et al., 2013; Dedeke, 2015; Schwartz, 2016). As unexpected hazards threaten the lives of traffic users during an AV moral dilemma, intuition and/or emotion is expected to be an important factor that impacts the moral judgment stage. Moreover, intuitive moral reasoning is the response to the individual's framed moral issue. Thus, intuitive moral reasoning mediates the issue framing stage and the rational moral judgment stage. The "Integrated AV ethical decision-making framework" suggests that both intuitive and cognitive reasoning take place, thus supporting the dual-process theory of ethical decision-making (Haidt, 2001; Dane and Pratt, 2007). This section explains the intuitive moral reasoning process. We propose the following proposition.

Link 3: The intuitive moral judgment stage mediates the relationship between the AV moral dilemma issue framing stage and the rational moral judgment stage.

Moral Intuition

Moral intuiting is a non-conscious cognitive process that occurs quickly and effortlessly Kahneman (2003), Evans (2008) when an individual perceives a morally salient context (Haidt, 2001; Reynolds, 2006; Schwartz, 2016). The dual-process theory explains that intuitive moral reasoning occurs automatically and effortlessly prior to slow and effortful moral reasoning (Greene et al., 2001; Haidt, 2001; Haidt and Joseph, 2004; Greene, 2007; Greene, 2009). However, there is a limitation of this theory. The dual-process theory interprets emotional processes as fast and unconscious, which oversimplifies the moral reasoning process and may neglect the possibility of conscious decision-making (Christensen and Sutton, 2012). Moreover, studies show that people make automatic and unconscious cognitive judgments based on their prior experiences (Greenwald and Farnham, 2000; Bargh et al., 2001; Dedeke, 2015). Consequently, this study does not distinguish intuitive processes as automatic and unconscious and cognitive moral reasoning as slow and conscious but acknowledges that both intuition and cognition can automatically occur during moral reasoning. In line with the previous findings, this framework expects that AV occupants who have not experienced AV accidents can automatically and effortlessly make both intuitive and cognitive responses during the AV moral dilemma because people have intuition and have preliminary moral knowledge in vehicle accidents that can be

extended to AV moral dilemma scenarios. In other words, when an AV user faces an AV moral dilemma, a reflexive pattern-matching process may be unconsciously started, and the best prototype that matches the novel context that also matches the user's values will be more acceptable or understandable for the user.

Moral Emotions

Moral emotion has been explicitly included in ethical decision-making (Gaudine and Thorne, 2001; Salvador and Folger, 2009). The following is a categorization of moral emotions that suggest direct relations to ethical decision-making (Eisenberg, 2000; Tangney et al., 2007), which can also be found during the AV moral dilemma: 1) "Prosocial" Moral emotions (e.g., empathy, sympathy, concern, or compassion). Prosocial behaviors such as providing support or help had a link between sympathy (e.g., Carlo et al., 2011), and compassion is activated when the suffering of others is viewed, which leads to altruistic moral actions (Goetz et al., 2010), 2) "Self-Conscious" Moral Emotions (e.g., guilt, shame, embarrassment). Emotions in this category are "evoked by self-reflection and self-evaluation" (Tangney et al., 2007, p. 347). Feeling guilt results from recognizing how the other party has been wronged, and thus leads to empathetic behaviors (de Hooge et al., 2007). 3) "Other-blame" Moral emotions (e.g., contempt, anger, and disgust). People who feel anger tend to attribute blame to others, thus aggregating aggressive behaviors (Dix et al., 1990; Keltner et al., 1993), because anger is often related to justice or fairness (Goldman, 2003). In addition, in a study that explored dual-process reasoning during the AV moral dilemma, moral emotions or related moral value codes in the context of AV ethics that fall into these categories were found (e.g., empathy, conscience, self-sacrifice, children-preservation, kin preservation, passenger preservation, fault liability of self, anger, and fault liability of others) (Rhim et al., 2020). Although AV accidents are a new phenomenon, moral emotions or emotional reactions toward a novel context will allow people to determine what is ethical or not. Therefore, it is crucial to include emotion as a mental process of moral reasoning during the AV moral dilemma. In summary, this framework adopts that individuals will respond to novel AV moral dilemma contexts depending on their emotional responses (Sonenshein, 2007; Dedeke, 2015).

Moral Reflection

In conventional crashes, moral reflection would rarely occur since most crash avoidance behaviors are reflexive actions without moral judgment. In contrast, when developing moral behaviors of AVs, the inclusion of the moral reflection stage is possible, which provides the opportunity to reflect upon contexts to minimize conflict that could occur (e.g., consequences vs. fairness). Dedeke (2015) explains that moral reflection focuses on the factual review process, and the role of moral reflection becomes more important when situations involve strong automatic responses, both emotional and cognitive. Thus, the following questions can be asked to reduce bias and minimize immediate reactions based on reflexive judgment. "Do I have all the facts to make my conclusion? Am I interpreting the facts in the correct way? Am I using the correct frame of reference?" (Dedeke, 2015, p.447). In this regard, the moral reflection stage during an AV moral dilemma will promote more accurate processing of information leading to more acceptable decisions for overall society.

Moral reflection occurs after reviewing facts that would occur during a moral dilemma (e.g., what will be the consequences of each decision? Whose liability will it be? What would be the fairest decision?). Reidenbach and Robin (1990) specified dimensions of moral reflection: The relativistic dimension evaluates whether a decision is traditionally acceptable or not and whether it is culturally appropriate or not. Further, the contractualism dimension evaluates whether unspoken promises or unwritten contracts are violated or not. These dimensions are derived from moral philosophies (Reidenbach and Robin, 1990). The relativistic and contractualism dimensions can be referenced in the AV moral reflection stage to induce more ethical and socially acceptable AV decisions. For example, one of the AV moral dilemma scenarios includes "Comply with road traffic laws which results in maximized overall harms" (Rhim et al., 2020, p. 44). An initial automatic intuition would perceive that following traffic rules is ethical. However, if the consequences result in multiple fatalities, the decision may not be ethical nor socially acceptable. As AVs can be preprogrammed, various consequences and reflections should be included in the algorithms. In summary, based on previous studies, this framework emphasizes the role of the moral rationalization process Dedeke (2015), Schwartz (2016), especially after reflexive moral reasoning, because reasoning that occurred quickly may not consider the full spectrum of the problem (Sonenshein, 2007).

Rational Moral Reasoning During AV Moral Dilemma

This study includes varying factors that impact intuitive and rational moral judgment either directly or potentially in ethical decision-making to explain the dual-process theory in the AV moral dilemma. How rational moral reasoning is shaped and impacts ethical intention will be explained in this section.

Rational Moral Judgment

In accordance with Dedeke's (2015) cognitive-intuitionist model, this framework provides an explanation of pluralistic moral reasoning judgment patterns. First, moral judgment could be mainly based on an AV user's intuitive reaction toward the framed moral issue. Second, moral judgment could be established mostly on rational judgment, in which intuition is less evoked. Third, a moral judgment could rely on both intuition and rational reasoning. In this case, the automatic reasoning process is the basis for moral reflection and rational reasoning process. In the AV moral dilemma, if one is directly impacted or involved in the AV accident, moral intuition would be more likely to be activated. For instance, if an AV user feels compassion toward pedestrians during an AV accident, he or she will tend to make moral judgments that could preserve pedestrians over other involved traffic users. Or if the decision-maker is a bystander of

an AV accident who is not impacted by the accidents, moral emotion would be less significant, and the rational reasoning process will become more dominant. For this reason, intuitive moral reasoning impacts the rational moral judgment process. Moreover, how the decision-maker frames the moral issue impacts the moral reasoning process (Dedeke, 2015). Moral issue frames can explain why people prefer utilitarian AVs, but do not want to buy such AVs. Utilitarian AVs, which intend to save the most lives, seem ethical from the observer's perspective. However, if the decision is made from the first-person perspective, there is a possibility that the decision-maker can be sacrificed to reduce overall harm. In other words, the moral judgment stage is impacted by how the specific AV moral dilemma is framed by an individual, which is impacted by PMI. Hence, the following propositions are developed:

Link 4a: Intuitive moral judgment processes impact rational moral judgment processes.

Link 4b: PMI impacts rational moral judgment. For stronger PMI, an AV user will face a more challenging moral reasoning process.

Ethical Behavioral Intent in AV Moral Dilemma

An AV user's contemplation in the moral judgment stage, whether intuitive, rational, or both, leads to the individual's intention to make either ethical or unethical behaviors during an AV moral dilemma. Researchers agree that emotions impact ethical decision-making. Bagozzi and Pieters (1998) explained that different emotions have discrete goals, thus leading to different behaviors. Moreover, different emotions lead to different moral actions or ethical behavioral intent (EBI) (Blasi, 1999). For instance, the empathy-altruism hypothesis explains that empathy evokes emotions of concern to others who are suffering, which is the driving motivation of altruistic or prosocial behaviors (Batson et al., 1988; Persson and Kajonius, 2016). Similar findings were found in AV moral dilemmas. The dominant moral emotions found for "Moral Altruist" were guilt and empathy. People in this group tend to make decisions that emphasize the safety of overall traffic users, including protecting negligent drivers (Rhim et al., 2020). In the case when cognition is more activated when making EBI, an individual will compare possible actions based on his or her moral principles Bastons (2008) and try to prioritize certain moral values over others to determine moral consequences (Melé, 2005; Craft, 2013). When applied to the AV moral dilemma, an individual's rational behavior intention would be to minimize overall harm, consider liability, follow road traffic rules, distribute responsibility, or protect a certain party (e.g., cyclists, pedestrians, and passengers in AV). In summary, this study postulates that understanding the impact of both intuition and cognition will provide a more concrete understanding of the connection between moral judgment and moral EBI. Hence, the following propositions are developed:

Link 5a: The intuitive moral judgment stage impacts EBI during the AV moral dilemma.

Link 5b: The rational moral judgment stage impacts EBI during the AV moral dilemma.

DISCUSSION

This study illustrates an "Integrative ethical decision-making framework for the AV moral dilemma" to provide an alternative perspective to the conventional trolley problem-based AV ethics. This framework fills in research gaps by explaining pluralistic nature of AV ethical decision-making patterns that reflect the public's perspectives, which in turn advances social value embedded AV ethics.

The following is the theoretical implication of this study. While many researchers agree with the need for an AV ethics framework to provide explanations of ethical behaviors of AVs, the existing models show only a limited aspect of AV moral reasoning. The "Integrated AV ethical decision-making framework" is one of the first models that provides a comprehensive explanation of the full ethical decision-making process by defining various variables related to the AV moral dilemma. The relationships among the constructs show the step-by-step ethical intention shaping process, which includes both intuitive and cognitive moral reasoning processes. Moreover, the detailed examples and propositions provided in this study overcome the limitation of studies adopting scenario-based methodologies. For instance, understanding the moral issue framing stage may aid in minimizing preconstructed interpretations in the scenarios (e.g., locus of control impacts moral judgment). Therefore, the framework in this study allows consideration of multiple aspects of the AV moral dilemma to discuss realistic AV ethics.

The social contributions of the study are as follows. First, a social value embedded AV ethics framework will provide explainable and transparent AV ethics for prospective users. Singhapakdi et al. (1999) explain that individuals could select ethically questionable decisions simply because they are unfamiliar with the moral issue. Similar trends can be found in AV moral dilemmas because not many people have experienced the novel context of AV involved crashes. Hence, AV instructions based on the framework may help potential users recognize frequently occurring morally salient situations. Moreover, clarification of which ethical decisions of AVs may be more appropriate is likely to enhance recognition of AV crashes with moral saliency and ultimately lead to less unethical AV crash selections. Second, regulators could develop more realistic AV ethical frameworks by considering alternatives to trolley problem-based ethics. Researchers advise that vague AV guidelines should be avoided (De Freitas et al., 2020b). Further, it is widely accepted that regulations are difficult to modify once implemented. Therefore, it is crucial to develop acceptable AVs in the first place. Consequently, establishing realistic and transparent AV ethics would facilitate communication with the public, which will, in turn, increase trust in AV systems. Ultimately, this will prepare the overall society to build socially acceptable AVs.

The following are the technological implications of this study. First, the model offers an alternative perspective to the trolley problem-based AV ethics, which often assumes one moral theory, such as utilitarianism. The propositions provided in this study bring to light that assumptions of ethical behaviors of AVs should

be reevaluated (e.g., different cultures will prefer different AV ethical behaviors). Toward addressing this issue, researchers have recently modeled three AV ethical decision-making algorithms (contractarian, utilitarian, and egalitarian) based on a Markov Decision Process (MDP) to react when moral dilemma situation is detected (De Moura et al., 2020). Although the AV decisions from the MDP provide an implementation of pluralistic AV moral behaviors, this model does not consider the intuitive aspect of users. Second, while it might not be feasible to directly program intuitions into AV algorithms, considering moral emotions and the intuiting process that occurs during the AV moral dilemma may enhance prospective users' acceptance and interpretation of AVs, as well as provide inspirations for engineers. For instance, current AVs are typically programmed with opaque, deep neural networks for fast, low-level processing, along with transparent conditional logic for high-level decision-making (Karpathy, 2020). The level at which to separate these two systems is still an active research topic, including the exploration of completely end-to-end System 1 approaches using reinforcement learning (Kuutii et al., 2020). An analysis of System 1 and System 2 in human ethical decision-making may be a way forward in designing systems that balance effectiveness and explanatory power. Third, human-centered AI (HCAI) provides clear goals to achieve reliable, safe, and trustworthy AI-embedded systems Shneiderman (2020), yet how to achieve these goals is unclear. The variables used in this study such as individual and cultural factors, perceived moral intensity, and possible decision-making patterns can aid engineers in considering machine translatable ethical AV behaviors. For example, in creating AV systems that may be deployed worldwide to different countries, AV developers could integrate tweakable parameters based on situationist vs. exceptionist differences, such as the ability to transgress rules of the road depending on the consequence to the group. As another example, surveys of AV users can be interpreted through the lens of individual factors such as education, age, and their expected moral responses, rather than taken as a whole.

The proposed "Integrative ethical decision-making framework for the AV moral dilemma" is not free of limitations. First, the framework is conceptual and suggests propositions that are not empirically tested. The detailed moral preferences cannot be measured. Future studies could empirically validate the framework Preferences for Precepts Implied in Moral Theories (PPIMT) instrument Dubljević et al. (2018), which "assess respondents' preference for the precepts implied in the three dominant moral theories" Dubljević (2020), can be used for empirical validation of AV users' moral judgment tendencies. Measuring PPIMT will provide a more concrete understanding of how the AV moral dilemma context activates users' preference of a specific ethical theory. Second, this model focused mainly on an individual AV user's moral judgment. However, AVs will be deployed in mixed traffic scenarios where multiple traffic users are involved (e.g., other AVs, conventional cars, pedestrians, passengers, and cyclists) (Nyholm and Smids, 2018; Ranasinghe et al., 2020). The framework or theory can be expanded to describe the interrelationship between multiple traffic users to understand accountable AV moral reasoning in a broader sense. A future study can reference the "Integrated AV ethical decision-making model" when developing social values embedded algorithms and user interfaces. Finally, while this study focused specifically on AV morality, AI-embedded technologies such as social robots will face similar moral conundrums. In the future, this framework may be extended to other related fields to provide a foundational theory to strengthen the field of AI ethics and roboethics.

CONCLUSION

This study attempts to fill in research gaps that appear in the existing AV ethics models by providing a comprehensive theoretical framework. It does so by defining key AV moral dilemma-related factors and merging them together into an integrative framework that includes both the intuitive and cognitive moral reasoning processes. More specifically, this study explains how an individual frames the AV moral dilemma, impacted by individual characteristics and PMP, which will in turn be the reference for intuitive and cognitive moral reasoning leading to EBI. The proposed integrated framework can be considered to reflect the "person-situation" interactionist perspective Trevino (1986) as well as the "cognitive-intuitionist" approach (Dedeke, 2015). Consequently, the framework embeds the dual-process theory and provides explanations for moral pluralism of AV ethics that includes the intuitive moral reasoning.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

JR is responsible for developing the framework and writing of the manuscript. AL provided key insights and completed multiple revisions. MC and JL provided valuable advice.

FUNDING

This work was supported under the Huawei-SFU Joint Lab Project (R569337).

ACKNOWLEDGMENTS

This article is an extension of one chapter in JR's doctoral thesis at the Korea Advanced Institute of Science and Technology (KAIST). The authors are grateful to Namwoo Kang, Wonjae Lee, Moon Choi, and Jeongmi Lee, who provided comments and insights.

REFERENCES

Ajzen, I. (2018). The theory of planned behavior. *Organ. Behav. Hum. Decis. Process.* 50, 179–211. doi:10.1016/0749-5978(91)90020-T

Allen, C., Smit, I., and Wallach, W. (2005). Artificial Morality: Top-Down, Bottom-Up, and Hybrid Approaches. *Ethics Inf. Technol.* 7 (3), 149–155. doi:10.1007/s10676-006-0004-4

Anderson, J. M., Kalra, N., Stanley, K. D., Sorensen, P., Samaras, C., and Oluwatola, O. A. (2014). *Autonomous Vehicle Technology: A Guide for Policymakers*. Rand Corporation. doi:10.7249/rb9755

Awad, E., Dsouza, S., Bonnefon, J.-F., Shariff, A., and Rahwan, I. (2020). Crowdsourcing Moral Machines. *Commun. ACM* 63 (3), 48–55. doi:10.1145/3339904

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., et al. (2018a). The Moral Machine Experiment. *Nature* 563 (7729), 59–64. doi:10.1038/s41586-018-0637-6

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., et al. (2018b). *Blaming Humans in Autonomous Vehicle Accidents: Shared Responsibility across Levels of Automation*. ArXiv:1803.07170 [Cs]. http://arxiv.org/abs/1803.07170.

Bagozzi, R. P., and Pieters, R. (1998). Goal-directed Emotions. *Cogn. Emot.* 12 (1), 1–26. doi:10.1080/026999398379754

Bansal, P., and Kockelman, K. M. (2017). Forecasting Americans' Long-Term Adoption of Connected and Autonomous Vehicle Technologies. *Transportation Res. A: Pol. Pract.* 95, 49–63. doi:10.1016/j.tra.2016.10.013

Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., and Trötschel, R. (2001). The Automated Will: Nonconscious Activation and Pursuit of Behavioral Goals. *J. Personal. Soc. Psychol.* 81 (6), 1014–1027. doi:10.1037/0022-3514.81.6.1014

Bastons, M. (2008). The Role of Virtues in the Framing of Decisions. *J. Bus Ethics* 78 (3), 389–400. doi:10.1007/s10551-006-9332-x

Batson, C. D., Dyck, J. L., Brandt, J. R., Batson, J. G., Powell, A. L., McMaster, M. R., et al. (1988). Five Studies Testing Two New Egoistic Alternatives to the Empathy-Altruism Hypothesis. *J. Personal. Soc. Psychol.* 55 (1), 52–77. doi:10.1037/0022-3514.55.1.52

Bellet, T., Hoc, J. M., Boverie, S., and Boy, G. A. (2011). "From human-machine interaction to cooperation: Towards the integrated copilot," in *Human-Computer Interaction in Transport*. Editors T. Kukutai and J. Taylor (Farnham, United Kingdom: Ashgate), 129–156.

Beu, D., and Buckley, M. R. (2001). The Hypothesized Relationship between Accountability and Ethical Behavior. *J. Business Ethics* 1 (34), 57–73. doi:10.1023/a:1011957832141

Bigman, Y. E., and Gray, K. (2020). Life and Death Decisions of Autonomous Vehicles. *Nature* 579 (7797), E1–E2. doi:10.1038/s41586-020-1987-4

Blasi, A. (1999). Emotions and Moral Motivation. *J. Theor. Soc. Behav.* 29 (1), 1–19. doi:10.1111/1468-5914.00088

Bonnefon, J.-F., Shariff, A., and Rahwan, I. (2016). The Social Dilemma of Autonomous Vehicles. *Science* 352 (6293), 1573–1576. doi:10.1126/science.aaf2654

Bonnemains, V., Saurel, C., and Tessier, C. (2018). Embedded Ethics: Some Technical and Ethical Challenges. *Ethics Inf. Technol.* 20 (1), 41–58. doi:10.1007/s10676-018-9444-x

Borenstein, J., Herkert, J., and Miller, K. (2019). Autonomous Vehicles and the Ethical Tension between Occupant and Non-occupant Safety. *Comput. Ethics - Philos. Enquiry (Cepe) Proc.* 2019 (1).

Carlo, G., Mestre, M. V., Samper, P., Tur, A., and Armenta, B. E. (2011). The Longitudinal Relations Among Dimensions of Parenting Styles, Sympathy, Prosocial Moral Reasoning, and Prosocial Behaviors. *Int. J. Behav. Dev.* 35 (2), 116–124. doi:10.1177/0165025410375921

Carlson, D. S., Kacmar, K. M., and Wadsworth, L. L. (2009). The Impact of Moral Intensity Dimensions on Ethical Decision-Making: Assessing the Relevance of Orientation. *J. Managerial Issues* 21 (4), 534–551.

Carsten, P., Andel, T. R., Yampolskiy, M., and McDonald, J. T. (2015). In-vehicle Networks: Attacks, Vulnerabilities, and Proposed Solutions. In Proceedings of the 10th Annual Cyber and Information Security Research Conference. 1–8.

Castelvecchi, D. (2016). Can We Open the Black Box of AI? *Nature* 538 (7623), 20–23. doi:10.1038/538020a

Cervantes, J.-A., López, S., Rodríguez, L.-F., Cervantes, S., Cervantes, F., and Ramos, F. (2020). Artificial Moral Agents: A Survey of the Current Status. *Sci. Eng. Ethics* 26 (2), 501–532. doi:10.1007/s11948-019-00151-x

Chatzidakis, A., Shaw, D., and Allen, M. (2018). A psycho-social approach to consumer ethics. *J. Consum. Cult.*. doi:10.1177/1469540518773815

Cherry, J. O. H. N., and Caldwell, J. A. M. E. S. (2013). Searching for the Origins of Consumer Ethics: Bridging the Gap between Intuitive Values and Consumer Ethical Judgments. *Marketing Manag. J.* 23 (2), 117–133.

Chowdhury, R. M. M. I. (2017). Emotional Intelligence and Consumer Ethics: The Mediating Role of Personal Moral Philosophies. *J. Bus Ethics* 142 (3), 527–548. doi:10.1007/s10551-015-2733-y

Christensen, W., and Sutton, J. (2012). *Reflections on Emotions, Imagination, and Moral Reasoning toward an Integrated, Multidisciplinary Approach to Moral Cognition*. Abingdon, Oxon, United Kingdom: Emotions, imagination, and moral reasoning, 327–347.

Cole, B. C., and Smith, D. L. (1996). Perceptions of Business Ethics: Students vs. Business People. *J. Bus Ethics* 15 (8), 889–896. doi:10.1007/bf00381856

Craft, J. L. (2013). A Review of the Empirical Ethical Decision-Making Literature: 2004–2011. *J. Business Ethics* 2 (117), 221–259. doi:10.1007/s10551-012-1518-9

Cushman, F., Young, L., and Hauser, M. (2006). The Role of Conscious Reasoning and Intuition in Moral Judgment. *Psychol. Sci.* 17 (12), 1082–1089. doi:10.1111/j.1467-9280.2006.01834.x

Dane, E., and Pratt, M. G. (2007). Exploring Intuition and its Role in Managerial Decision Making. *Amr* 32 (1), 33–54. doi:10.5465/amr.2007.23463682

Danielson, P. (2015). Surprising Judgments about Robot Drivers: Experiments on Rising Expectations and Blaming Humans. *Etikk i Praksis-Nordic J. Appl. Ethics* 1, 73–86. doi:10.5324/eip.v9i1.1727

De Freitas, J., Cikara, M., and Cikara, M. (2021). Deliberately Prejudiced Self-Driving Vehicles Elicit the Most Outrage. *Cognition* 208, 104555. doi:10.1016/j.cognition.2020.104555

De Freitas, J., Anthony, S. E., Censi, A., and Alvarez, G. A. (2020a). Doubting Driverless Dilemmas. *Perspect. Psychol. Sci.* 15 (5), 1284–1288. doi:10.1177/1745691620922201

De Freitas, J., Censi, A., Di Lillo, L., Anthony, S. E., and Frazzoli, E. (2020b). *From Driverless Dilemmas to More Practical Ethics Tests for Autonomous Vehicles*.

de Hooge, I. E., Zeelenberg, M., and Breugelmans, S. M. (2007). Moral Sentiments and Cooperation: Differential Influences of Shame and Guilt. *Cogn. Emot.* 21 (5), 1025–1042. doi:10.1080/02699930600980874

De Moura, N., Chatila, R., Evans, K., Chauvier, S., and Dogan, E. (2020). *Ethical Decision Making for Autonomous Vehicles*. Las Vegas, NV, USA: IEEE Intelligent Vehicles Symposium (IV), 2006–2013. doi:10.1109/IV47402.2020.9304618Ethical Decision Making for Autonomous Vehicles

Dedeke, A. (2015). A Cognitive-Intuitionist Model of Moral Judgment. *J. Bus Ethics* 126 (3), 437–457. doi:10.1007/s10551-013-1965-y

Deshpande, S. P. (1997). Managers' Perception of Proper Ethical Conduct: The Effect of Sex, Age, and Level of Education. *J. Business Ethics* 16 (1), 79–85. doi:10.1023/a:1017917420433

Dix, T., Reinhold, D. P., and Zambarano, R. J. (1990). Mothers' Judgment in Moments of Anger. *Merrill-Palmer Q.* 36 (4), 465–486.

Dogan, E., Chatila, R., Chauvier, S., Evans, K., Hadjixenophontos, P., and Perrin, J. (2016). "Ethics in the Design of Automated Vehicles: The AVEthics Project," in *EDIA@ ECAI*, 10–13.

Dubinsky, A. J., and Loken, B. (1989). Analyzing Ethical Decision Making in Marketing. *J. Business Res.* 19 (2), 83–107. doi:10.1016/0148-2963(89)90001-5

Dubljević, V., Sattler, S., and Racine, E. (2018). Deciphering Moral Intuition: How Agents, Deeds, and Consequences Influence Moral Judgment. *PloS one* 13 (10), e0204631. doi:10.1371/journal.pone.0204631

Dubljević, V. (2020). Toward Implementing the ADC Model of Moral Judgment in Autonomous Vehicles. *Sci. Eng. Ethics* 26 (5), 2461–2472. doi:10.1007/s11948-020-00242-0

Dubljević, V., and Racine, E. (2014). The ADC of Moral Judgment: Opening the Black Box of Moral Intuitions with Heuristics about Agents, Deeds, and Consequences. *AJOB Neurosci.* 5 (4), 3–20. doi:10.1080/21507740.2014.939381

Eisenberg, N. (2000). Emotion, Regulation, and Moral Development. *Annu. Rev. Psychol.* 51 (1), 665–697. doi:10.1146/annurev.psych.51.1.665

Eizenberg, E., and Jabareen, Y. (2017). Social Sustainability: A New Conceptual Framework. *Sustainability* 9 (1), 68. doi:10.3390/su9010068

Evans, J. S. B. T. (2008). Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annu. Rev. Psychol.* 59 (1), 255–278. doi:10.1146/annurev.psych.59.103006.093629

Fagnant, D. J., and Kockelman, K. (2015). Preparing a Nation for Autonomous Vehicles: Opportunities, Barriers and Policy Recommendations. *Transportation Res. Part A: Pol. Pract.* 77, 167–181. doi:10.1016/j.tra.2015.04.003

Ferrell, O. C., and Gresham, L. G. (1985). A Contingency Framework for Understanding Ethical Decision Making in Marketing. *J. Marketing* 49 (3), 87–96. doi:10.1177/002224298504900308

Ferrell, O. C., LeClair, D. T., and Ferrell, L. (1998). The Federal Sentencing Guidelines for Organizations: A Framework for Ethical Compliance. *J. Business Ethics* 17 (4), 353–363. doi:10.1023/a:1005786809479

Forsyth, D. R. (1980). A Taxonomy of Ethical Ideologies. *J. Personal. Soc. Psychol.* 39 (1), 175–184. doi:10.1037/0022-3514.39.1.175

Forsyth, D. R. (1985). Individual Differences in Information Integration during Moral Judgment. *J. Personal. Soc. Psychol.* 49 (1), 264–272. doi:10.1037/0022-3514.49.1.264

Forsyth, D. R. (1992). Judging the Morality of Business Practices: The Influence of Personal Moral Philosophies. *J. Bus Ethics* 11 (5–6), 461–470. doi:10.1007/bf00870557

Forsyth, D. R., Nye, J. L., and Kelley, K. (1988). Idealism, Relativism, and the Ethic of Caring. *J. Psychol.* 122 (3), 243–248. doi:10.1080/00223980.1988.9915511

Forsyth, D. R., O'Boyle, E. H., and McDaniel, M. A. (2008). East Meets West: A Meta-Analytic Investigation of Cultural Variations in Idealism and Relativism. *J. Bus Ethics* 83 (4), 813–833. doi:10.1007/s10551-008-9667-6

Forsyth, D. R., and Pope, W. R. (1984). Ethical Ideology and Judgments of Social Psychological Research: Multidimensional Analysis. *J. Personal. Soc. Psychol.* 46 (6), 1365–1375. doi:10.1037/0022-3514.46.6.1365

Forte, A. (2005). Locus of Control and the Moral Reasoning of Managers. *J. Business Ethics* 1 (58), 65–77. doi:10.1007/s10551-005-1387-6

Gaudine, A., and Thorne, L. (2001). Emotion and Ethical Decision-Making in Organizations. *J. Business Ethics* 31 (2), 175–187. doi:10.1023/a:1010711413444

Gawronski, B., and Beer, J. S. (2017). What Makes Moral Dilemma Judgments "Utilitarian" or "Deontological"? *Soc. Neurosci.* 12 (6), 626–632. doi:10.1080/17470919.2016.1248787

Goetz, J. L., Keltner, D., and Simon-Thomas, E. (2010). Compassion: An Evolutionary Analysis and Empirical Review. *Psychol. Bull.* 136 (3), 351–374. doi:10.1037/a0018807

Gogoll, J., and Müller, J. F. (2017). Autonomous Cars: In Favor of a Mandatory Ethics Setting. *Sci. Eng. Ethics* 23 (3), 681–700. doi:10.1007/s11948-016-9806-x

Goldman, B. M. (2003). The Application of Referent Cognitions Theory to Legal-Claiming by Terminated Workers: The Role of Organizational Justice and Anger. *J. Manag.* 29 (5), 705–728. doi:10.1016/s0149-2063_03_00032-1

Goodall, N. J. (2014b). Ethical Decision Making during Automated Vehicle Crashes. *Transportation Res. Rec.* 2424 (1), 58–65. doi:10.3141/2424-07

Goodall, N. J. (2014a). "Machine Ethics and Automated Vehicles," in *Road Vehicle Automation* (Springer), 93–102. doi:10.1007/978-3-319-05990-7_9

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., et al. (2013). "Moral Foundations Theory," in *Advances in Experimental Social Psychology* (Academic Press), 47, 55–130. doi:10.1016/b978-0-12-407236-7.00002-4

Greene, J. D. (2009). Dual-process Morality and the Personal/impersonal Distinction: A Reply to McGuire, Langdon, Coltheart, and Mackenzie. *J. Exp. Soc. Psychol.* 45 (3), 581–584. doi:10.1016/j.jesp.2009.01.003

Greene, J. D. (2016). Our Driverless Dilemma. *Science* 352 (6293), 1514–1515. doi:10.1126/science.aaf9534

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. (2001). An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science* 293 (5537), 2105–2108. doi:10.1126/science.1062872

Greene, J. D. (2007). Why Are VMPFC Patients More Utilitarian? A Dual-Process Theory of Moral Judgment Explains. *Trends Cogn. Sci.* 11 (8), 322–323. doi:10.1016/j.tics.2007.06.004

Greenwald, A. G., and Farnham, S. D. (2000). Using the Implicit Association Test to Measure Self-Esteem and Self-Concept. *J. Personal. Soc. Psychol.* 79 (6), 1022–1038. doi:10.1037/0022-3514.79.6.1022

Groves, K., Vance, C., and Paik, Y. (2008). Linking Linear/Nonlinear Thinking Style Balance and Managerial Ethical Decision-Making. *J. Bus Ethics* 80 (2), 305–325. doi:10.1007/s10551-007-9422-4

Guo, J., Kurup, U., and Shah, M. (2018). *Is it Safe to Drive? An Overview of Factors, Challenges, and Datasets for Driveability Assessment in Autonomous Driving*. ArXiv Preprint ArXiv:1811.11277.

Haidt, J., and Joseph, C. (2004). Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues. *Daedalus* 133 (4), 55–66. doi:10.1162/0011526042365555

Haidt, J. (2001). The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychol. Rev.* 108 (4), 814–834. doi:10.1037/0033-295x.108.4.814

Hevelke, A., and Nida-Rümelin, J. (2015). Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis. *Sci. Eng. Ethics* 21 (3), 619–630. doi:10.1007/s11948-014-9565-5

Ho, F. N., Vitell, S. J., Barnes, J. H., and Desborde, R. (1997). Ethical Correlates of Role Conflict and Ambiguity in Marketing: The Mediating Role of Cognitive Moral Development. *J. Acad. Mark. Sci.* 25 (2), 117–126. doi:10.1007/bf02894347

Hofstede, G. H., Hofstede, G. J., and Minkov, M. (2005). *Cultures and Organizations: Software of the Mind*, 2. New York: McGraw-Hill.

Hofstede, G., and McCrae, R. R. (2004). Personality and Culture Revisited: Linking Traits and Dimensions of Culture. *Cross-cultural Res.* 38 (1), 52–88. doi:10.1177/1069397103259443

Hulse, L. M., Xie, H., and Galea, E. R. (2018). Perceptions of Autonomous Vehicles: Relationships with Road Users, Risk, Gender and Age. *Saf. Sci.* 102, 1–13. doi:10.1016/j.ssci.2017.10.001

Hunt, S. D., and Vitell, S. (1986). A General Theory of Marketing Ethics. *J. Macromarketing* 6 (1), 5–16. doi:10.1177/027614678600600103

Hunt, S. D., and Vitell, S. J. (2006). The General Theory of Marketing Ethics: A Revision and Three Questions. *J. Macromarketing* 26 (2), 143–153. doi:10.1177/0276146706290923

Hunt, T. G., and Jennings, D. F. (1997). Ethics and Performance: A Simulation Analysis of Team Decision Making. *J. Business Ethics* 16 (2), 195–203. doi:10.1023/a:1017987224590

Imenda, S. (2014). Is There a Conceptual Difference between Theoretical and Conceptual Frameworks? *J. Soc. Sci.* 38 (2), 185–195. doi:10.1080/09718923.2014.11893249

Inglehart, R., and Baker, W. E. (2000). Modernization, Cultural Change, and the Persistence of Traditional Values. *Am. Sociological Rev.* 65 (1), 19–51. doi:10.2307/2657288

Jabareen, Y. (2009). Building a Conceptual Framework: Philosophy, Definitions, and Procedure. *Int. J. Qual. Methods* 8 (4), 49–62. doi:10.1177/160940690900800406

Jones, T. M. (1991). Ethical Decision Making by Individuals in Organizations: An Issue-Contingent Model. *Acad. Manag. Rev.* 16 (2), 366. doi:10.2307/258867

Kahneman, D. (2003). A Perspective on Judgment and Choice: Mapping Bounded Rationality. *Am. Psychol.* 58 (9), 697–720. doi:10.1037/0003-066x.58.9.697

Kallioinen, N., Pershina, M., Zeiser, J., Nosrat Nezami, F., Stephan, A., Pipa, G., et al. (2019). Moral Judgements on the Actions of Self-Driving Cars and Human Drivers in Dilemma Situations from Different Perspectives. *Front. Psychol.* 10, 2415. doi:10.3389/fpsyg.2019.02415

Karcher, J. N. (1996). Auditors' Ability to Discern the Presence of Ethical Problems. *J. Bus Ethics* 15 (10), 1033–1050. doi:10.1007/bf00412045

Karnouskos, S. (2020a). Self-Driving Car Acceptance and the Role of Ethics. *IEEE Trans. Eng. Manage.* 67 (2), 252–265. doi:10.1109/tem.2018.2877307

Karnouskos, S. (2020b). *The Role of Utilitarianism, Self-Safety, and Technology in the Acceptance of Self-Driving Cars*. London, United Kingdom: Cognition, Technology & Work. doi:10.1007/s10111-020-00649-6

Karpathy, A. (2020). *U.S. Patent No. WO/2020/056331*. Washington, DC: U.S. Patent and Trademark Office.

Keltner, D., Ellsworth, P. C., and Edwards, K. (1993). Beyond Simple Pessimism: Effects of Sadness and Anger on Social Perception. *J. Personal. Soc. Psychol.* 64 (5), 740–752. doi:10.1037/0022-3514.64.5.740

Kochupillai, M., Lütge, C., and Poszler, F. (2020). Programming Away Human Rights and Responsibilities? "The Moral Machine Experiment" and the Need for a More "Humane" AV Future. *NanoEthics* 14 (3), 285–299. doi:10.1007/s11569-020-00374-4

Kracher, B., Chatterjee, A., and Lundquist, A. R. (2002). Factors Related to the Cognitive Moral Development of Business Students and Business Professionals

in India and the United States: Nationality, Education, Sex and Gender. *J. Business Ethics* 35 (4), 255–268. doi:10.1023/a:1013859404733

Kruegel, S., and Uhl, M. (2020). *Autonomous Vehicles and Moral Judgments under Risk*. Available at SSRN 3686613.

Kumfer, W., and Burgess, R. (2015). Investigation into the Role of Rational Ethics in Crashes of Automated Vehicles. *Transportation Res. Rec.* 2489 (1), 130–136. doi:10.3141/2489-15

Kuutti, S., Bowden, R., Jin, Y., Barber, P., and Fallah, S. (2020). *A Survey of Deep Learning Applications to Autonomous Vehicle Control*. IEEE Transactions on Intelligent Transportation Systems. doi:10.1109/icra40945.2020.9197351

Latif, D. A. (2001). The Relationship between Pharmacists' Tenure in the Community Setting and Moral Reasoning. *J. Business Ethics* 31 (2), 131–141. doi:10.1023/a:1010771103427

Leikas, J., Koivisto, R., and Gotcheva, N. (2019). Ethical Framework for Designing Autonomous Intelligent Systems. *JOItmC* 5 (1), 18. doi:10.3390/joitmc5010018

Lerner, N. D. (1993). "Brake perception-reaction times of older and younger drivers," in Proceedings of the human factors and ergonomics society annual meeting. Los Angeles, CA: SAGE Publications, Vol. 37, 206–210.

Liehr, P., and Smith, M. J. (1999). Middle Range Theory: Spinning Research and Practice to Create Knowledge for the New Millennium. *Adv. Nurs. Sci.* 21 (4), 81–91. doi:10.1097/00012272-199906000-00011

Lin, P. (2015). "Why Ethics Matters for Autonomous Cars," in *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte*. Editors M. Maurer, J. C. Gerdes, B. Lenz, and H. Winner (Springer), 69–85. doi:10.1007/978-3-662-45854-9_4

Luetge, C. (2017). The German Ethics Code for Automated and Connected Driving. *Philos. Technol.* 30 (4), 547–558. doi:10.1007/s13347-017-0284-0

Lund, D. B. (2000). An Empirical Examination of Marketing Professionals' Ethical Behavior in Differing Situations. *J. Business Ethics* 24 (4), 331–342. doi:10.1023/a:1006005823045

Malle, B. F. (2016). Integrating Robot Ethics and Machine Morality: The Study and Design of Moral Competence in Robots. *Ethics Inf. Technol.* 18 (4), 243–256. doi:10.1007/s10676-015-9367-8

May, D. R., and Pauli, K. P. (2002). The role of moral intensity in ethical decision making: A review and investigation of moral recognition, evaluation, and intention. *Bus Soc.* 41, 84–117. doi:10.1177/0007650302041001006

McGaghie, W. C., Bordage, G., and Shea, J. A. (2001). Problem Statement, Conceptual Framework, and Research Question. *Acad. Med.* 76 (9), 923–924. doi:10.1097/00001888-200109000-00021

Melé, D. (2005). Ethical Education in Accounting: Integrating Rules, Values and Virtues. *J. Business Ethics* 57 (1), 97–109. doi:10.1007/s10551-004-3829-y

Mermet, B., and Simon, G. (2016). *Formal Verication of Ethical Properties in Multiagent Systems*. Lyon, France: 1st Workshop on Ethics in the Design of Intelligent Agents.

Metcalfe, J., and Mischel, W. (1999). A Hot/cool-System Analysis of Delay of Gratification: Dynamics of Willpower. *Psychol. Rev.* 106 (1), 3–19. doi:10.1037/0033-295x.106.1.3

Mikhail, J. (2007). Universal Moral Grammar: Theory, Evidence and the Future. *Trends Cogn. Sci.* 11 (4), 143–152. doi:10.1016/j.tics.2006.12.007

Mordue, G., Yeung, A., and Wu, F. (2020). The Looming Challenges of Regulating High Level Autonomous Vehicles. *Transportation Res. Part A: Pol. Pract.* 132, 174–187. doi:10.1016/j.tra.2019.11.007

Novak, T. P. (2020). A Generalized Framework for Moral Dilemmas Involving Autonomous Vehicles: A Commentary on Gill. *J. Consumer Res.* 47 (2), 292–300. doi:10.1093/jcr/ucaa024

Nyholm, S., and Smids, J. (2018). Automated Cars Meet Human Drivers: Responsible Human-Robot Coordination and the Ethics of Mixed Traffic. *Ethics Inf. Tech.* 22 (4), 1–10. doi:10.1007/s10676-018-9445-9

Nyholm, S., and Smids, J. (2016). The Ethics of Accident-Algorithms for Self-Driving Cars: An Applied Trolley Problem?. *Ethic Theor. Moral Prac* 19 (5), 1275–1289. doi:10.1007/s10677-016-9745-2

O'Fallon, M. J., and Butterfield, K. D. (2013). "A Review of the Empirical Ethical Decision-Making Literature: 1996–2003," in *Citation Classics from the Journal of Business Ethics: Celebrating the First Thirty Years of Publication*. Editors A. C. Michalos and D. C. Poff (Springer Netherlands), 213–263.

Persson, B. N., and Kajonius, P. J. (2016). Empathy and Universal Values Explicated by the Empathy-Altruism Hypothesis. *J. Soc. Psychol.* 156 (6), 610–619. doi:10.1080/00224545.2016.1152212

Peterson, D., Rhoads, A., and Vaught, B. C. (2001). Ethical Beliefs of Business Professionals: A Study of Gender, Age and External Factors. *J. Business Ethics* 31 (3), 225–232. doi:10.1023/a:1010744927551

Pratt, C. B. (1991). PRSA Members' Perceptions of Public Relations Ethics. *Public Relations Rev.* 17 (2), 145–159. doi:10.1016/0363-8111(91)90052-m

Ranasinghe, C., Holländer, K., Currano, R., Sirkin, D., Moore, D., Schneegass, S., et al. (2020). *Autonomous Vehicle-Pedestrian Interaction across Cultures: Towards Designing Better External Human Machine Interfaces (eHMIs)*. Denver, CO, United States: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, 1–8.

Reidenbach, R. E., and Robin, D. P. (1991). A Conceptual Model of Corporate Moral Development. *J. Bus Ethics* 10 (4), 273–284. doi:10.1007/bf00382966

Reidenbach, R. E., and Robin, D. P. (1990). Toward the Development of a Multidimensional Scale for Improving Evaluations of Business Ethics. *J. Bus Ethics* 9 (8), 639–653. doi:10.1007/bf00383391

Rest, J. R. (1986). *Moral Development: Advances in Research and Theory*.

Reynolds, S. J. (2006). A Neurocognitive Model of the Ethical Decision-Making Process: Implications for Study and Practice. *J. Appl. Psychol.* 91 (4), 737–748. doi:10.1037/0021-9010.91.4.737

Rhim, J., Lee, G.-b., and Lee, J.-H. (2020). Human Moral Reasoning Types in Autonomous Vehicle Moral Dilemma: A Cross-Cultural Comparison of Korea and Canada. *Comput. Hum. Behav.* 102, 39–56. doi:10.1016/j.chb.2019.08.010

Rhodes, N., and Pivik, K. (2011). Age and Gender Differences in Risky Driving: The Roles of Positive Affect and Risk Perception. *Accid. Anal. Prev.* 43 (3), 923–931. doi:10.1016/j.aap.2010.11.015

Robin, D. P., Reidenbach, R. E., and Forrest, P. J. (1996). The Perceived Importance of an Ethical Issue as an Influence on the Ethical Decision-Making of Ad Managers. *J. Business Res.* 35 (1), 17–28. doi:10.1016/0148-2963(94)00080-8

Ruedy, N. E., Moore, C., Gino, F., and Schweitzer, M. E. (2013). The Cheater's High: The Unexpected Affective Benefits of Unethical Behavior. *J. Personal. Soc. Psychol.* 105, 531–548. doi:10.1037/a0034231

Saltzstein, H. D., and Kasachkoff, T. (2004). Haidt's Moral Intuitionist Theory: A Psychological and Philosophical Critique. *Rev. Gen. Psychol.* 8 (4), 273–282. doi:10.1037/1089-2680.8.4.273

Salvador, R., and Folger, R. G. (2009). Business Ethics and the Brain: Rommel Salvador and Robert G. Folger. *Bus Ethics Q.* 19 (1), 1–31. doi:10.5840/beq20091911

Savulescu, J., Kahane, G., and Gyngell, C. (2019). From Public Preferences to Ethical Policy. *Nat. Hum. Behav.* 3 (12), 1241–1243. doi:10.1038/s41562-019-0711-6

Schoettle, B., and Sivak, M. (2014). *Public Opinion about Self-Driving Vehicles in China, India, Japan, the U.S., the U.K., and Australia*. UMTRI-2014-30Article UMTRI-2014-30. doi:10.1109/iccve.2014.7297637

Schwartz, M. S. (2016). Ethical Decision-Making Theory: An Integrated Approach. *J. Bus Ethics* 139 (4), 755–776. doi:10.1007/s10551-015-2886-8

Shariff, A., Bonnefon, J.-F., and Rahwan, I. (2017). Psychological Roadblocks to the Adoption of Self-Driving Vehicles. *Nat. Hum. Behav.* 1 (10), 694–696. doi:10.1038/s41562-017-0202-6

Shneiderman, B. (2020). Bridging the Gap between Ethics and Practice. *ACM Trans. Interact. Intell. Syst.* 10 (4), 1–31. doi:10.1145/3419764

Singhapakdi, A., Vitell, S. J., and Franke, G. R. (1999). Antecedents, Consequences, and Mediating Effects of Perceived Moral Intensity and Personal Moral Philosophies. *J. Acad. Marketing Sci.* 27 (1), 19–36. doi:10.1177/0092070399271002

Smith, B. (2019). Personality Facets and Ethics Positions as Directives for Self-Driving Vehicles. *Tech. Soc.* 57, 115–124. doi:10.1016/j.techsoc.2018.12.006

Sonenshein, S. (2007). The Role of Construction, Intuition, and Justification in Responding to Ethical Issues at Work: The Sensemaking-Intuition Model. *Amr* 32, 1022–1040. doi:10.5465/amr.2007.26585677

Sparks, J. R., and Hunt, S. D. (1998). Marketing Researcher Ethical Sensitivity: Conceptualization, Measurement, and Exploratory Investigation. *J. Marketing* 62 (2), 92–109. doi:10.2307/1252163

Stilgoe, J., Owen, R., and Macnaghten, P. (2013). Developing a Framework for Responsible Innovation. *Res. Pol.* 42 (9), 1568–1580. doi:10.1016/j.respol.2013.05.008

Taddeo, M., and Floridi, L. (2018). How AI Can Be a Force for Good. *Science* 361 (6404), 751–752. doi:10.1126/science.aat5991

Tangney, J. P., Stuewig, J., and Mashek, D. J. (2007). Moral Emotions and Moral Behavior. *Annu. Rev. Psychol.* 58 (1), 345–372. doi:10.1146/annurev.psych.56.091103.070145

Tenbrunsel, A. E., and Messick, D. M. (1999). Sanctioning Systems, Decision Frames, and Cooperation. *Amproc* 1999 (1), C1–C6. doi:10.5465/apbpp.1999.27621841

Tenbrunsel, A. E., and Smith-Crowe, K. (2008). 13 Ethical Decision Making: Where We've Been and where We're Going. *Acad. Manag. Ann.* 2 (1), 545–607. doi:10.1080/19416520802211677

Trevino, L. K. (1986). Ethical Decision Making in Organizations: A Person-Situation Interactionist Model. *Amr* 11 (3), 601–617. doi:10.5465/amr.1986.4306235

Trevino, L. K. (1992). Moral Reasoning and Business Ethics: Implications for Research, Education, and Management. *J. Bus Ethics* 11 (5), 445–459. doi:10.1007/bf00870556

Vitell, S. J. (2003). Consumer Ethics Research: Review, Synthesis and Suggestions for the Future. *J. Business Ethics* 43 (1), 33–47. doi:10.1023/a:1022907014295

Vitell, S. J., Lumpkin, J. R., and Rawwas, M. Y. A. (1991). Consumer Ethics: An Investigation of the Ethical Beliefs of Elderly Consumers. *J. Bus Ethics* 10 (5), 365–375. doi:10.1007/bf00383238

Vitell, S. J., Rallapalli, K. C., and Singhapakdi, A. (1993). Marketing Norms: The Influence of Personal Moral Philosophies and Organizational Ethical Culture. *Jams* 21 (4), 331–337. doi:10.1007/bf02894525

Vrščaj, D., Nyholm, S., and Verbong, G. P. (2020). *Is Tomorrow's Car Appealing Today? Ethical Issues and User Attitudes beyond Automation*. Germany: AI & SOCIETY, 1–14.

Wallach, W., Franklin, S., and Allen, C. (2010). A Conceptual and Computational Model of Moral Decision Making in Human and Artificial Agents. *Top. Cogn. Sci.* 2 (3), 454–485. doi:10.1111/j.1756-8765.2010.01095.x

Wallach, W. (2010). Robot Minds and Human Ethics: The Need for a Comprehensive Model of Moral Decision Making. *Ethics Inf. Technol.* 12 (3), 243–250. doi:10.1007/s10676-010-9232-8

Webb, J., Wilson, C., and Kularatne, T. (2019). Will People Accept Shared Autonomous Electric Vehicles? A Survey before and after Receipt of the Costs and Benefits. *Econ. Anal. Pol.* 61, 118–135. doi:10.1016/j.eap.2018.12.004

Weber, J., and Gillespie, J. (1998). Differences in Ethical Beliefs, Intentions, and Behaviors. *Business Soc.* 37 (4), 447–467. doi:10.1177/000765039803700406

Williams, A. F. (2003). Teenage Drivers: Patterns of Risk. *J. Saf. Res.* 34 (1), 5–15. doi:10.1016/s0022-4375(02)00075-0

Wimalasiri, J. S., Pavri, F., and Jalil, A. A. K. (1996). An Empirical Study of Moral Reasoning Among Managers in Singapore. *J. Bus Ethics* 15 (12), 1331–1341. doi:10.1007/bf00411818

Winfield, A. F., Michael, K., Pitt, J., and Evers, V. (2019). Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems [Scanning the Issue]. *Proc. IEEE* 107, 509–517. doi:10.1109/jproc.2019.2900622

Wu, C.-F. (2003). A Study of the Adjustment of Ethical Recogntion and Ethical Decision-Making of Managers-To-Be across the Taiwan Strait before and after Receiving a Business Ethics Education. *J. Business Ethics* 45 (4), 291–307. doi:10.1023/a:1024167503455

Yacout, O. M., and Vitell, S. (2018). Ethical consumer decision-making: The role of need for cognition and affective responses. *Bus. Ethics. Eur. Rev.* 27, 178–194. doi:10.1111/beer.12178

Yan, X., Harb, R., and Radwan, E. (2008). Analyses of Factors of Crash Avoidance Maneuvers Using the General Estimates System. *Traffic Inj. Prev.* 9 (2), 173–180. doi:10.1080/15389580701869356

Zollo, L., Pellegrini, M. M., and Ciappei, C. (2017). What Sparks Ethical Decision Making? the Interplay between Moral Intuition and Moral Reasoning: Lessons from the Scholastic Doctrine. *J. Bus Ethics* 145 (4), 681–700. doi:10.1007/s10551-016-3221-8

Zollo, L. (2020). The Consumers' Emotional Dog Learns to Persuade its Rational Tail: Toward a Social Intuitionist Framework of Ethical Consumption. *J. Business Ethics* 168, 1–19. doi:10.1007/s10551-019-04420-4

Zollo, L., Yoon, S., Rialti, R., and Ciappei, C. (2018). Ethical Consumption and Consumers' Decision Making: the Role of Moral Intuition. *Md* 56 (3), 692–710. doi:10.1108/md-10-2016-0745

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Robotics and AI

ORIGINAL RESEARCH
published: 10 May 2021
doi: 10.3389/frobt.2021.637125

Drivers of Automation and Consequences for Jobs in Engineering Services: An Agent-Based Modelling Approach

Hildegunn Kyvik Nordås[1,2]* *and Franziska Klügl*[3]

[1]Economics and Statistics, School of Business, Örebro University, Örebro, Sweden, [2]Council on Economic Policies (CEP), Zürich, Switzerland, [3]Machine Perception and Interaction Lab, AASS, School of Science and Technology, Örebro University, Örebro, Sweden

New technology is of little use if it is not adopted, and surveys show that less than 10% of firms use Artificial Intelligence. This paper studies the uptake of AI-driven automation and its impact on employment, using a dynamic agent-based model (ABM). It simulates the adoption of automation software as well as job destruction and job creation in its wake. There are two types of agents: manufacturing firms and engineering services firms. The agents choose between two business models: consulting or automated software. From the engineering firms' point of view, the model exhibits static economies of scale in the software model and dynamic (learning by doing) economies of scale in the consultancy model. From the manufacturing firms' point of view, switching to the software model requires restructuring of production and there are network effects in switching. The ABM matches engineering and manufacturing agents and derives employment of engineers and the tasks they perform, i.e. consultancy, software development, software maintenance, or employment in manufacturing. We find that the uptake of software is gradual; slow in the first few years and then accelerates. Software is fully adopted after about 18 years in the base line run. Employment of engineers shifts from consultancy to software development and to new jobs in manufacturing. Spells of unemployment may occur if skilled jobs creation in manufacturing is slow. Finally, the model generates boom and bust cycles in the software sector.

Keywords: technology uptake, employment, automation, economic modelling, agent-based simulation

OPEN ACCESS

Edited by:
Martim Brandão,
King's College London,
United Kingdom

Reviewed by:
Matthew Studley,
University of the West of England,
United Kingdom
Taewoo Nam,
Sungkyunkwan University, South Korea
Tania Treibich,
Maastricht University, Netherlands

***Correspondence:**
Hildegunn Kyvik Nordås
hildegunn.kyvik-nordas@oru.se

Specialty section:
This article was submitted to Ethics in Robotics and Artificial Intelligence, a section of the journal Frontiers in Robotics and AI

Received: *02 December 2020*
Accepted: *21 April 2021*
Published: *10 May 2021*

Citation:
Kyvik Nordås H and Klügl F (2021) Drivers of Automation and Consequences for Jobs in Engineering Services: An Agent-Based Modelling Approach. Front. Robot. AI 8:637125. doi: 10.3389/frobt.2021.637125

1 INTRODUCTION

Due to recent advances in algorithms and technology based on Artificial Intelligence (AI), intelligent automation systems are rapidly moving into the workplace. AI technologies such as Deep Learning have become accessible to industry as a result of growing digitisation, the consequent availability of data, computation power and powerful tools, propelled by research by the leading technology companies. Nevertheless, the adoption of technology is gradual, often with long lags between innovation and adoption[1]. A survey of firms' use of AI released by Statistics Sweden in November

[1]Over the past 200 years in 166 countries it has taken 45 years on average from innovation to adoption of technology (Comin and Hobijn, 2010).

2020, for example, finds that only 5.4% of the firms surveyed use AI. In the United States, the 2018 Annual Business Survey found that 10.3% of firms use at least one of the advanced business technologies classified as AI[2]. Against this backdrop, it is clear that to assess the impact of AI on the future of work, one first needs to understand what determines the uptake of AI in firms. This paper contributes to filling this gap. It studies the uptake of AI-based automation and its determinants as a market interaction between developers and users of AI-based automation software.

It contributes to the literature in three major ways.

First, studies on the adoption of AI are few despite the observed long lags between innovation and adoption. It is well documented in the literature that adoption of new technology goes together with investment in intangible assets, including skills and organisational innovations (Brynjolfsson and Hitt, 2000; Rock, 2019). Nevertheless, standard models of technology adoption do not feature organisational changes. Our novel approach contributes to filling this gap by modelling AI-adoption as a switch in business model. Before AI-adoption, engineers serve their manufacturing clients through face-to-face, on-site interaction. AI-adoption implies using AI technology to automate the engineering tasks. Among the technologies we have in mind are machine learning, intelligent planning, automated reasoning, text mining and natural language generation. Many of these AI approaches have become applicable outside purely academic contexts due to accessible tools and platforms[3]. These automation technologies are embedded in software, such as intelligent systems for computer-assisted design (CAD), systems supporting additive manufacturing (3D printing), software for advanced construction of digital twins, software performing advanced data analysis and complex tests for verifying control software. Engineers switch from a consultancy model to developing, maintaining and licensing such software to clients. Manufacturers switch to a more skills-intensive software-supported production technology. The driving forces that we analyse are economic and institutional, notably uncertainties about user costs and benefits of the new technology, the switching costs to a different business model, the need for skills upgrading, and regulatory incentives or disincentives. Our focus on the demand side of technology diffusion provides new insights that can inform a balanced R&D, skills- and labour market policy.

Second, the study focuses on AI-adoption in services, noting that professional services are at the cusp of using AI-enabled automation[4]. Indeed, the Swedish AI adoption survey found that services sectors that produce and use ICT intensively have the highest AI-adoption rate in the economy. While AI is on its way into most professional services, engineering has a long history of developing technology for modern manufacturing, for instance through computer assisted design (CAD) feeding into computer assisted manufacturing (CAM). Here, with advanced image processing and new approaches combining data-driven learning and (spatial) reasoning, AI-based software can automate knowledge-intensive services previously performed by specifically educated engineers. The vision of Industry 4.0 (Lasi et al., 2014; Wang et al., 2018; Rock, 2019) further drives these developments. Today, civil engineers top the list of occupations most affected by AI while three other engineering occupations feature among the top 20 (Felten et al., 2019). Despite the susceptibility to automation, engineers are among the occupations with the fastest job growth in recent years[5]. Engineering is therefore of particular interest for understanding the relationship between AI and jobs in high-skilled services occupations.

A recent EU enterprise survey on the use of technologies based on AI found that about 60% of AI-using firms buy software or ready-to-use systems from external services suppliers[6]. Our modelling strategy reflects this empirical observation. Thus, engineering firms are the external suppliers of AI-enabled software and ready-to-use systems, engaging in market interactions with manufacturers. Most existing studies focus on the impact of robotics for automation in manufacturing. One reason for this is that while data on robot use is readily available, data on AI-enabled software use is not.

This leads to our third major contribution, which is to develop an agent based model (ABM) to study the joint adoption of AI in services and manufacturing. ABMs are particularly suitable for dynamic processes where outcomes are uncertain and agents interact (Dawid, 2006). Furthermore, it is apposite when the future is likely to be qualitatively different from the past such as during technological transitions (Farmer and Foley, 2009). Our ABM captures the interactions between the agents and the environment in which they operate and generates important insights on the trajectory of AI adoption. Notably, the model generates the boom and bust cycles often observed during the early stages of technology adoption. The combination of traditional economic modelling and the rigorous agent-based perspective used in our study results in a rather complex, yet comprehensive model. Using a stringent agent-based perspective, we avoid the "invisible hand" that automatically and instantly clears markets. Instead agents decide strictly based own experience, perception and individual economic reasoning, allowing us to trace out the process of technology adoption step by step. Agent-based approaches to economic modelling per se are not new (Tesfatsion, 2006; Hamill and Gilbert, 2016; Gatti et al., 2018), yet still far from being a mainstream approach in economic modelling.

Our model has two types of agents, engineering firms and manufacturing firms; and two business models, which we label consultancy and software respectively. Consultancy is the

[2]See https://www.scb.se/contentassets/4d9059ef459e407ba1aa71683fcbd807/uf0301_2019a01_br_xftbr2001.pdf for Sweden and Zolas et al. (2021) for the US.

[3]Accessibility of AI technology is not just improved by platforms for Machine Learning or Deep Learning, such as (Keras or Pytorch), but also by initiatives such as OpenAI https://beta.openai.com/or AI4EU https://www.ai4eu.eu/.

[4]See for instance (Baldwin and Forslid, 2020).

[5]see https://ec.europa.eu/eurostat/web/main/data/database.

[6]See https://digital-strategy.ec.europa.eu/en/library/european-enterprise-survey-use-technologies-based-artificial-intelligence.

traditional business model where engineering firms deploy consultants to clients, working with them on-site and face-to-face to solve problems. In the software model consultants are replaced by in-house engineers working with intelligent systems for automating engineering services. Manufacturers buy such software through licensing agreements, paying a license fee, or they may opt for cloud-based software-as-a-service, paying an annual subscription rate. The model generates a change of business model when engineers have gathered sufficient experience to create software solutions that automate services that were previously provided by consultants. Gathering this experience is modelled as learning-by-doing and represents dynamic economies of scale. Manufacturers decide whether to license software or stick to the consultancy model based on the expected costs and benefits of doing so. The benefits are uncertain at the time of the decision. Our analysis shows that it is hesitance on the part of manufacturers that holds back the uptake of AI-based software.

The rest of the paper is organised as follows: Section two discusses related research while section three develops a conceptual framework that captures the interaction between engineering firms, their clients and the environment in which they operate. The framework is coded into a dynamic ABM in section four. Section five presents the simulation results, while section six summarises and concludes.

2 RELATIONS TO PREVIOUS WORK

The literature on adoption of AI in the workplace is new and to the best of our knowledge this is the first paper to simulate the adoption of AI in business services. It builds on the theoretical literature on technology diffusion and adoption pioneered by Nelson and Phelps (1966); Rosenberg (1972); Davies and Davies, (1979); Stoneman and David (1986) and others. The theory is inspired by the stylised fact that the adoption of new technology follows an S-curve with slow uptake at an early stage, followed by a sharp rise in adoption when a critical mass is reached, until the market is saturated and the curve flattens (Gort and Klepper, 1982; Hall and Khan, 2003)[7].

Two different classes of theoretical models can explain such a pattern. The first envisages technology diffusion as the propagation of information, using models similar to those explaining epidemics. Observing that technology spreads much slower than epidemics and information, a learning process is added to the theory. Thus, firms learn by using new technology, and some of the accumulated tacit knowledge enters the public domain over time (Rogers, 1995).

The other major theory of technology adoption focuses on the characteristics of early technology adopting firms. Such models feature differences in firm size, productivity and abilities as explanatory variables. A new technology is fraught with uncertainty about its potential benefits, which introduces expectations as an important component of the theory. Furthermore, to reap the full benefit from a new technology, complementary investments in skills, reorganisation of production and rearranging relations to suppliers and customers are needed[8]. Therefore, the largest, most productive or otherwise most capable firms adopt new technology first (Davies and Davies, 1979; Rogers, 1995).

Recent survey data from the US and Sweden finds that it is indeed the largest and most productive firms that adopt AI. This explains the early, slow diffusion part of the S-curve. The subsequent acceleration in uptake may stem from standardisation of the technology as experience with using it accumulates, substantially reducing uncertainty over time. Network effects can also be important when the benefits from adopting the technology depends on suppliers or customer adopting it too. Then, the switching cost to the new technology declines as the number of users increases. Our model builds on the second strand featuring firms that differ in productivity, uncertainty about the benefits of new technology and switching costs. Our model also features network effects as well as learning by doing that reduces uncertainty and adoption costs over time. It generates the S-curve predicted by the theoretical literature in a setting of interaction between supply and demand and technology that has the features of AI-driven software, i.e. substituting for skilled workers, high cost of software development but zero marginal cost of adding another user (Varian, 2019).

Turning to the literature on technology and jobs, the most common approach to studying the impact of AI-related technology on jobs is to break jobs down to tasks and analyse the task content of different occupations (Autor et al., 2003; Acemoglu and Restrepo, 2018; Neves et al., 2019). The approach is to identify tasks that can be automated, vs. tasks that complement AI, and make predictions about the future of work from these metrics. In our context, this would generate business models where engineers may offer both software and consultancy, or it could generate deeper specialisation in the engineering sector where automatable tasks are performed by software while new tasks are performed by engineering consultants. However, this literature assumes that all tasks that can be automated are automated instantly, and thus assumes away adoption costs. Our contribution to the literature is precisely to focus on the scenario where existing technology is not instantly adopted, which is clearly the empirically most relevant case. The scenario is mentioned in Acemoglu and Restrepo (2018), but is not further developed. We explore and endogenise the uptake of technology as a function of wages, the cost of switching to AI-driven software, including the cost of reorganising production, and the expected gains from switching to new technology. Our model also features reallocation of engineering jobs across activities from consultancy to software development and maintenance, and to employment in

[7]See also (Geroski, 2000) for a review of the literature.

[8]Such complementary investment can be up to an order of magnitude larger than the initial investment in technology such as computers and other information technology (Milgrom and Roberts, 1990; Brynjolfsson and Hitt, 2000; Bessen, 2002; Bresnahan et al., 2002).

manufacturing where engineers work on technical problem solving using AI-driven software.

On methodology, our paper relates to Agent-based Modelling and Simulation that has become an established micro-simulation approach in social sciences, economics (Gallegati and Richiardi, 2009; Hamill and Gilbert, 2016), ecology and for modelling complex systems in general (Klügl and Bazzan, 2012). The underlying metaphor of such a model is a set of interacting agents—that can be basically seen as situated intelligent, autonomous actors (Wooldridge, 2009). A model captures agents' decision making in their individual environmental context which may be changing and influenced by multi-level feedback loops. During simulation, overall dynamics are generated. Consequently, agent-based simulation is particularly apt for modelling endeavours which involve heterogeneous agents, with transient dynamics and without the necessity of an equilibrium-based model. Technology adoption, which has all these features, is best understood through the lens of interacting agents. Our paper integrates insights from economics and Agent-Based Modelling by assigning decision making rules from economic theory to individual interacting agents within the framework of an ABM.

3 THE MODEL

3.1 Intuition

We propose a dynamic model consisting of two types of agents: engineering firms and their manufacturing clients. Manufacturers produce final goods according to a production function which combines production workers employed by the manufacturer and services inputs sourced from engineering firms. We distinguish two types of relationships between the engineering firms and the manufacturer: consultancy and software.

The consultancy model involves engineers working with the client, on-site, face to face, to solve problems and provide necessary services for production. The problems and services are client-specific and the ability to solve them rests with the consultant. The engineering firm and the manufacturer enter a contract which specifies the tasks the consultants are to perform as well as the payment, which is an annual fee per consultant. Contracts are setup anew every year; the number of consultants needed depends on the productivity and size of the manufacturer. Engineers are also explicitly modelled as discrete entities with individual experience that increases when working for a highly productive manufacturer.

In the software model the engineering firm establishes an R&D department where assigned engineers develop software that automates services adopting available AI technology such as machine learning or reasoning based on the problem solving experience of the engineers. The R&D activity requires a given number of engineers; their salaries constitute a fixed cost which the engineering firm recuperates through the licensing of the resulting software. Once developed, the software can be licensed to an unlimited number of manufacturers.

Each engineering firm offers its unique variety of the service, and thus distinguishes its product from competitors. Such product differentiation implies that the engineering firms may charge customers a premium and mark up their price over marginal cost. In the case of occupational licensing, engineers have exclusive rights to perform a predefined set of tasks. Furthermore, they may limit the number of licensed engineers and thereby charge a higher mark-up.

Manufacturers are heterogeneous in terms of size and productivity. Productivity is a measure of how effectively the firm transform inputs into outputs. Thus, the more productive firms use less engineering services per unit of output. Switching business model from relying on external consultants to using software, supported by in-house engineers, involves restructuring of production for a seamless interface between fabrication and the software. This requires upgrading of machinery and skills, creating jobs for engineers to manage the interface between the software and machinery, supervise production workers, support management in technical decision making, and govern the licensing contract with the engineering firm[9]. The dynamics of the model consist of learning by doing on the part of engineers working on problem-solving in manufacturing firms and network effects in the adoption of software.

3.2 Formal Model

Manufacturers, indexed i, are heterogeneous in terms of productivity denoted θ_i, which follows a Pareto distribution. The probability density function of the Pareto distribution is given by $g(\theta) = k(\theta_{\min})^k(\theta)^{-(k+1)}$ where $\theta_{\min}$ is the scale parameter, which we set to unity, and k is the shape parameter, which we tentatively set to 2.2[10]. The corresponding cumulative density function is $1-(\theta_{\min}/\theta)^k$. The manufacturers produce final output, denoted Y using production workers indexed l and engineering services. Total costs for the consultancy and software models respectively at time t are:

$$TC_{i,c} = \left[\frac{w_i}{\alpha} + \frac{\varphi w_s}{\beta_c \theta_i}\right] Y_i \tag{1}$$

$$E_t\left[TC_{i,softw,t}\right] = \frac{A_t}{\theta_i} w_l^{1-\beta} w_s^{\beta} Y_i + \delta + \gamma \tag{2}$$

TC represents total cost of production. The two business models are indexed c and *softw* respectively. In both cases we apply constant elasticity of substitution production- and cost functions. In the consultancy model we use the extreme case of a Leontief specification where production factors are perfect complements, while in the software model we apply the Cobb-Douglas functional form where the elasticity of substitution between factors is unity. These particular functional forms are not critical for the results, but serve to distinguish between more and less flexible technologies in the two business models.

[9]There is ample evidence that ICT and AI complement skills in the workplace. See for instance (Berman et al., 1998; Autor et al., 2003; Bessen et al., 2018; Brynjolfsson et al., 2019).

[10]This is close to empirical estimates of the shape parameter of productivity distribution from firm level data (Feenstra, 2018). A shape parameter larger than two ensures that the variance of the distribution can be identified.

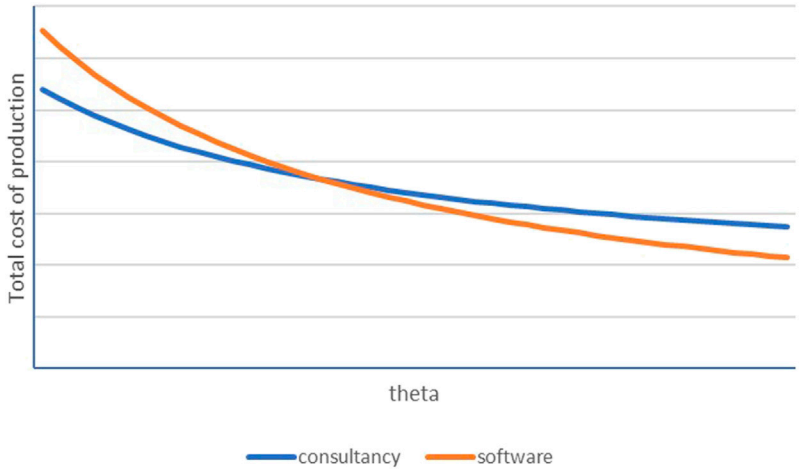

FIGURE 1 | Total cost comparison, consultancy and software.

Variables and parameters: α represents the production worker intensity while β_c depicts the consultancy input intensity of manufacturing production in the consultancy business model. Wage rates for production workers and engineers are denoted by w_l and w_s respectively, while the mark-up rate that engineering firms obtain for their consultancy services is φ. A scale parameter A, the license fee for software δ, and a stochastic element γ are additional parameters in the cost function for manufacturers that opt for the software model. The stochastic element γ is normally distributed $\gamma \sim N(0, \sigma^2)$. Manufacturers will be in the market for software if the expected cost of switching to software is lower than continuing with the consultancy model. **Figure 1** illustrates the two cost functions where the horizontal axis represents manufacturing firms' productivity and the vertical axis total cost. Clearly, software represents the lowest cost for high-productivity firms, while consultancy is the better option for low-productivity manufacturers[11].

There are network effects related to the switch to the software model as adopters reorganise production, including relations to suppliers and customers around the software. Also professional organisation's investment into competence development speeds up technology adoption. We capture this by modelling the scale parameter A to be a declining function of the number of firms that have switched to software. The network effect works with one period lag.

$$A_t = \frac{A_{t-1}}{n^{\mu}_{softw,t-1}} \tag{3}$$

where $0 < \mu < 1$. Demand for engineering consultancy services from each manufacturing firm choosing that model is given by:

$$C_i = \frac{Y_i}{\beta_c \theta_i} \tag{4}$$

[11]The figure is drawn for the parameter values depicted in **Table 2**.

Manufacturers that have switched to the software model will seek to employ engineers according to the demand function:

$$S_{i,t} = \frac{A_t}{\theta_i}\left[\frac{\beta}{1-\beta}\frac{w_l}{w_s}\right]^{\beta} Y_i \tag{5}$$

Engineering firms, indexed over j hire engineers which are deployed to client firms on a contractual basis in the consultancy model. The contract covers one period and its value varies across clients, depending on their size and productivity as indicated in the demand function, **Eq. 4**. The engineering firms incur wage costs only and they sell consultancy services with a mark-up factor of $\varphi > 1$. The consultancy revenue is thus $\varphi w_s \sum C_i$. We choose units such that one unit of consultancy services corresponds to the input of one full-time consultant for one period. Profits from the consultancy model at time t are:

$$\pi_{c,j,t} = (\varphi - 1)\sum C_{i,t} \tag{6}$$

where the number of manufacturing clients changes over time. In the software model, engineering firms establish an R&D department and divert $S_F/\lambda_{j,t}$ engineers to staff it. The R&D department uses available data sets and experience from previous consultancy efforts to create AI-based software that provides the services that are otherwise done by consultants. Thus, engineers accumulate experience from working with clients, and harness this experience into software that automates the consultants' work.

It is assumed that a minimum number of experienced engineers is needed to successfully develop the software. So, experience accumulated over years of working with clients is an advantage when developing software, assuming that experience helps to identify appropriate machine learning architectures and to formalise knowledge for automated reasoning. We model this by introducing the experience of the engineer, denoted λ in the cost of developing the software. The total cost of switching for the engineering firm is the wage costs

for the engineers working in the R&D department and the foregone profits from no longer deploying them to clients as consultants. Revenue in the software model will be the licence fee δ times the number of manufacturers that license the software from company j; $n_{softw,j,t}\delta$. Expected profits from the software model at time t is thus:

$$E_t\left[\pi_{softw,j,t}\right] = E_t\left[n_{softw,j,t}\right]\delta - \frac{w_s S_F}{\lambda_{j,t}}\varphi \tag{7}$$

The engineering firm knows the cost of developing software, but at the point of decision whether to develop it, the number of clients that will take up the software is unknown. The engineering firm does, however, observe the productivity of the manufacturers and thus can estimate how many of them are sufficiently productive to gain from switching to software. Engineering firms base their decision to develop software on expectations about how many clients they may capture from the mass of manufacturers that are sufficiently productive to benefit from switching to the software model.

After software is available, manufacturers that decided to switch their business model to software, randomly select engineering companies that offer software. Random selection is weighted by experience of the software provider assuming that more experienced firms produce higher quality software. Since the marginal cost of servicing another client is zero, it is conceivable that one engineering firm could corner the market.

It is clear from **Eq. 7** that profits from switching to the software model are lower the higher the mark-up factor φ, predicting that engineering firms operating in a less competitive market, for instance a small market with occupational licensing, are less innovative than firms operating in a competitive market which limits the ability to charge a high price[12]. Engineering firms will develop the software if expected profits as defined in seven is positive.

After the initial investment into software development, the software life-cycle contains a number of periods with software maintenance. It is assumed that data-driven software is depreciating fast, and lasts for T periods. Each period between its development and obsolescence a fraction ζ of the number of engineers that are needed to develop the software, is sufficient to maintain it. After T periods, the engineering firm needs to invest again into full software development. We assume no influence of the age or status of the software on its licensing fee[13].

Experience accumulates from working on-site and face to face with manufacturing clients. Furthermore, engineers gain more experience from working with the most productive manufacturers. An engineering firm j's accumulated experience is thus a function of the productivity of the manufacturers it has worked with as follows:

$$\lambda_{j,t}\int_0^t f\left(\theta_{ij}\right)\mathrm{d}\theta \tag{8}$$

These eight equations, representing supply and demand for engineering services in two business models constitute the conceptual core of the ABM. The forces that drive the adoption of software are engineers' accumulated experience from working with clients and network effects from its adoption. What holds back the development of software is comfortable profits from the consultancy model, uncertainty about how many manufacturers will buy the software once the cost of developing it is sunk on the part of the engineering firms, and uncertainty about the gains from the switch to software on the part of manufacturers. These countervailing forces ensure a gradual adoption of software in the economy. The speed depends on the size of the economy, the endowment of production workers and engineers, the level and dispersion of productivity among manufacturing companies as well as policy-induced factors including occupational licensing and protection of intellectual property rights.

4 THE AGENT BASED MODEL SETUP

The agents and their role and actions are presented in **Table 1**.

The environment consists of supply of production workers and engineers, a set of exogenous parameters and decision rules as spelled out in the model presented in **Section 3**. All agents act in parallel and go through their individual processes within one period. **Figure 2** illustrates what happens in one period including the synchronisation points between the activities that each engineering company and each manufacturer agent perform in parallel. So, manufacturer agents first determine their service needs—this happens in parallel when each engineering company either publishes their consultancy offer or offers software to be licensed (only after period two in the simulation). Then, manufacturers evaluate the offers and enter contracts or license software. After the next synchronisation step, production happens, partially with the help of consultants. The next steps with different synchronisation points are devoted to decision making for both manufacturers and engineering companies. First the manufactures reason about profitability of using software instead of hiring consultants and signal their interest. This is observed by the engineering firms who—with the information on potential size of the market for software, decide about whether they want to produce software or continue offering consultancy services. All decisions have consequences on employment of engineers.

The simulation runs through the following phases[14]:

- In phase 0—during initialisation –, manufacturer agents draw their productivity level from a Pareto distribution.

[12]International trade in engineering services would also limit the ability to charge a high mark-up and thus spur innovation. Adding space and different wages across countries could further exploit this point in future work.

[13]By explicitly integrating the software lifecycle into the model, we also capture the idea that technology is not static, but needs to be updated from time to time.

[14]We labelled the phases after analysing all simulation runs, also illustrating the shared observations. All experiments start with a situation in which all engineers are employed at engineering companies who exclusively offer consultancy services.

TABLE 1 | Summary of the modelled entities, their roles and activities.

Agents	Status	Role and actions	
		Consultancy	Software
Manufacturing firms	Active	Employ production workers Enter consultancy contract Produce final output	Employ production workers License software Employ engineers Produce final output
Engineering firms	Active	Employ engineers Enter consultancy contract	Employ engineers Develop and maintain software License out software
Production workers	Passive	Work in manuf. firms	Work in manuf. firms
Engineers	Passive	Work in eng. firms	Work in any firm
Authorities	Passive	Occupational licensing, IPR protection	

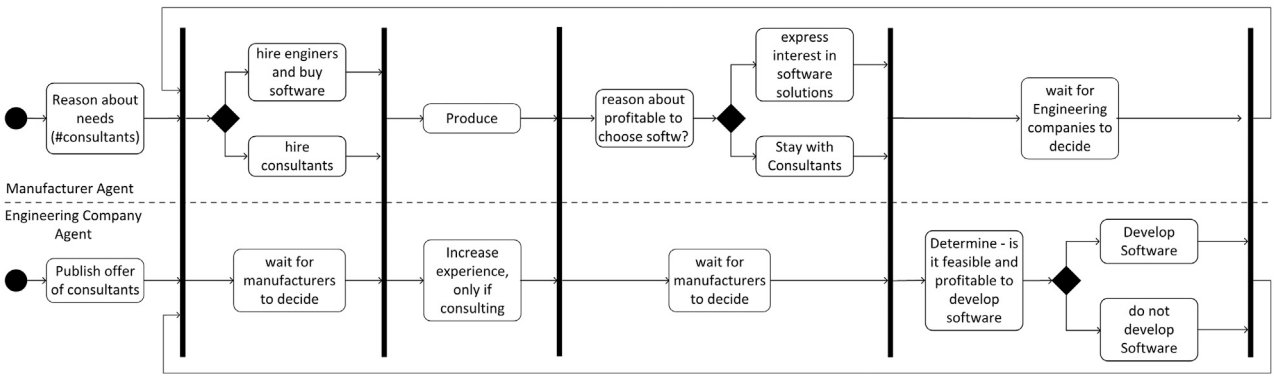

FIGURE 2 | Activities of the different agent types and their coordination. Note: The black vertical lines form synchronization bars meaning that all agents need to have finished the activities before an individual agent can continue with the next activity after the bar.

Manufacturers hire production workers, which are matched to firms randomly, but the number of employees is proportional to the firms' productivity. Engineering firms hire engineers, which are randomly matched to engineering firms.

- In phase 1 – first year—all engineering firms adopt the consultancy business model. Engineering firms and manufacturers are matched randomly and manufacturers produce final output.
- At the end of phase 1, all active firms observe their profits. Engineering firms' experience parameter is updated. Engineering firms then consider, whether to develop software and automate their service or continue with offering consultancy services. The decision is based on expected number of clients ready to switch to the software model, and the cost of developing the software. The cost is lower for the more experienced firms. For deciding about the potential market for their software, the engineering firms observe how many manufacturer agents would be interested in software. They expect to sell to a random subset of those manufacturers who are ready to switch. If expected profits from selling software is positive, engineering firms will establish an R&D department which will work on software development. Redundant engineers, that means those not engaged in the software development, are laid off. Manufacturers decide whether actually to switch to the software model. The decision is conditioned on software being available as well as there being engineers available on the market to hire in the new jobs created during the switch to the software model. As a consequence, the most productive manufacturers are the first to switch to software. If expected profit from the software model is smaller than that for continuing with the consultancy model for all engineering firms, phase 1 is repeated and consulting engineers gain more experience during each repetition.
- In phase 2 at least one engineering firm has developed software and earns a positive profit from licensing it. In this phase the two business models coexist. Some manufacturer agents having switched business model, license software from a random supplier and hire engineers to integrate the software into the production

TABLE 2 | Exogenous variables and parameters.

Symbol	Description	Value in the baseline case
	Number of engineering firms	30
	Number of manufacturers	100
L	Number of production workers	3000
S	Number of engineers	1000
w_l	Salary, production worker	1
w_s	Salary, engineer	1.5
α	Production worker intensity, manufacturing, consultancy model	1
β_c	Consultant intensity, manufacturing, consultancy model	1.5
β	Engineer intensity, manufacturing, software model	0.2
θ_i	Productivity level, manufacturing firm *i*	Pareto distributed
A_0	Scale parameter, manufacturing, software model	3
μ	Strength of network effects of using software	0.02
δ	License fee, software	10
γ	Stochastic switching cost, manufacturing	normally distributed
λ_0	Initial experience, engineers	1
η	Update factor for λ	0.1
φ	Mark-up rate, consultancy	1.3
S_F	Number of engineers needed to develop software	18
ζ	Software maintenance cost relative to development cost	0.5

process, other manufacturer agents continue hiring consultants. Manufacturers that do not license software and do not find consultants, do not produce output, all others do. Not all engineering firms developing software may be profitable. Making a loss from software development, causes engineering firms to immediately return to offering consultancy services.

- Manufacturers' cost of switching to software is adjusted by the network effect given by **Eq. 3**. The more manufacturers use software, the cheaper it becomes for latecomers to switch, and eventually also the less productive manufacturers can afford software. Manufacturer agents who cannot recruit consultants nor can afford software, do not produce in the current cycle, but wait for opportunities in the next period. Software is maintained (bug fixes, new, minor features in small updates) at a cost ζS_F with $0 < \zeta < 1$. When a software has reached obsolescence, the engineering firm decides again whether in a changed market, it could generate profit when re-developing software. Experience of engineers working as consultants is updated. Phase 2 continues until all manufacturing firms have switched to software.
- In phase 3 all firms have switched to software. There is a churning of engineering firms as software becomes obsolete and new software is developed to replace it. At this stage, engineers no longer gain experience from working directly with clients, but more are employed to support the software usage at the manufactures. There is still some dynamic ongoing at the engineering firms, as manufacturers re-select software in each period—we do not assume commitment to a particular software product. As a consequence, even when producing software in a market in which every manufacturer uses software, some software firms may lose customers to competitors, and possibly make losses on their investments.

Exogenous variables and parameters are summarised in **Table 2**.[15]

The ABM was implemented using the SeSAm platform[16] which is a fast prototyping environment for agent-based simulations providing an activity diagram-like way of implementing complex agent behaviour.

5 RESULTS

We start by running the simulations with baseline parameters as reported in **Table 2**, including sensitivity analysis on the overall size of the sector and the ratio of production workers to engineers. We experiment next with policy relevant parameters: 1) the mark-up rate, which is related to the strength of competition in the engineering services market and 2) the license fee, which is partly related to the strength of intellectual property rights protection and partly to the strength of competition in the market for software. Eventually, we want to explain what are the relevant factors influencing how fast intelligent automated solutions distribute in a market characterised by the parameters

[15]The parameter values reflect empirical relations observed in OECD countries. Wages of production workers are the numeraire in the model and set to unity. Data on employment by occupation and sector is not readily available, but the share of university educated workers in the total labour force is about 20% in the EU and 23% in the US https://ilostat.ilo.org/topics/employment/. The wage premium for professionals relative to plant and machine operators was about 1.4 in Sweden and 1.7 in the US in 2019 according to ILO statistics. https://ilostat.ilo.org/topics/wages/. Technicians and associate professionals account for 16% of all employees in manufacturing in the European Union, while computer, mathematical, architecture and engineering professionals account for another 9% https://ec.europa.eu/eurostat/databrowser/view/lfsa_eisn2/default/table?lang=en. The mark-up rate is also consistent with rates in the literature, while the other parameters are used to calibrate the model and to experiment with different scenarios.

[16]www.simsesam.org.

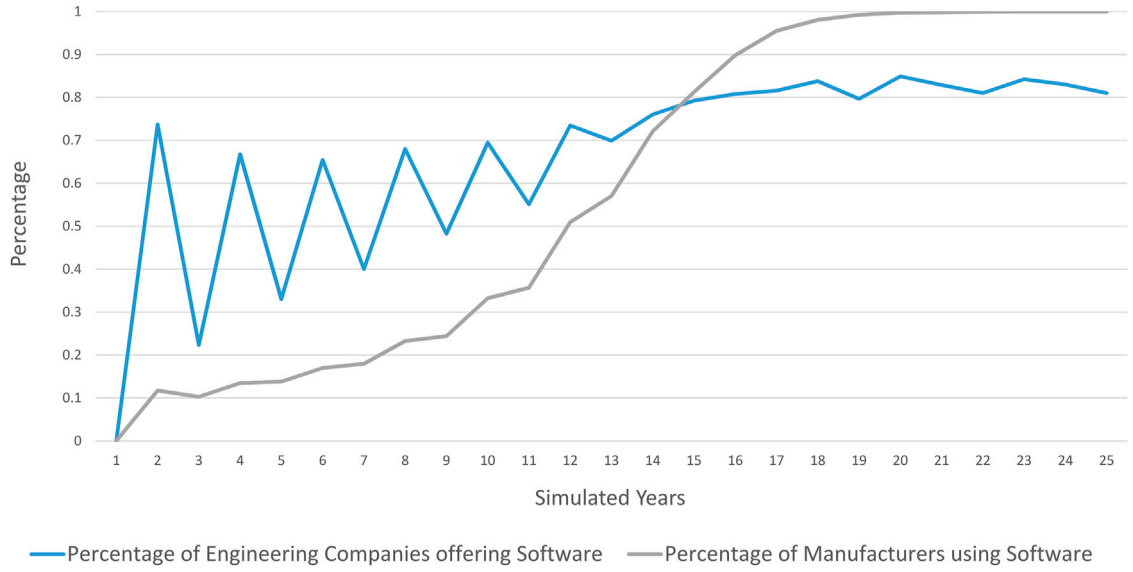

FIGURE 3 | Percentage of companies having switched to the software model Note: Baseline setting as described in **Table 2**.

above as well as what is the actual impact and dynamics on the employment of highly qualified engineers.

5.1 Baseline

We start by simulating the baseline scenario[17]. As described, we start with a scenario where all firms are in the consultancy model. Firms next decide whether to switch to the software model and look for a supplier or customer for software. As **Figure 3** indicates, a few manufacturing firms already switch to software in the second year. All engineering firms anticipate the market opportunity these firms constitute, and a large share of them decides to develop software. However, the customers are few, competition is fierce, and most early software developers fail. As a consequence, those failing firms give up to offer software[18].

The uptake of software in manufacturing is gradual and about half of all manufacturers have switched to software after 11 years. The uptake does, however, accelerate after about a third of all manufacturers have switched, and levels off when about 90% of firms have switched[19]. During the first decade of relatively slow uptake, there is a competitive fringe of engineering firms that develops software, fails and exits as indicated by the zigzagging of the blue line in the chart. After all manufacturers have switched to the software business model, about 80% of engineering firms offer software. There remains a competitive fringe of engineering firms that exit when a loss from software happens, when a new development is necessary, but too expensive or when simply not a sufficient number of software licenses were acquired. A start-up seeks consultancy contracts, but realises that demand for such services is close to zero and quickly starts to develop software as well.

What we see in our simulation shown in **Figure 3** is a largely demand-driven adoption of automation software, and a boom and bust cycle in the automation software sector. The booms are driven by all engineering firms simultaneously forming expectations about the number or clients that will shift to software (**Eq. 7**). However, not all software firms will find customers for their software. Those who do not, exit the software market and reestablish as consultant. This cycle is similar to the so-called *dot.com* bubble that could be observed in the 1990s when adoption of ICT took off, although in that case the financial market amplified the cycle[20].

For explaining this overall behaviour, a look into the dynamics on the agent level is helpful. **Figure 4** depicts the lifeline of two randomly selected engineering firm agents. They both start out as consultants and earn a positive profit. They both end up profitably licensing software, and they both have at least one unsuccessful attempt at switching to software. The first company has two spells of consultancy after a commercially successful

[17]We repeat every simulation 30 times. If not otherwise stated, diagrams show averaged values. Where suitable, we also give the standard distribution which is naturally higher in the transient phase 2 and low in phase 3.

[18]Technically, this may also be modelled as an exit of the firms that fail to sell the software they have developed, while start-up engineering firms use the consultancy model, or as a single firm switching between business models. The results are the same either way.

[19]Recall that the model captures innovation using existing AI technology to develop software. Considering software evolution with explicit software maintenance does not account for technological game change, rather for small, yet continuous improvements. Had the underlying AI-technology changed, a steady state might not occur. The authors thank an anonymous referee for making this point.

[20]See for instance (Doms et al., 2004) for a study of the *dot.com* bubble in the United States. Our model does not have a financial market, but still generates a boom-bust cycle due to expectations and herd behaviour. The authors thank an anonymous referee for making this point.

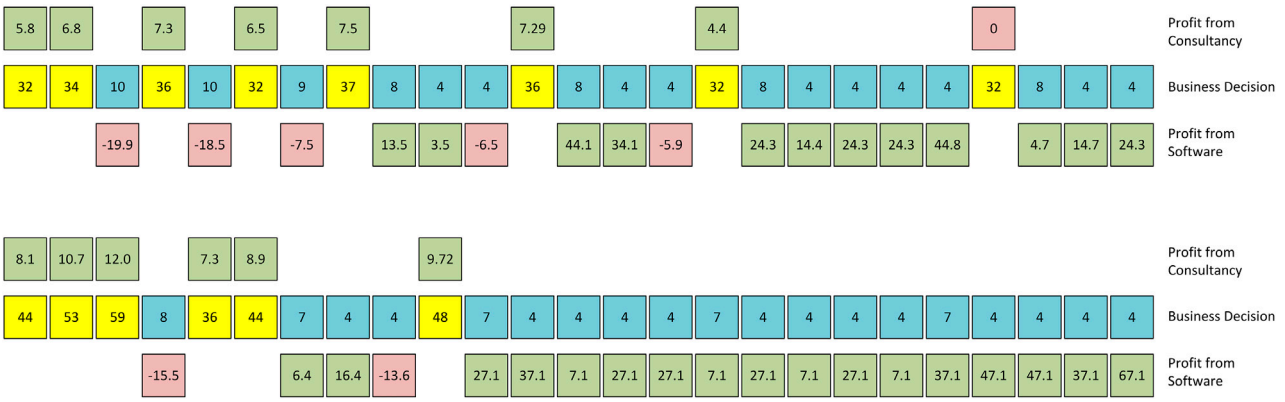

FIGURE 4 | Two examples of the lifeline of randomly selected engineering firms. Note: The middle row shows the selected business model (consultancy in yellow, software in blue). Numbers in these cells denote the numbers of employed engineers in that year. The upper cells contain the profit the company made in cases of consultancy, the lower cells profit or loss when trying to sell software.

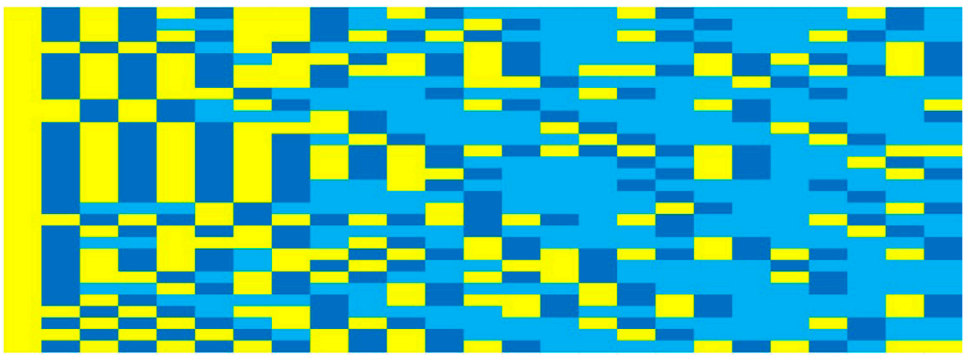

FIGURE 5 | Business model dynamics of all engineering companies over 25 simulated years. Note: Yellow means that the agent offers consulting, dark blue: it develops software, blue: it maintains software.

software becomes obsolete, while the second company experiences only one such event.

Figure 5 shows the business model dynamics for all engineering firms over the complete simulation run. We observe that they have all entered the software model after two years, but only three are successful and continue in the third year to maintain and develop their software. As time passes, the dynamics turn increasingly toward a shifting between developing new and maintaining existing software, but all firms experience occasional failures in the market for automated software.

In addition to the technology uptake, we also want to analyse dynamics of the composition of engineer employment. **Figure 6** depicts the dynamic impact of technology adoption on employment of engineers. All engineers work as consultants in the first year. Consistent with the changing business model, they gradually move to the R&D department in the engineering firm where they develop and subsequently maintain software. Consultants that cannot find a job in the R&D department are laid off. Most of them find new jobs in manufacturing firms that have switched to software and are looking for engineers to fill new jobs created during the transition to a more sophisticated and skills-intensive production process. Finally, some of the laid off engineers do not find a new job immediately, and become unemployed. We notice that with substantial economies of scale in software development, the number of workers needed to develop and maintain software is relatively small. Our simulations thus predict that most of the changes in employment are from external consultants to engineers working in manufacturing[21].

An interesting parameter is β, the engineer intensity influencing how many engineers are needed to support complex software usage at the manufacturer (**Eq. 5**).

The unemployment rate among engineers following the transition to software depends crucially on the ratio of

[21]Our model has a fixed number of workers and engineers. As software and learning by doing reduces the unit cost of production, employment in manufacturing and engineering firms may decline and unemployed workers may seek work in other sectors. Transition of workers to other sectors is not directly captured by our model. However, by keeping wages fixed, we implicitly capture an outside option at the going wage for workers.

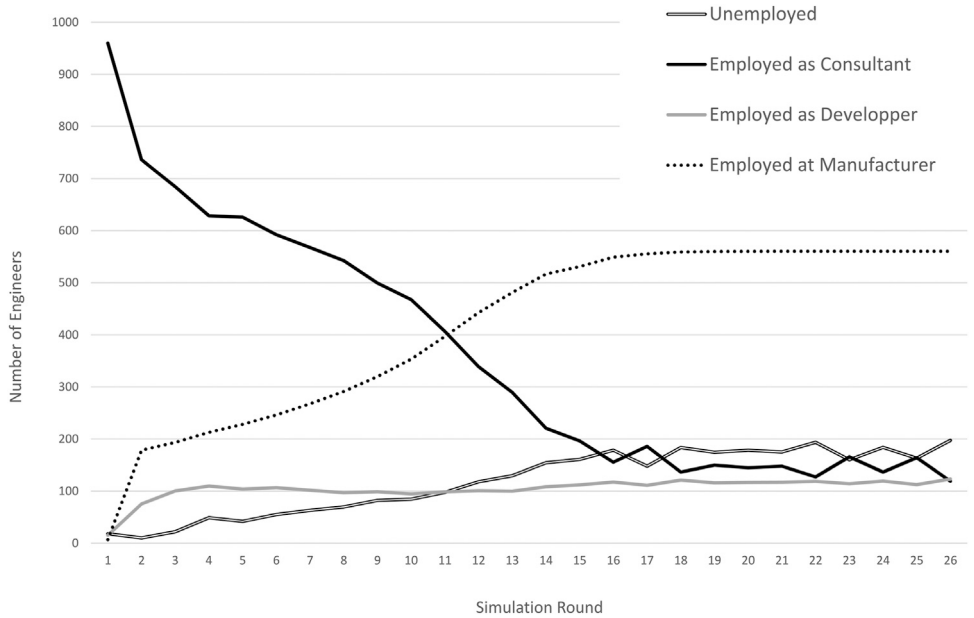

FIGURE 6 | Development of employment over simulation time. Overall number of engineers is 1,000.

production workers to engineers in the labour market and the desired skills composition of employees in manufacturing firms that have switched business model. Sensitivity analyses depicted in **Figure 7** shows that there will be full employment of engineers at the end of the transition period if β is larger than about a quarter. Sensitivity analyses also show that with fewer engineers in the market relative to production workers there could also be shortage of engineers at lower levels of β. Our results reflect the S-curve of technology adoption predicted by the theoretical literature e.g. (Gort and Klepper, 1982; Hall and Khan, 2003). It is also compatible with recent shifts in employment patterns where the share of professional jobs in manufacturing has increased from 5.7 to 9.4% from 2008 to 2019 in the European Union, and the share of technicians and associate professionals have increased from 13.4 to 15.6% during the same period[22]. Finally, our results reflect work by Andrews et al. (2015) which shows that the most productive firms are the first to adopt new technology.

5.2 Experiments, the Mark-Up Rate

The mark-up rate reflects the strength of competition in the market for consultant engineering services. High mark-up rates may stem from occupational licensing that gives licensed engineers exclusive rights to perform a defined set of engineering tasks, a small market closed to foreign competition, or simply a shortage of engineers for instance due to low education capacity for engineers or a limited number of engineering licenses issued.

From **Eqs. 1** and **2** we see that a high mark-up rate makes consultants relatively more expensive than software. On the other hand a higher mark-up rate yields higher profits for the engineers in the consultancy model (**Eq. 6**). Thus, manufacturers are more likely to switch to software the higher the mark-up rate, while engineering firms are less likely to switch the higher the mark-up rate. It follows that if adoption of the software model is driven from the demand side, the adoption rate increases as the mark-up rises. If on the other hand the uptake is driven by a supply push, then we would expect it to be delayed for longer the higher the mark-up rate. **Figure 8** clearly shows that this is a demand pull story.[23]

Figure 9 shows employment of engineers by sector and activity after 25 periods as a function of the mark-up rate. We first notice that employment of engineers in manufacturing is largely unaffected by the mark-up rate. After 25 periods all manufacturers have switched to the software model and pay engineers the going wage w_s, rather than the marked-up consultancy fee, so this is no surprise. Employment as software developer is also largely unaffected by the mark-up rate. Where we do see a significant difference is on the employment of consultants and the unemployment rate for engineers. The employment as consultant is actually very brittle as there is practically no market for consultancy services. Engineering firms that made a loss with software provision, try to re-establish with consultancy, yet there is

[22]Unfortunately, more detailed employment data by occupation and sector is not publicly available. However, EU aggregates should be representative for developed countries.

[23]There are parameter values where this may not be the case. These are however outside the scope of what is reasonable considering available data.

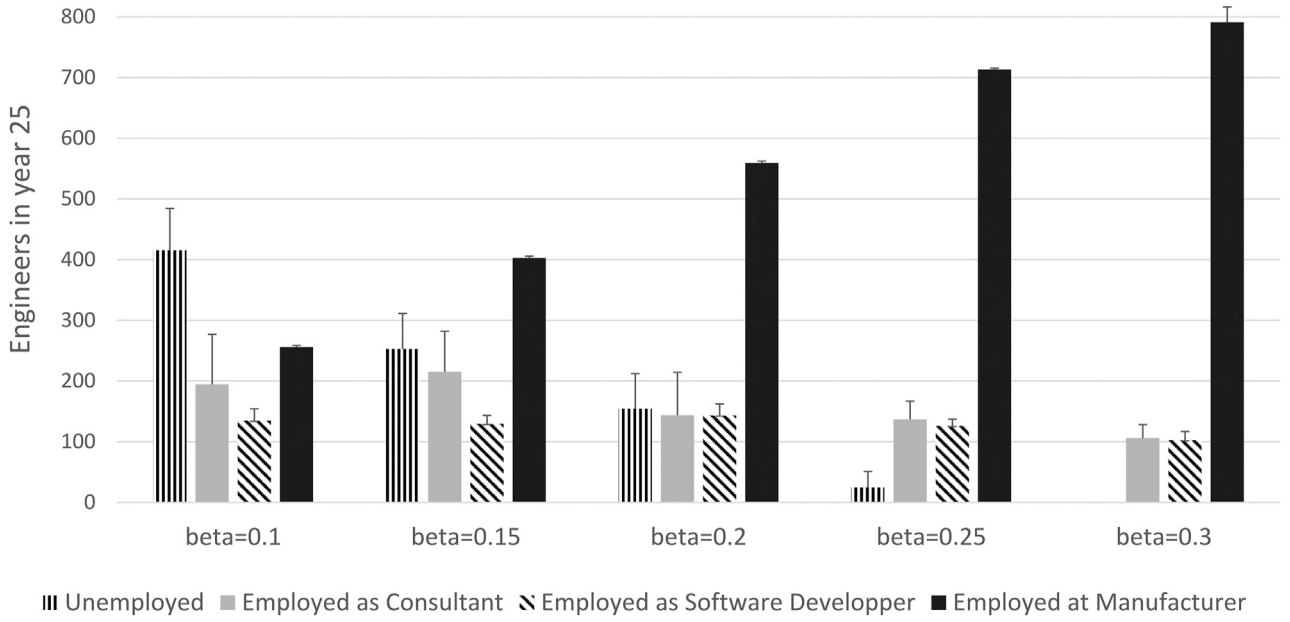

FIGURE 7 | Influence of β on the employment structure.

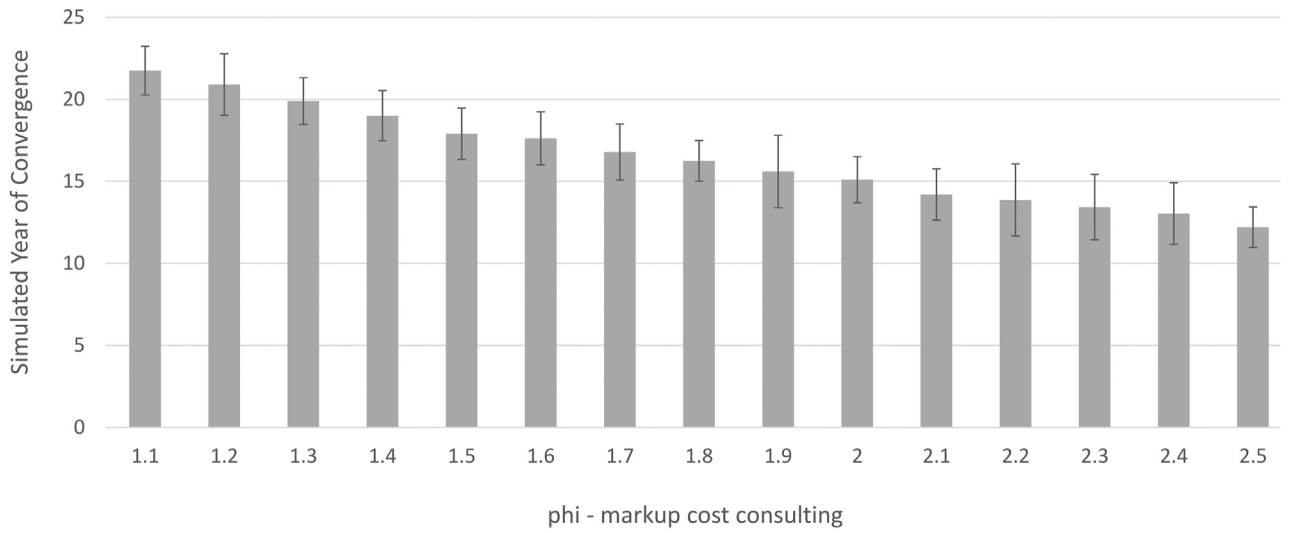

FIGURE 8 | Mark-up rates and simulated year in which 100% manufacturers were using software. Note: Averaged over 30 simulation runs, standard deviation shown as error bars. In the case of φ = 1.1, 1.4 and 2.4 two runs was omitted as not converged. With φ = 1.3, 1.6, 1.8 and 2.5 convergence was not reached in one single run. We consider those cases as outliers with randomly generated manufacturers that are particularly small and omitted those from the analysis.

hardly any demand for consultancy services and thus the profit from consultancy is zero. Such a company lays off all the newly recruited engineers, but tries to re-recruit them again in the next cycle, when deciding about producing software or offering consultancy again. Yet, depending on the competition in recruiting engineers, the full number may not be available any more.

An important policy implication of the simulations is that the potentially harmful delay of the uptake of technology due to occupational licensing does not materialise in a demand-driven market. This conclusion holds when the mark-up rate is unrelated to the software license fee and thus, exclusive rights do not extend to software licensing. We now turn to an experiment where we let the license fee vary.

5.3 Experiments, the License Fee

As **Figure 10** indicates, the adoption rate of software is slower, the higher the license fee δ is set. From **Eq. 2** we observe that cost of

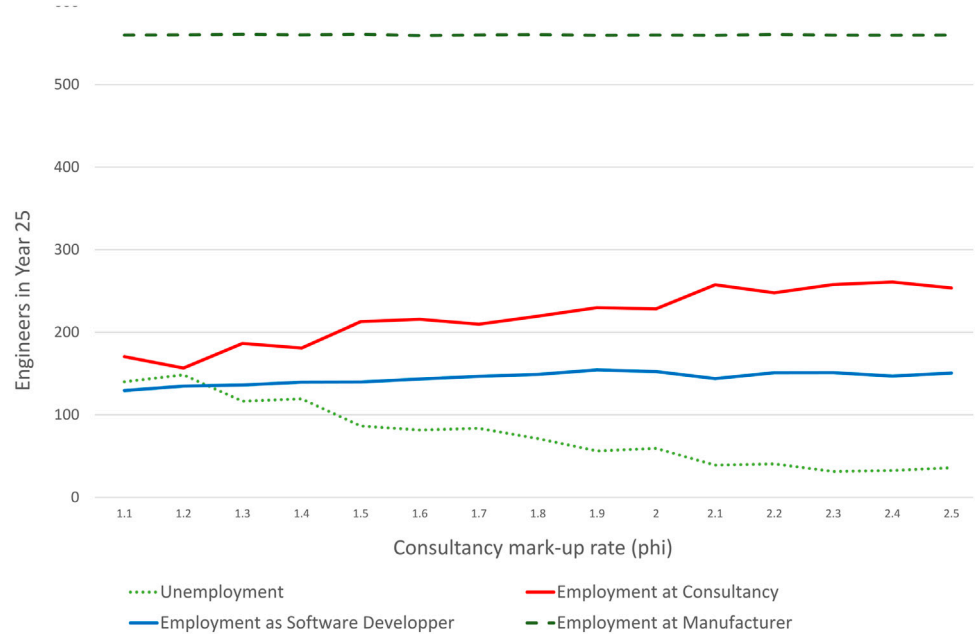

FIGURE 9 | Employment in simulated year 25 depending on the consultancy mark-up rate. Averaged over 30 simulation runs.

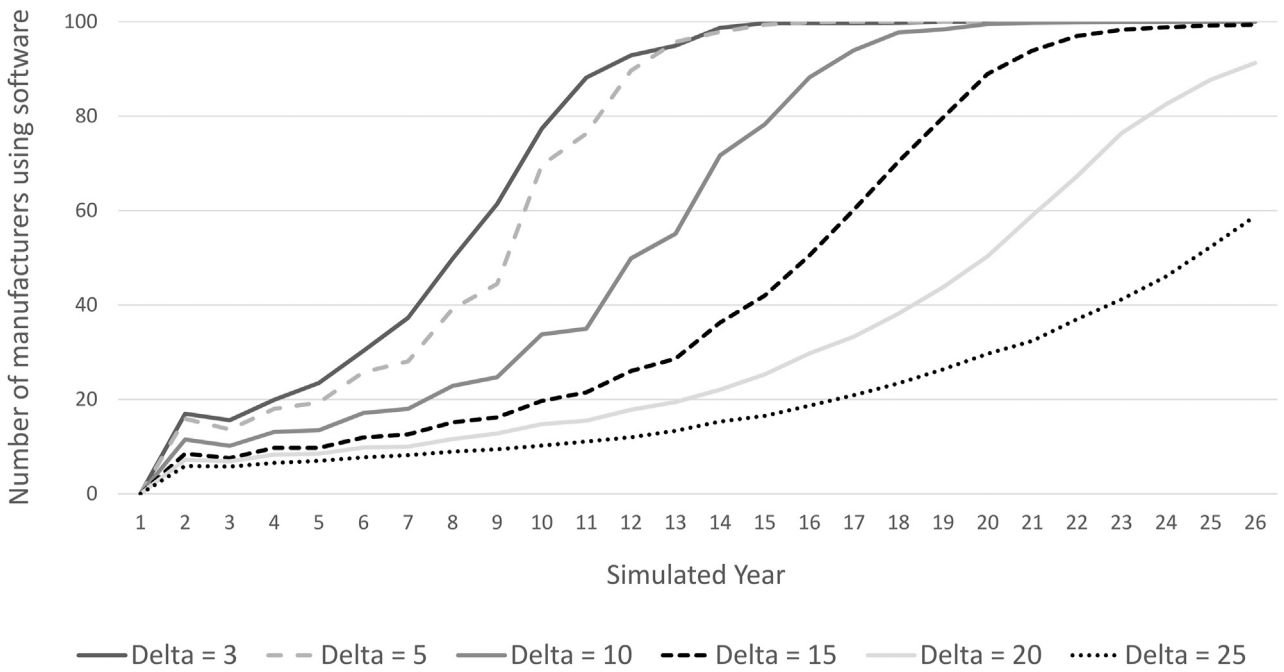

FIGURE 10 | Number of manufacturers who use software over simulated time with different settings of the licence fee Delta.

production is higher for manufacturers the higher is δ, so there will be fewer takers of software the higher is δ. This also results in weaker network effects, further slowing down the uptake over time (see **Eq. 3**). On the other hand, as depicted in **Eq. 7** the revenue of the engineering firms providing software is higher the higher is δ, all else equal. Thus, the slower rate of transition to the software model stems from the demand side.

A convergence toward a situation in which all manufacturers use software happens also in scenarios with high licence fees. Longer simulation runs with $\delta \geq 20$ confirm that the adoption rate

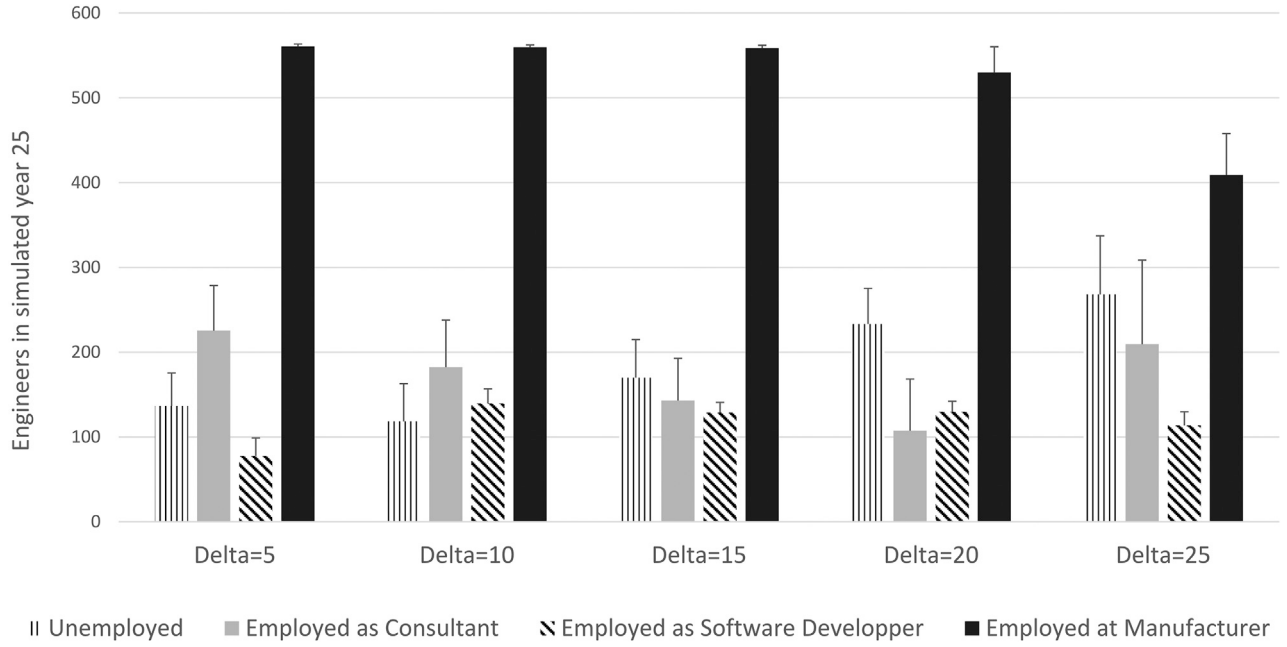

FIGURE 11 | Employment of engineers in the simulated year 25. Note: Average over 30 runs, the error bar shows the standard deviation between runs. Delta is the licence. All other parameter according to baseline scenario.

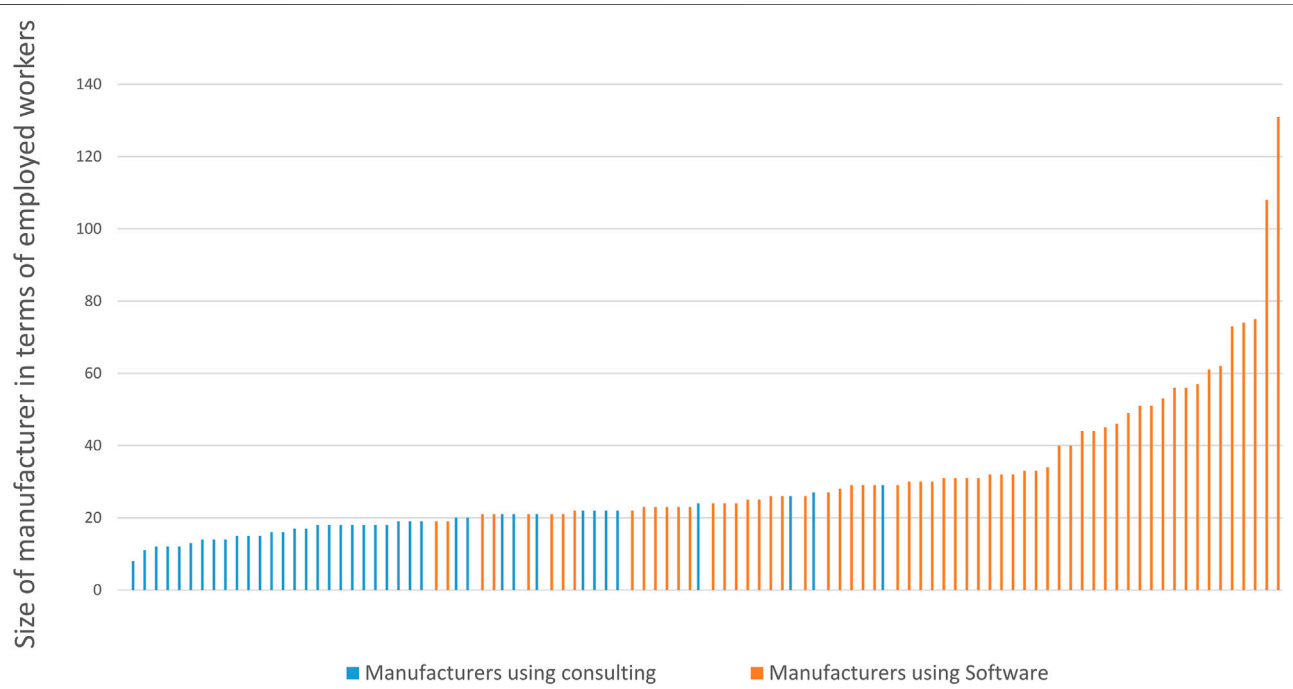

FIGURE 12 | Illustration of the dependency between size of a manufacturer and its decision to use software. Note: Delta = 50, in year 25. With γ, there is a random element in the decision making: for sizes between 19 and 29 employed workers, both decisions are observed, yet with corresponding tendencies.

eventually approaches 100%. For example, with $\delta = 25$, the population of firms converges into a stable situation with slightly less than 100% of manufacturers using software between years 50 and 60. In these simulations, two engineering firms are too small to start software development and try to hold on to the consultancy model.

A higher license fee also results in a higher rate of unemployment among engineers during and after the transition to software as illustrated in **Figure 11**. Consultancy jobs are lost, and job creation in the R&D department to develop and maintain software together with engineering jobs created in manufacturing is insufficient to absorb the idle consultants. However, sensitivity analyses with a higher β show that unemployment among engineers is substantially reduced or even eliminated when also δ is higher than in the baseline scenario.

Finally, **Figure 12** shows the number of manufacturing firms that take up software in a scenario where software is very expensive and the two business models co-exist also in the long run. It illustrates that the first adopters are the largest and most productive manufacturing firms. Further, since firms may have different risk assessments related to switching to software, there is a mix of software adopters and consultancy users in the middle range of firm size and productivity levels.

6 CONCLUSION

Economic history documents that the adoption of technology is gradual with long delays. Furthermore, it is amply documented in the business literature that the adoption of new technology in firms requires organisational changes and new skills, which constitute significant switching costs for individual firms. Nevertheless, recent literature on AI and the future of jobs overlooks or abstracts from such switching costs and assume that AI-based service automation technology is adopted as soon as it is invented, with dramatic effects on jobs. To understand, predict and prepare for the labour market implications of AI on jobs a much better grasp on what drives the *adoption* of technology is needed. Our paper contributes to filling this gap, studying the adoption of AI-based automation jointly in engineering services firms and their manufacturing customers.

Our simulations generate results that resonate with insights from economic history. First, AI-based automation, like general purpose technologies before, is adopted gradually. It starts at a slow pace, and accelerates after reaching a critical mass of adoption. Second, switching costs on the user side is the most important factor holding back the adoption of new technology. Third, technology does indeed destruct jobs, but it also generates new high-skilled jobs in the technology-using sectors. Finally, our simulations generate a boom and bust cycle on the supply side of the technology sector, which resembles what we have observed in the past, for instance during the *dot.com* bubble. This is not often observed in the literature and is thus an important contribution to new insight.

A policy implication of our findings is that innovation policy is not enough to foster technical progress. New technology is of little use if it is not adopted. We find that the early adopters are the largest and most productive manufacturing firms and that network effects of technology adoption can be strong. Furthermore, we find that adoption of AI-based automation is associated with demand for more skilled labour in using sectors. Policies aiming at fostering technical progress therefore need to focus more on switching costs on the user-side and on education and skills to make sure that the potential users of new technology can find the skills needed to restructure production around the technology.

The importance of the demand-side also suggest that occupational licensing does not necessarily constitute a drag on technology adoption as long as at least one engineering firm offers software. However, if exclusive rights to offer a service extends to software that automates the same service, the license fee is likely to be higher than in a competitive market, and the adoption rate may be substantially slowed down.

Finally, our results are relevant for other occupations and sectors. First, AI-enabled automation software in engineering is also relevant for the construction sector in a similar manner as in manufacturing. Second, other high-skilled business services occupations such as architects and management consultants face similar technological changes as the ones simulated here for engineering. Although these professions are currently way behind engineering in using AI-based automation, they are susceptible to such automation in the future. The accelerated digital transformation during the Covid-19 crisis may, however, have brought us closer to the steep part of the adoption curve for some of these services. Developments in the engineering sector modelled in this paper could thus be a harbinger of things to come in other professions going forward.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

HN developed the economic model, FK implemented and experimented with the ABM. Both authors together discussed simulation results, model updates and assumptions built into the implementation. Both authors contributed to the text, including result presentation.

ACKNOWLEDGMENTS

The authors would like to thank Gaurav Nayyar and three referees for useful comments and the Örebro University AI-Econ Lab for discussions and support.

REFERENCES

Acemoglu, D., and Restrepo, P. (2018). The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment. *Am. Econ. Rev.* 108, 1488–1542. doi:10.1257/aer.20160696

Andrews, D., Criscuolo, C., and Gal, P. N. (2015). Frontier Firms, Technology Diffusion and Public Policy: Micro Evidence from OECD Countries. OECD Productivity Working Papers, 2015-02. Paris: OECD Publishing. doi:10.1787/5jrql2q2jj7b-en

Autor, D. H., Levy, F., and Murnane, R. J. (2003). The Skill Content of Recent Technological Change: An Empirical Exploration. *Q. J. Econ.* 118, 1279–1333. doi:10.1162/003355303322552801

Baldwin, R., and Forslid, R. (2020). Globotics and Development: When Manufacturing is Jobless and Services are Tradable. Tech. Rep. Cambridge, MA: National Bureau of Economic Research. doi:10.3386/w26731

Berman, E., Bound, J., and Machin, S. (1998). Implications of Skill-Biased Technological Change: International Evidence. *Q. J. Econ.* 113, 1245–1279. doi:10.1162/003355398555892

Bessen, J. E., Impink, S. M., Reichensperger, L., and Seamans, R. (2018). *The Business of Ai Startups*. Boston: Boston Univ. School of Law, Law and Economics Research Paper. doi:10.3386/w24235

Bessen, J. (2002). Technology adoption Costs and Productivity Growth: The Transition to information Technology. *Rev. Econ. Dyn.* 5, 443–469. doi:10.1006/redy.2001.0152

Bresnahan, T. F., Brynjolfsson, E., and Hitt, L. M. (2002). Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence. *Q. J. Econ.* 117, 339–376. doi:10.1162/003355302753399526

Brynjolfsson, E., and Hitt, L. M. (2000). Beyond Computation: Information Technology, Organizational Transformation and Business Performance. *J. Econ. Perspect.* 14, 23–48. doi:10.1257/jep.14.4.23

Brynjolfsson, E., Rock, D., and Syverson, C. (2019). "Artificial intelligence and the Modern Productivity Paradox", in *The Economics of Artificial Intelligence an Agenda*. Editors A. Agrawal, J. Gans, and A. Goldfarb (Chicago: University of Chicago Press) 23–57.

Comin, D., and Hobijn, B. (2010). An Exploration of Technology Diffusion. *Am. Econ. Rev.* 100, 2031–205910. doi:10.1257/aer.100.5.2031

Davies, S., Davies, G., et al. (1979). *The Diffusion of Process Innovations*. (Cambridge, United Kingdom: CUP Archive).

Dawid, H. (2006). Chapter 25 Agent-Based Models of Innovation and Technological Change. *Handbook Comput. Econ.* 2, 1235–1272. doi:10.1016/s1574-0021(05)02025-3

D. D. Gatti, G. Fagiolo, M. Gallegati, M. Richiardi, and A. Russo (2018). *Agent-Based Models in Economics - A Toolkit*. (Cambridge, United States: Cambridge University Press).

Doms, M., et al. (2004). The Boom and the Bust in information Technology investment. San Francisco, United States: Economic Review-Federal Reserve Bank of San Francisco, 19–34.

Farmer, J. D., and Foley, D. (2009). The Economy Needs agent-Based Modelling. *Nature* 460, 685–686. doi:10.1038/460685a

Feenstra, R. C. (2018). Restoring the Product Variety and Pro-competitive Gains from Trade with Heterogeneous Firms and Bounded Productivity. *J. Int. Econ.* 110, 16–27. doi:10.1016/j.jinteco.2017.10.003

Felten, E. W., Raj, M., and Seamans, R. (2019). *The Occupational Impact of Artificial Intelligence: Labor, Skills, and Polarization*. New York, NY: NYU Stern School of Business.

Gallegati, M., and Richiardi, M. G. (2009). "Agent Based Models in Economics and Complexity," in *Encyclopedia of Complexity and Systems Science*. Editor R. A. Meyers (New York, NY: Springer New York), 200–224. doi:10.1007/978-0-387-30440-3_14

Geroski, P. A. (2000). Models of Technology Diffusion. *Res. Pol.* 29, 603–625. doi:10.1016/s0048-7333(99)00092-x

Gort, M., and Klepper, S. (1982). Time Paths in the Diffusion of Product innovations. *Econ. J.* 92, 630–653. doi:10.2307/2232554

Hall, B. H., and Khan, B. (2003). Adoption of New Technology. *Tech. Rep*. Cambridge, United States: National bureau of economic research. doi:10.3386/w9730

Hamill, L., and Gilbert, N. (2016). *Agent-Based Modelling in Economics*. New York, United States: John Wiley & Sons.

Klügl, F., and Bazzan, A. L. C. (2012). Agent-based Modeling and Simulation. *AIMag* 33, 29. doi:10.1609/aimag.v33i3.2425

Lasi, H., Fettke, P., Kemper, H.-G., Feld, T., and Hoffmann, M. (2014). Industry 4.0. *Bus Inf. Syst. Eng.* 6, 239–242. doi:10.1007/s12599-014-0334-4

Milgrom, P., and Roberts, J. (1990). The Economics of Modern Manufacturing: Technology, Strategy, and Organization. *Am. Econ. Rev.* 80, 511–528.

Nelson, R. R., and Phelps, E. S. (1966). Investment in Humans, Technological Diffusion, and Economic Growth. *Am. Econ. Rev.* 56, 69–75.

Neves, F., Campos, P., and Silva, S. (2019). Innovation and Employment: An agent-Based approach. *J. Artif. Societies Soc. Simulation* 22. 8. doi:10.18564/jasss.3933

Rock, D. (2019). Engineering Value: The Returns to Technological Talent and investments in artificial intelligence. Available at SSRN 3427412.

Rogers, E. M. (1995). Diffusion of innovations: Modifications of a Model for Telecommunications. *Die diffusion von innovationen in der telekommunikation*. New York, United States: Springer, 25–38. doi:10.1007/978-3-642-79868-9_2

Rosenberg, N. (1972). Factors affecting the Diffusion of Technology. *Explorations Econ. Hist.* 10, 3–33. doi:10.1016/0014-4983(72)90001-0

Stoneman, P. L., and David, P. A. (1986). Adoption Subsidies vs information Provision as instruments of Technology Policy. *Econ. J.* 96, 142–150. doi:10.2307/2232977

Tesfatsion, L. (2006). "Chapter 16 Agent-Based Computational Economics: A Constructive Approach to Economic Theory,". *Handbook of Computational Economics*. Editors L. Tesfatsion and K. L. Judd. 1 edn (Philadelphia, New York: Elsevier), 2, 831–880. doi:10.1016/s1574-0021(05)02016-2

Varian, H. (2019). *16. Artificial Intelligence, Economics, and Industrial Organization*. Chicago, United States: University of Chicago Press.

Wang, J., Ma, Y., Zhang, L., Gao, R. X., and Wu, D. (2018). Deep Learning for Smart Manufacturing: Methods and applications. *J. Manufacturing Syst.* 48, 144–156. doi:10.1016/j.jmsy.2018.01.003

Wooldridge, M. (2009). *An Introduction to Multiagent Systems*. 2nd Edn. New York, United States: John Wiley & Sons.

Zolas, N., Kroff, Z., Brynjolfsson, E., McElheran, K., Beede, D. N., Buffington, C., et al. (2021). *Advanced Technologies Adoption and Use by US Firms: Evidence From the Annual Business Survey. Tech. Rep*. New York, United States: National Bureau of Economic Research.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Robotics and AI

ORIGINAL RESEARCH
published: 16 June 2021
doi: 10.3389/frobt.2021.654298

Robot Care Ethics Between Autonomy and Vulnerability: Coupling Principles and Practices in Autonomous Systems for Care

Alberto Pirni[1*†], *Maurizio Balistreri*[2†], *Marianna Capasso*[1†], *Steven Umbrello*[2†] *and Federica Merenda*[1†]

[1]*Sant'Anna School of Advanced Studies, Pisa, Italy,* [2]*Department of Philosophy and Educational Science, University of Turin, Turin, Italy*

OPEN ACCESS

Edited by:
Martim Brandão, King's College London, United Kingdom

Reviewed by:
David Gunkel, Northern Illinois University, United States
Alistair Niemeijer, University of Humanistic Studies, Netherlands

***Correspondence:**
Alberto Pirni
alberto.pirni@santannapisa.it

These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Ethics in Robotics and Artificial Intelligence, a section of the journal Frontiers in Robotics and AI

Received: *15 January 2021*
Accepted: *03 June 2021*
Published: *16 June 2021*

Citation:
Pirni A, Balistreri M, Capasso M, Umbrello S and Merenda F (2021) Robot Care Ethics Between Autonomy and Vulnerability: Coupling Principles and Practices in Autonomous Systems for Care. Front. Robot. AI 8:654298. doi: 10.3389/frobt.2021.654298

Technological developments involving robotics and artificial intelligence devices are being employed evermore in elderly care and the healthcare sector more generally, raising ethical issues and practical questions warranting closer considerations of what we mean by "care" and, subsequently, how to design such software coherently with the chosen definition. This paper starts by critically examining the existing approaches to the ethical design of care robots provided by Aimee van Wynsberghe, who relies on the work on the ethics of care by Joan Tronto. In doing so, it suggests an alternative to their non-principled approach, an alternative suited to tackling some of the issues raised by Tronto and van Wynsberghe, while allowing for the inclusion of two orientative principles. Our proposal centres on the principles of autonomy and vulnerability, whose joint adoption we deem able to constitute an original revision of a bottom-up approach in care ethics. Conclusively, the ethical framework introduced here integrates more traditional approaches in care ethics in view of enhancing the debate regarding the ethical design of care robots under a new lens.

Keywords: applied ethics, care ethics, bioethics, robotics, care robots, autonomy, vulnerability

INTRODUCTION

Developments in robotics and automation technologies are rapidly changing many aspects of our lives. The field of (health) care has been no exception, promising many boons while also bringing about controversial ethical questions. This paper takes care robots for the elderly as an object of analysis, evaluating the existing literature on their ethical and responsible design. In particular, we aim to discuss the existent approach to the ethical design of care robots by Aimee van Wynsberghe (2012); van Wynsberghe (2013a); van Wynsberghe (2016) that relies principally on the work on care ethics by Joan Tronto while also exploring the viability of a care ethics approach that is fundamentally non-principled, such as those expounded by Tronto (1993), Tronto (2010) in view of possibly envisaging a conciliation between the two alternative proposals (§ 1).

Tronto argues that general principles are too broad to constitute a sufficiently stable justification for actions consequent to care ethics. However, in recent years, the literature on care ethics has been trying to identify principles that can have an informative and a justificatory role in making moral decisions and carrying out moral actions in care practices (Collins, 2015) (§ 2). Though such an approach constitutes just one of the many different understandings of care ethics, it becomes

particularly relevant as a theoretical basis for care robot programming, that is when the ultimate goal of philosophical research is to program machines able to interact with human beings in ways that are acceptable from a care ethics perspective. To this end, this paper explores the possibility of integrating 1) a care-ethical perspective based on the consideration of particular care relationships, their contextual levels and the importance of needs, emotions and sympathetic modes of deliberation with 2) a principlist approach to care.

Such an approach asserts that orientative principles, rather than constitutive ones, may have a justificatory role in grounding proper forms of action and would constitute one to be included in the category of the so-called "hybrid approaches" similar to the one proposed by Van Rysewyk and Pontier (2015) but with substantial differences that will be explored. According to a Kantian approach, the proposed principles are argued to be not mutually exclusive and contribute to identifying a more comprehensive account of care ethics (§ 3).

In our view, this approach to care ethics can be operationalized through an innovative account of two basic orientative principles and their systemic interrelation: autonomy, on the one hand, and the principle of vulnerability, on the other (§ 4). If successful, the practical implications of this approach pave the way for a revision of how care ethics is treated within the domain of engineering and design and subsequently a reimagination of how to translate these types of orientative principles into tangible design requirements. And this las point is a fundamental point with regards to the "design turn in applied ethics" (van den Hoven et al., 2017) given that the traditional top-down and bottom-up approaches have dominated the field of robotic design. Still, this paper does not delve into this issue, but rather provides the conceptual framework to springboard new discussions in engineering ethics for how to go about designing care robots according to the approach we discuss below.

ISSUES AND APPROACHES TO DESIGNING CARE ROBOTS

The Top-Down Approach

This article aims to analyse how care robots–i.e., machines used in care practice–can be designed to support and promote the fundamental values in care practices. There is already a wide variety of autonomous machines currently used in assistance and care: My Spoon is a robot able to spoon-feed an assisted person, Sanyo, to wash and rinse him. Further, robots such as RIBA (Robot for Interactive Body Assistance) can move patients from one place to another, while Care-o-bots do likewise with objects in a room. And as well as robots to monitor people's health and wellbeing, there are nursebots, used to reminds the elderly of certain routine activities (from eating and drinking to taking medicine and washing their teeth) and accompany their movements within a space–while Pepper, NAO, Kabochan, Brian 2.1, and Nexi 2 are humanoid robots that can not only move their arms, dance, and answer questions but also gather information through a camera and microphone and entertain the assisted person with basic games. This section intends to discuss the top-down and bottom-up strategies to design artificial moral agents (AMAs). "Top-down approaches to this task involve turning explicit theories of moral behaviour into algorithms. Bottom-up approaches involve attempts to train or evolve agents whose behaviour emulates morally praiseworthy human behaviour" (Allen et al., 2006: 149).

The "top-down" approach may look the easiest from an engineering perspective because it consists of programming the machine according to general behavioural principles (or laws). As noted by Van Rysewyk and Pontier (2015), such an approach is particularly apt to operationalize utilitarian or deontological ethical perspectives. It also follows a long-standing moral tradition that identifies the correct behaviour with that conforming to the law. Asimov's three laws of robotics are an example of this type of solution, in that they bind the machine to act according to general principles at all times (Anderson, 2008). Some attempts to program robots to be "good" using a principlist approach have been reported in the literature. Winfield et al. (2014) discuss research in which robots are programmed to achieve a goal and to prevent other robots (as proxy humans) from getting hurt (for example, by falling into a hole or ending up in a dangerous area). According to Winfield et al. (2014, p. 5), this is an example of a robot that "appears to match remarkably well with Asimov's first law of robotics: A robot may not injure a human being or, through inaction, allow a human being to come to harm." Arkin (2009) have also have proposed a moral system able to adhere to the International law of war (LOW) and rules of engagement (ROE) and to distinguish between unethical and ethical actions based on their compliance or not with international law (Arkin 2009, p. 1).

The problem with a "top-down" approach is that the laws or general principles constitute overly generic moral references which may be hard to apply (or interpret) in complex real-life situations. Furthermore, each case is different and cannot be reduced *a priori* to law, which must be interpreted. Thus, an intelligent robot limiting itself to apply the instructions rigidly has received risks to interpret them inadequately and to the letter. For example, suppose we program a robot to serve,[1] obey, and protect human beings. In that case, this could have disastrous consequences for humanity, in that they could think they are morally obliged to stop us from doing anything–because the less we do, the fewer chances we have of getting hurt–or to inflict serious cerebral damage on us as well, so that we perceive less pain.

Further, to safeguard and promote a patient's wellbeing, a robot programmed to carry out care activities may feel justified in violating their personal sphere and refusing to obey them and meet their needs. For example, a robot could inform the health operators of their patient's intention to put an end to their life, or not even help them die after an explicit request, even when the patient's existence has become unbearable (Tonkens, 2015: 207–222). Furthermore, it merits questioning whether a care robot programmed to promote a patient's good would stop

[1]Which can also be understood in the negative sense, i.e., allowing for a certain degree of (negative) freedom to do risky things.

the surgeon about to operate on them (Wallach and Allen, 2015: 92) or deceive them about their health when the prognosis is terminal.

Finally, the greater the number of rules and moral principles to be respected by the moral agent, the higher the risk that, in certain situations, two or more principles conflict. To this end, we should design autonomous machines to confront these situations and know what–in the case of conflict–is the principle that must prevail because it is "stronger." That is, these types of dilemmas, or moral overload, create inextricable computational roadblocks. The way to resolve such issues computationally is to have higher-order principles that can be used to address such dilemmas (Allen et al., 2006: 150; Goodall, 2014). As new scenarios and new circumstances present themselves, we should add new moral rules and principles (that can resolve the existing conflicts) and specify their application. But this would not be a solution either. Even if it is always possible to determine when one norm has precedence over another, we cannot imagine all possible scenarios. "So even if deontological ethics can provide a guide in many situations, it cannot be used as a complete ethical system, due to the incompleteness of any group of rules and the difficulty of articulating human ethics in its complexity in a list of rules" (Goodall, 2014). That is, autonomous systems could end up in situations that were not foreseen nor even foreseeable. Moreover, even if we were able to formulate explicit criteria allowing an artificial moral agent (AMA) to override a rule, "any such criteria would very likely produce other dilemmas" (Wallach and Allen, 2015).

For this reason, it seems preferable to find/adopt a single moral principle or more general and abstract principles from which all other particular principles (or rules) can be derived. For example, utilitarian ethics stating that it is right to maximize happiness, wellbeing, preferences (or informed preferences), or pleasure for the greatest number of those involved. A strength of utilitarianism is its apparent ability to quantify goods and harms. The issue is that calculations could be highly complex and that an engineering model of ethics (i.e., do you have a dilemma? Apply the principle) is inadequate for moral life. Imagine, for example, that we wish to programme machines according to utilitarian ethics. We want "intelligent" care robots to promote the patients' wellbeing as much as possible, but how can wellbeing be calculated? Only in terms of life years (the more, the better)? Or do we also consider economic, psychological, and social wellbeing? Significant disagreements could emerge regarding the objective to maximize, in that some would think it proper to maximize "happiness" while others would view it right to maximize "wellbeing." Still, others may consider maximizing "preferences" (or "informed preferences") while others again could give value just to "pleasure."

Further, to calculate wellbeing, should we bear in mind only the most immediate consequences or take long-term effects (but how far away?) into consideration? This would also mean deciding which perspective the intelligent care robots should assume: whether to protect the interests of the owner or user (that is, the patient) at all times or consider a more stable, general point of view and protect the collective interest. Imagine, for example, that a patient finds themselves at home and is about to have guests over. Suddenly, a fire breaks out in their kitchen and spreads into the other rooms of the house: who must the care robot save? If we consider things morally, it seems fitting that the robot should not only worry about its owner but assume a general point of view: my wellbeing counts, but so does anyone else's. However, still, certain decisions could end up as counterproductive. For example, programming care robots to protect general interests could reduce (or even extinguish) interest in these machines and slow down their adoption (at least in the short term). The most appropriate solution seems to minimize the consequences, considering the interests of both the care robot's owners and anyone else (i.e., the general interest). But there is no single way to bring together and balance these different interests: we could give more value to the general interest, do the opposite, or consider them equally.

Utilitarianism presents the same problems as the so-called deontological ones. They appear to be unable to derive from highly abstract, generic premises; solutions can orientate people in concrete situations when choosing (Williams, 2011). Utilitarianism seems to permit the construction of scientific ethics. Still, the same precise and rigorous conditions cannot be achieved in practice as in science. Further, an engineering model of morality does not solve but intensifies moral contrasts in that it reduces all differences to divergences that are of principle disjunctions, hence harder to overcome. That is, "taking for granted that the only way to face ethical disputes is to apply this deductive and axiomatic model of practical rationality ends up making it almost impossible to overcome disagreement" (Lecaldano, 2005: 16).

The Bottom-Up Approach

The alternative to the top-down model is the one we have called bottom-up: here, the strategy to make machines moral does not consist in giving them laws or general principles. The bottom-up approach allows artificial intelligence to learn morality (that is, what ethical behaviour is) through experience and learning without the need for general principles. Allen et al. (2006) liken this approach to the way a human child learns, and van Rysewyk and Pointier define it as an approach that creates "a series of learning situations through which a machine works its way toward a level of ethical understanding acceptable by the standards humans define" (2015: 99). Small pieces of knowledge gained through experience, manipulated by programmers as new challenges and tensions arise, all done within a learned social context in which the AMA is situated and able to grow (Allen et al., 2006: 151).

According to Aimee van Wynsberghe, the most promising manner of programming a care robot based on an alternative approach to the top-down one is to integrate the traditional value sensitive design (VSD) approach with normative criteria and elements from care ethics (van Wynsberghe, 2015; van Wynsberghe, 2016). Care ethics appears as an alternative perspective to the strategies inspiring a lot of modern moral philosophy based on an appeal to universal, abstract, and impersonal rules (principles that may be assumed to be valid for humanity overall), which should regulate the behaviour of separate, independent individuals. According to care ethics, it is

wrong to reduce the moral to merely obeying norms and principles imposed on real life and people's experience (Tronto, 1993; Botti, 2015; Collins, 2015). A long-standing philosophical tradition has linked morality with the ability to assume a completely detached perspective from our particular interests, the concreteness of the situation and relations, and the ties in which we are involved. Nevertheless, morality means not conforming to or applying general principles (or moral laws). Still, rather it corresponds to the ability to develop dispositions and practices of care and attention towards others (considered not as abstract individuals, but perceived in their concreteness and particularity). That is, from this perspective, morals are not the mere execution of a mechanical task (the simple application of a law or general principle to a particular case), but a practice requiring sensitivity (that is, attention)–and of course empathy–towards others. Likewise, it is also the subject's awareness of their relation to others and developing an ability to listen and sentimental communication (Gilligan, 1982; Botti, 2015; Collins, 2015).

The considerations above should suffice to highlight that care ethics involves the importance of connection (Noddings, 1984; Noddings, 2002) and the relationship between individuals, their choices, and the context where they find themselves situated (Botti, 2018: 16). Dependence is the sign of vulnerability, but it is possible through this dependence to feel a responsibility of care towards one's neighbour and to be able to pay more attention towards others (Gilligan, 1982; Noddings, 1984; Tronto, 1993; Held, 2006). Similarly, care is not perceived as a pre-set ethical perspective, ready for use in any context, but, as Wynsberghe (1,025) argues, it is a starting lens to recognize the other person's dignity and begin to look after one's neighbour. Care is already current practice in our lives, from birth to the moment of death (Noddings, 1984; Held, 2006). So, according to the supporters of ethical care, there is no need to justify it, but rather to take its importance into account to place it at the centre of morality (Botti, 2015).

In this view, starting from Gilligan's original formulations, care ethics is not to be conceived as something which regards female subjects in as much as they can become mothers: it does not correspond to "maternal" ethics, but it does constitute a valid moral paradigm for any person (Slote, 2010; Slote, 2011). According to Virginia Held, care is first and foremost a form of emotional, reflective commitment (Held, 1993)–including sensitivity, solicitude, and worry, but also empathetic responsiveness, attention to specific needs and contexts, as well as relationships–which of course has a biological basis. However, this basis cannot be described as an animal- and/or human-type "instinct." It is a social practice, a cultural transformation of something we already find in maternal care. Attention is sensitivity, which means that they are central aspects in responding to others' needs. But these "natural" abilities may and (within an appropriate care process) must be refined and corrected through communication and dialogue exchanges (Held, 2006). Held does not dwell on this process of refining moral sensitivity. Still, from her perspective, emotions such as sympathy, empathy, sensitivity, and responsiveness are moral emotions that any individual should learn to cultivate to approach other people's condition and do that which morality recommends. Maternal relationships may be the starting point for care activities, and, for Held, they remain an authoritative reference model. However, the activity of caring for another is only learnt through practice and experience. In this way, Slote (2010); Slote (2011) maintains, we may also develop a broadened, mediated empathy which stands in relation to situations not immediately present, through which we can even imagine the feelings of those farthest away.

In Joan C. Tronto's opinion, care activity–"[o]n the most general level, we suggest that caring be viewed as a species activity that includes everything that we do to maintain, continue, and repair our "world" so that we can live in it as well as possible. That world, which includes our bodies, ourselves, and our environment, all of which we seek to interweave in a complex, life-sustaining web" (Tronto, 1993: 103)–plays out in four phases:

1. "Caring about," where we recognize that care is necessary and we perceive the existence of a person's need that must be satisfied. "Caring about–writes Tronto–will often involve assuming the position of another person or group to recognize their needs" (Tronto, 1993, p. 106).
2. "Taking care of," that is, the moment in which we assume responsibility towards that need and consider what can be done, bearing the situation in mind. There is, then, recognition of the possibility of acting towards the identified need (Tronto, 1993: 106–107)
3. "Caregiving" is committing ourselves to satisfy the need through work requiring that the one giving care comes into contact with its recipients (Tronto, 1993: 107).
4. The fourth phase of the process or care activity is the "receiving care", because care should also be measured in terms of appropriateness of the basis of the response from its recipient: "[u]nless we realize that the object cared for responds to the care received, we may ignore the existence of these dilemmas, and lose the ability to assess how adequately care is provided" (Tronto, 1993: 108).
5. There is, finally, a fifth phase of care–that is, *caring with*–which is specific to a democratic society in which the citizens are constantly involved in taking care, not individually–as autonomous, self-sufficient subjects–but together with other people as vulnerable subjects who need care and can trust and rely on other people. Care (the activity of care) is present in any society. Still, in a democratic society–writes Tronto–we have the best care activity because only in a democratic society is it possible to have institutions promoting caring *with* (Tronto, 2013; pp. 154–155).

These moments correspond to specific moral qualities: attentiveness, responsibility, competence, responsiveness, and trust (and solidarity). Attention is required because the caregiver must have the ability to perceive the continual changes in the situation and the needs of the person they are taking care of, in that there can be no care unless there is attention to others' needs. Recognizing others' needs is a challenging task, but this is precisely why it is a moral element, and ignoring others' needs is without any doubt a moral evil (Tronto, 1993: 127),

whilst responsibility is the ability to take on others' perceived needs: it is not a promise; nor is it a commitment to act according to pre-set formal rules. In other terms, it is the ability to recognize that we must, based on the role we occupy and the skills we have, do something to change other peoples' situation. Also, competence is the ability to consider the effectiveness of our actions, because "clearly, making certain that the caring work is done competently must be a moral aspect of care if the adequacy of the care given is to be a measure of the success of care" (Tronto, 1993:133). That is, Tronto states, truly responsible (and appreciable) cannot be uninterested in the consequences, because–she adds–"from a perspective of care, we would not permit individuals to escape from responsibility for their incompetence by claiming to adhere to a code of professional ethics" (Tronto, 1993: 134).

Then, responsiveness marks the importance of the care recipient's response and the caregiver's duty to pay attention to the "responses" of those cared for. Good care–writes Tronto–requires the four care process phases and appropriate integration of the different skills, or rather, moral elements necessary to perform it: "Such an integration of these parts of caring into a moral whole–states Tronto–is not simple. Care involves conflict; to resolve this conflict will require more than an injunction to be attentive, responsible, competent, and responsive" (Tronto, 1993: 136). Finally, trust results from people's awareness that they can count on others' participation in their care and care activities. At the same time, solidarity is built when citizens know that they can dispense care with others better (Tronto, 2013).

Following the care ethicist Joan Tronto (1993), Aimee van Wynsberghe identifies four fundamental values of care to be promoted in the design of (autonomous) systems: 1) attentiveness, as the capacity to recognize the needs of the care-receiver; 2) responsibility, which implies the caregiver's concern for meeting the needs of the care-receiver; 3) competence, as the capacity of executing an action to fulfill the needs of the care-receiver; and 4) responsiveness or reciprocity, as the capacity of the care receiver to guide the caregiver and the instauration of a reciprocal interaction (van Wynsberghe, 2012; van Wynsberghe, 2013a; van Wynsberghe, 2013b; van Wynsberghe, 2016). Van Wynsberghe insists that these four elements are crucial in any care practice that impacts caregivers and care receivers due to the ethical importance they assign to the relationship and distribution of roles and responsibilities (Tronto, 2010). We may add a fifth element, including trust and solidarity as the caregiver's ability to collaborate with others, in a democratic society, in care activity, enjoying trust in the willingness to participate, and collaborate with other people (van Wynsberghe, 2021).

Artificial Morality and Moral Training

What we are asking is how to make care robots sufficiently moral, a question that has been at the centre of ethical concerns already for some time and widely disseminated by individuals like David Gunkel, who has reconstructed the different perspectives developed in philosophy over the relation between morality, artificial intelligence and robotics (Gunkel, 2012). Building on what we have discussed thus far, we can conclude that their design must consider both attentiveness and empathy in whichever care practice setting.

However, robots are machines: we can make them increasingly intelligent (and hence able to respond–or to react–to stimuli of human beings more and more appropriately), but they remain incapable–at least for the moment–of "sympathizing" with others' needs and interests. So, this questions the possibility of building robots able to perform care activity integrating a traditional VSD (Value Sensitive Design) approach with elements from a care ethics perspective. Starting from Gilligan, care ethics has underlined that appropriate care activity does not consist of the ability to detach or abstract ourselves from the particular context or take distance from the actual features rationally, but instead of developing sensitivity and solicitude towards other people. And it is that attention to others' specific, particular needs, of which Tronto also speaks, which is, in this view, reached through practice, experience, and specific abilities: the willingness to listen to the other and communicate with them (Tronto, 2013; Tronto, 2015). But, as Gilligan states (1982), sentimental communication, that is, empathy or sympathy, is also needed (Noddings, 1984; Noddings, 2002). Meaning the ability to let ourselves be influenced by other people's emotions and feelings. For this reason, robots' lack of any moral sensitivity seems to exclude from the start the possibility of attributing a minimal form of moral ability to them (Dumouchel and Damiano, 2017).

Yet, the fact that robots are not capable–at least at present–of feeling sympathy is not problematic: as suggested by Coeckelbergh (2020), while against the wishes of both cognitivists and feeling theorists of emotions until robots will have a consciousness, they will not be able to feel emotions properly, what Coeckelbergh calls "the appearance of emotions" and of being entirely moral can be attained. This is because we can program such sentiments synthetically rather than biologically. Even though others' feelings cannot substantially influence a machine, it could learn to modify its behaviour and emotions based on others' reactions and approval (or disapproval), developing this way a synthetic kind of nature-nurture interaction which resembles that building up to human moral development (Coeckelbergh, 2020). In this way, we may expect it to become a moral (or virtuous) machine over time; able, that is, to take into account not only the present situation and people but also the needs and interests of those it interacts and relates with, taken on in their particularity. The sentiments of love (or esteem) and hate (or contempt) for our fellows–David Hume (2007) and Adam Smith (1976) outline in their works–are the most potent motors of morality, in that we want others to like us and appreciate our behaviour and the passions and sentiments we have (Baier, 1991). Robots might be intelligent, but they are insensitive to others' reactions because they are not conscious. However, they could be programmed to regulate their behaviour (and their "sentiments") based on the esteem or contempt they receive from others. For example, moral approval (expressed by human beings) could be a reason to repeat specific behaviour, and moral disapproval a cause to modify/change it or make it more acceptable. We could call this ability to adapt "synthetic

sensitivity" in that, like "biological" sensitivity, it denotes a disposition to put oneself in tune with the sentiments of other people, which is, as van Wynsberghe recalls, the essential quality for a good care provider: "Being in tune with the delicacy of the situation, and how to address it, can also be referred to as ethical sensitivity or "tinkering." The former adheres to the idea of care as caring about while the latter is closely linked with care as caring for, albeit they are not mutually exclusive" (van Wynsberghe, 2015: 35).

Stating that robots can only become moral through practice and not thanks to abstract training simulations like those proposed in the Silicon Coppélia experiment mentioned by Van Rysewyk and Pontier, (2015)– find themselves, that is, interacting with people and becoming subject to their moral evaluation–means suggesting a different model of morality from that indicated by those who think it is possible to make a machine moral simply by programming it to obey certain principles.

This is not a defect but a strength in the approach we are suggesting. It recognizes the importance of experience and education for moral training and the inappropriateness of reducing the complexity of moral life to a few principles.

As announced at the very beginning, in our view, it is possible to integrate a 1) care-ethical perspective based on the consideration of particular care relationships, their contextual levels and the importance of needs, emotions, and sympathetic modes of deliberation with 2) a principlist approach to care. These two perspectives are not mutually exclusive, as it has been thought. They contribute to identifying a more comprehensive account of care practices that can be operationalized through an innovative interpretation of two fundamental and orientative principles and their systemic interrelation: the principle of vulnerability, on the one hand, and the principle of autonomy, on the other.

In Tronto's words, care is not only an activity, but also a flair: "we insist that the activity of caring is largely defined culturally, and will vary among different cultures. Fourth, we see caring as ongoing. Care can characterize a single activity, or it can describe a process. In this regard, caring is not simply a cerebral concern or a character trait but the concern of living, active humans engaged in the processes of everyday living. Care is both a practice and a disposition" (Tronto, 1993: 103–104). For the Aristotelian ethics of virtue, dispositions appear as functions or abilities belonging to human nature.

In contrast, sentimentalist ethics consider the dispositions as individual character traits that are subject to approval or disapproval due to the consequences they produce. Referring, then, to Julia Driver's definition, we can state that, "A character trait is a moral virtue if it is a disposition to produce (i.e., it tends to produce) intentional action that is systematically productive of the good (Driver, 2001:107). In other terms, dispositions are those personality or character traits that do not end in action because they represent principles, that is, stable conduct motifs (Baier, 1991; Baier, 1995): so we may call them qualities characterizing a person's character or mind. (Hume, 2007: 3.2.1.2.).

A psychological disposition is made up of accepting a distinctive fan of considerations as reasons for action and a tendency to have a certain feeling or combination of emotions, often driving us to action. A robot cannot be moved by certain feelings–nor by combining feelings and passions–but it can still be programmed to act based on particular orientative principles and consequently manifest the disposition to behave in the way we prefer. On the basis, that is, of the ethical conception we are referring to, the ideal would be to have a care robot with the necessary sensitivity to respond appropriately to the feelings and emotions of the people he is called to care for. Indeed, for a robot to empathize with the people it interacts with, it would be easier to establish how to discharge its tasks. Yet, at least for the moment, hoping to build a robot endowed with sensitivity and our empathetic ability is unthinkable. Given the impossibility of counting on a compassionate robot, we can–as we have already said–consider making it synthetically (or artificially) empathetic through a programme allowing it to respond considering others' judgement. But beyond this, programming it with orientative principles, we could also attribute to it a disposition "to be interested, look after and provide care when there is an unsatisfied need". If the robot could have a character, we would not need to programme it with principles: but the robot cannot have a character, so the orientative principles could allow us to influence/condition its character sufficiently appropriately for our needs. (Dumouchel and Damiano, 2017). From a practical perspective, programming a machine to follow a few orientative principles could be advantageous. It would permit not only to control the machine's behaviour but also to limit its autonomous space.

Furthermore, the robot would not need to learn to behave from scratch, in that it would already be programmed to follow certain principles, hence ways of behaviour (Allen et al., 2006; Wallach and Allen, 2015: 114–115). Nor would there be the problem characterizing "bottom-up" learning, which can be a prolonged, mistake-ridden process (Van Rysewyk and Pontier, 2015). There is still the risk that the overall principles are too general or poorly interpreted or that the robot does not know how to behave. Yet, in the terms described above, a machine sensitive to the reactions and responses of those it interacts with would be less subject to these problems. It could learn from practice and experience to correct its behavior. So even though it may misapply the moral principles, it could still always correct itself, taking into account the reactions of those it interacts with. Both Held (1993), Held (2006), Tronto (1993) stress the difficulty in grasping other people's need for care, and for this reason, they emphasize the importance of dialogue and communication, as well as–naturally–the finetuning of our empathic abilities. Our capacity to take care of others' interests and needs is limited: sometimes we do not perceive their suffering nor realize that we are causing them harm; certain forms or ways of life are invisible or at least remain opaque. A robot could 'be born' with our same defects, but, like us, could still, through experience/practice, become an appreciable "person".This demonstrates that it is possible, in the case of robots, to integrate 1) a care-ethical perspective based on the consideration of particular care relationships, their contextual levels and the importance of needs, emotions and sympathetic modes of deliberation with 2) a principlist approach to care. These two perspectives are not mutually exclusive, as it has been

thought, and contribute to individuate a more organic account of care practices which can be operationalized through an innovative understanding of two basic principles and their systemic interrelation: the principle of vulnerability, on the one hand, and the principle of autonomy, on the other.

FUNDAMENTAL ORIENTATIVE PRINCIPLES WITHIN AND BEYOND THE CARE ETHICS APPROACH

Preliminary Orientative Lines

Based on the premises laid out in the preceding section, the main research question can be rephrased: how can we formulate a comprehensive approach that can frame the human-robot interaction overcoming the objective difficulties discussed above in terms of empathy? We are used to communicating such issues both from the point of view of human beings and machines. If we take the first issue discussed in the preceding section, what we are doing is referring to the possibility–in some cases welcomed, in others indeed feared–that specific groups of human beings might develop feelings for robots. More specifically, there is a typology of relationships that emerge, i.e., by persons with mental impairments or by elderly people with affective difficulties or, still, by persons addicted to robot companion and/or sex robots (Sharkey, 2014; Bendel, 2017; Balistreri, 2018; Ostrowski et al., 2019; Bisconti Lucidi and Piermattei, 2020; Jecker, 2020).

On the other hand, we refer to the objective difficulty of human-robot interaction by considering the second viewpoint: the tension between a robot executing a programmed behaviour and an empathetic behaviour towards a human subject. Of course, we can consider a situation in which the programmer is committed to programming a sort of *synthetic empathy* that can resemble or replicate, as much as possible, empathic behaviour that is observed in human-human interactions. Again, we are aware that this is one of the most promising frontiers of the intersection of big data enquiry and artificial intelligence (Cavallo et al., 2018).

We are thus perfectly aware of the existing and expanding debate on such issues. Nonetheless, it is not necessarily to expand on what are indeed promising theoretical directions in what follows. Instead, we would aim to propose a preliminary outline of what has been explored up to now. This is a sort of theoretical framework that, in a sense, acts as a foundation, and that thus we propose as one to be inserted into the care-ethics approach we presented above.

The final expected goal of this insertion (and of the entire research project which this paper is part of) is to propose a renewed care-ethics approach that shall integrate the VSD approach articulated by van Wynsberghe. In what follows, we will offer a comprehensive argument to ground the legitimacy of such a theoretical framework by justifying the insertion of two basic principles whose argumentative role is identifying an avenue along which trying to integrate a possible renewed program in care ethics–we attempted to articulate a preliminary attempt to frame the implementation issue of an integrated VSD in Umbrello et al. (2021). In turn, we would like to deepen the same linkage between that framework and the more traditional care ethics approach in a further step.

Methodological Remarks

In approaching such a framework, some preliminary and methodological remarks should be clarified.

Firstly, in line with the objections expressed above (§ 2.1), it merits reiterating that referring to two principles does not mean that we implicitly affirm a top-down approach. We remain convinced that such a top-down approach risks substantiating nothing other than general principles that do not offer concrete guidance in specific situations or actual settings.

Secondly, introducing a unique principle might imply operationalizing such a principle by underestimating the contradictory or aporetic effects derived from it. Some of them, related to the utilitarian principle of happiness, have been analyzed above. This tension can be exacerbated by introducing other relevant principles like "dignity" or "respect", for example. One can already imagine the conflict that would undoubtedly emerge through a haphazard combination of two or more principles. A striking exemplar would be between managing the moral overload between two principles like happiness and dignity.

Furthermore, introducing several (all fundamental) principles can't exclude the possibility of creating conflicts and dilemmas that are undoubtedly difficult to solve within specific operational situations. This is already true for human beings, but it remains even more evident for robots. Thus, being fully aware of the risks of implementation and the conflicts related to a top-down approach, we indicate two principles that must be framed not as antithetical but rather complementary.

To methodologically avoid affirming a top-down approach, we instead actively aim at constructing a *revision* of a bottom-up approach. This, of course, does not mutually exclude other approaches like the so-called "mid-level ethical theories" that have been proposed by van den Hoven (2010), Jacobs and Huldtgren (2018), and Cenci and Cawthorne (2020) following the theoretical path traced by the likes of Martha Nussbaum, Amartya Sen, and John Rawls. We have elsewhere employed such an approach in application to care robots (Umbrello et al., 2021). What characterizes this revision of a bottom-up approach is marked by a straightforward adoption of theories of relationality as they pertain directly to vulnerability and autonomy, something that those other approaches do not undertake. That is the line of argumentation that will be followed here. Nonetheless, some methodological remarks need to be explained in a revised bottom-up approach as the one we aim to construct here.

We introduce four remarks here. The first remark is related to the specific theoretical usage of the term "principle." Selecting care ethics as the primary focus implies some particular difficulties in inserting a discourse on principles since care ethics originated from the need to avoid an "ethical principle" in the traditional sense of the term. However, the meaning of the word "principle" we suggest and deem appropriate here is not to be understood as a normative term that is external/independent

to the situation and subsequently commands in the universalistic sense of the word. Simply put, we are not thinking of a principle that reproduces the same normative constraints that the Kantian categorical imperative entails.

Instead, and like Kant, we envision the usage of the term "principle" as analogous to the one he employed in the *Critique of Pure Reason*, where he refers to the "transcendental ideas" that pure reason faces in the "Transcendental Dialectic" (Kant, 1998, 394–408). Here he argues that these are not to be understood as strictly "binding principles"; rather, they are instead "orientative principles."

That expression refers to principles that must be considered asymptotic lines that can aggregate and gather patterns of behaviours that might *prima facie* be considered divergent. They are the peak of normative purity that trigger a constant interest for human reason and which individual agency is inescapably addressed. In Kantian terms, they are principles that cannot be the object of knowledge. Principles, then, are impossible for human beings to experience. We do not know such principles nor experience them in the total sense of the term. We can rationally imagine them as a point of orientation, a focus that orients our action and systematizes it along a coherent path.[2]

If we defer to the Kantian perspective, this comprehensive qualification of the "transcendental ideas" as "principles" is valid for both the cognitive and the practical realm. In the present context, we can imagine the same extended perspective. Nonetheless, what is more relevant for this level of discussion–and this is the second methodological remark–is that those principles should be considered orientative for and within any context of interrelation. The implications of this point are twofold.

1. On the one hand, we affirm that those principles should be considered orientative lines for any possible setting or state of affairs in which a form of exchange occurs that we can call "interaction." Phenomenologically speaking, we might agree on a minimal meaning of the term "interaction" by affirming the coexistence of two conditions: 1) this term describes a univocal spacetime frame in which at least two agents are present; 2) they are or become aware of the (effects of an) action that the first is doing as addressed to the second.
2. On the other hand, we would like to suggest that those principles should orient the pattern of action that guides any possible relationship among beings that share a "status of subjectivity." As a primary point of reference that can be commonly shared, with this expression, we mean a being that can start a "state of affairs" they are responsible for. We don't want to enter here further by embarking on a theoretical account of the subjectivity of robots [for a preliminary framing, see Stradella et al. (2012)]. Taking for granted that we are speaking about robots that are complex enough to consider themselves as starting points of possible (patterns of) actions, we are just alluding to the possible relationship network of humans-humans and humans-machines (by leaving open the possibility to imagine robot-robot relationships–which we do not investigate here).

Thirdly, by considering the principles we will articulate as orientative ones, we are paving the way for a different line of inquiry. We are thinking of principles that are intense, pervasive, and flexible enough to both inform and orient at the same time, as well as function in any context of the interrelation of each possible "subject" involved in it–again, by attributing such a status to both humans and robots. In more synthetic terms, we maintain that such a kind of principle can be adequate for orienting both the contextual settings or situation (embedding perspective) and the forms of subjectivity involved in them (embodying perspective). We will integrate this point in the following paragraph.

Last but not least, as our fourth methodological remark, the comprehensive approach here used can be described as hermeneutical. We can attribute this evocative word to the same twofold meaning that Hans-Georg Gadamer intended (Gadamer, 2013)–"fusions of horizons" on the one hand and as "history of effects" on the other. We would like to allude that any single situation or context of interrelation has its characteristics, spatial/temporal borders and constraints, and it embeds specific actors in it. There are no privileged or external points of view that can account for it with sufficient accuracy–and, consequently–that can take appropriate decisions outside the subjects involved in it. If this affirmation may be considered accurate for any interaction situation, it becomes still more evident by referring to forms of relationships mediated by technologies (Pirni and Carnevale, 2013; Pirni and Carnevale , 2014; Pirni et al., 2017). Again, this is another way of excluding from the very beginning the possibility for any top-down (now: external-internal) approach. Moreover, accuracy in making decisions is in direct proportion to the habits of interrelating with those specific agents or issues. In short, it is the result of intersubjective historicity that is lived in common.

AUTONOMY AND VULNERABILITY: A DUALITY OF PRINCIPLES FOR A RENEWED CARE-ETHICS APPROACH

By moving forward in designing the proposed approach, we must outline a comprehensive definition of each of the two principles we want to offer in this context. Again, to avert any reference to a top-down approach, we avoid the insertion of any unique or unifying principle. Instead, the insertion of the systemic linkage of two orientative principles (in the sense outlined above) we are going to illustrate has to be considered within a sharp argumentative lie that can be summarized in what follows:

1. Both principles are to be considered on the same plane or level. No one of them should be regarded as a priority, even in extreme conflictual situations;

[2]For a parallel approach that uses "insight" as the key-concept instead of "principle," in order to give space to a similar need for an orientative perspective in care ethics, see Leget et al. (2019).

2. They are constitutively complementary: neither an account of (a focus on) autonomy without considering vulnerability nor an asymmetric opposite account (focus) is admitted. They are two, but none of them can be neither subjected to the other nor avoided or underestimated in favour of the other.

Autonomy and Care: A Preliminary Outline

Accordingly, the comprehensive account of each of the two principles should be considered within a strong linkage with the other. This is precisely the meaning through which the same hermeneutic/phenomenological framework outlined above articulates. If we try and grasp the most phenomenologically evident meaning of both principles, we can affirm that they are opposed. What we are alluding to here is a theoretical understanding, which is related to an experience from the first-person point of view, more than linguistic antonyms (the subsequent paragraphs are devoted to expand and widely ground this point). Rather than being a weakness of our argument, their constituting opposite principles are, in fact, an intended characteristic that substantiates our reasoning. By adopting both principles, on the one hand, through *autonomy*, we set as an objective the guarantee of the maximum conceivable extent of independence to the individual. At the same time, by also contemplating the principle of *vulnerability*, we aim at the full possible degree of relationality and dependence of a subject from (an)other one (-s).

Accordingly, to try and outline a meaning of the principle of *autonomy* that is constitutively open towards and provide a potential integration to the principle of *vulnerability*, certainly the common understanding of "autonomy" must be finetuned in some of its primary and "classic" characteristics.

First of all, we must distinguish *autonomy* from *arbitrium*, distinguishing the former from arbitrary/unconstrained agency. Being autonomous does not mean "to do whatever one wants" nor whatever is conceivable and possible, according to one's own overall capability to act/to avoid acting in any context in which one's action might take place and be oriented to any (subject or thing) present in it. Such a definition would correspond to the very meaning of the concept of *arbitrium*. Here we indeed wish to differentiate from the first principle we were introducing-that is autonomy.

Instead, the concept of autonomy we are searching for is a principle that systematically opens up the possibility of acting in a context in which other subjects are acting or might act and, therefore, endowed with a *relational* dimension. In this understanding, the individual will is structurally open to any possible principle of concrete acting. Yet, it must select and choose among possible principles of acting while keeping in mind a general meta-rule that is shaped in line with the Kantian third formulation of the categorical imperative. This imperative would mean that anytime you are on the verge of acting, try to articulate, select, and put into practice just. Only that principle of action can you rationally imagine that any other subject may want and affirm on their behalf. In other words, try to act by *orienting* your agency while having in mind a systemic approval of your action and the subjective principle that guided it by any other subject who might act or might be the recipient of your activity in the same context. As we can see, such a formulation of the categorical imperative is inextricably tied to a conception of the individual, which is far from being solipsistic, like other different understandings of the concept of autonomy suggest.

Of course, in line with a relational conception of subjectivity, the target of our conceptual framework corresponds both to the individual as the subject who acts and to the individual as the subject who "receives" the action. In this perspective, then, being autonomous does not mean being a solipsistic agent pursuing their own goal whatever the conditions, whatever other subjects and correlative goals they may encounter. Rather, being autonomous means finding a systemic and dynamic balance between the need for self-sufficiency and the capability to start a state of affairs by one's own will on the one hand, and the need to take care of the analogous need and capability that guides and orientates the agency of any other subject in any specific context, on the other.

Ultimately, according to this definition, being autonomous means taking care of the autonomy of others as well as of the potential fragility and vulnerability that is endowed in each and every one of us (Pirni, 2006; Pirni, 2013; Pirni, 2016).

Vulnerability and Care: A Preliminary Outline

Given the understanding of autonomy we just outlined, its interplay with the principle of vulnerability is less problematic than one would expect. The endorsement of vulnerability can now be understood as coherent with but even necessary to the realization of the idea of autonomy we set up in the previous section. To define vulnerability, as a first approximation, we can build on a phenomenologically evident (instead of at-first-sight contradictory, as with our definition of autonomy) understanding of the concept. According to this understanding of vulnerability, the basic situation becomes one in which no individual can either live or survive, nor can they pursue their own goals alone. Relationality, in this view, is not just courtesy, nor a possible or socially acceptable behaviour, but rather an intrinsic characteristic of the subject. Such subjects do not stand alone but are permanently embedded in a relational net with others for survival and fulfilment. Relationality is a matter of systemic and vital necessity.

It merits noting here that we are not alluding to unique situations that bear clear evidence, like the condition of people with impairments. Instead, the perspective we adopt is in line with theories that see subjects as embedded in a net of relations and in contextual circumstances that try to manage - but, in the end, are forced to confirm-our constitutive dependence at the individual level as well as at the systemic one (Kittay, 1999). Such perspective is shared by normative frameworks such as relational egalitarianism (See Voigt, 2020). More importantly, the present research is first and foremost built upon the understanding of subjectivity expounded by feminist thinking (See, *inter alia*, Butler, 2004). Thus, we are not referring to a vulnerable subject in the sense that they are practically dependent on others to fulfill daily needs or perform basic activities. Instead, we refer to the basic constitutive situation of vulnerability shared

by every human being as an embedded feature of humanity, which is accurate and operating for each subject capable of acting. Such a subject is solely *prima facie* independent or autonomous. They are forced to act in a world shared with other subjects, within definite boundaries and facing a series of limitations in terms of lack of resources or deficiency of time. Following Arendt (1958), we can say that "Men [human beings] and not Man, live on the earth and inhabit the world" and that one defining characteristic of humanity is plurality. Further, they can be cognitively/ethically vulnerable, in the sense of not being equipped with sufficient knowledge as well as the ethical competencies to overcome specific difficulties, constraints, and limitations that interfere with both the most linear pursuit of their tasks and the due care to the autonomy/vulnerability of others.

In sum, the final achievement of this provisional theoretical path is a constitutive interdependence between the two principles. This might offer a challenging and potentially open theoretical "platform" to relaunch a care ethics perspective more in line with the demanding and urgent reshaping of any possible integration between human and machine.

CONCLUSION

Various approaches have been undertaken in an attempt to integrate ethical principles and practices in care ethics. This has similarly been an approach applied to the design and development of robotic technologies that fall within the domain of care (§ 1). This paper has taken these approaches as a starting point, illustrating how they have been employed and their shortcomings. In particular, we showed how both the traditional top-down and bottom-up approaches have fundamental misgivings (§ 2). This, consequently, is inextricably linked with foundational ethical issues. To address these issues, we propose a revision of the bottom-up approach as the most salient starting point for rethinking care ethics as it is applied to robots. The central innovative contribution of this paper is the proposal of rehabilitation of two orientative principles that can surround the entire theoretical building of any care ethics approach.

These principles were selected following a specific methodology (§ 3), which led to identifying an ethical horizon where the interplay between autonomy and vulnerability includes both humans and machines on a single plane. On the one hand, this horizon enhances the potential autonomy of both, but it also highlights their respective and constitutive vulnerability. On the other, this opens up the possibility of a new relational dimension (§ 4). In doing so, the central contribution of this approach aims to provide a framework that promises a more salient interplay, and possibly a novel integration, that is directed towards the future of our "living togetherness,"

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

REFERENCES

Allen, C., Smit, I., and Wallach, W. (2005). Artificial Morality: Top-Down, Bottom-Up, and Hybrid Approaches. *Ethics Inf. Technol.* 7, 149–155. doi:10.1007/s10676-006-0004-4

Anderson, S. L. (2008). Asimov's "three Laws of Robotics" and Machine Metaethics. *AI Soc.* 22 (4), 477–493. doi:10.1007/s00146-007-0094-5

Arendt, H. (1958). *The Human Condition*. Chicago: University of Chicago Press.

Arkin, R. (2009). *Governing Lethal Behavior in Autonomous Robots*. Boca Raton, FL: CRC Press. doi:10.1201/9781420085952

Baier, A. (1991). *A Progress of Sentiments. Reflections on Hume's Treatise*. Boston, MA: Harvard University Press.

Baier, A. (1995). *Moral Prejudices. Essays on Ethics*. Boston: Harvard University Press.

Balistreri, M. (2018). Sex Robot. *L'amore Al Tempo Delle Macchine*. Roma: Fandango.

Bendel, O. (2017). "Sex Robots from the Perspective of Machine Ethics," in *Love and Sex with Robots. LSR 2016. Lecture Notes in Computer Science*. Editors A. Cheok, K. Devlin, and D. Levy (Cham: Springer), Vol. 10237, 17–26. doi:10.1007/978-3-319-57738-8_2

Bisconti Lucidi, P., and Piermattei, S. (2020). "Sexual Robots: The Social-Relational Approach and the Concept of Subjective Reference," in *Human-Computer Interaction. Multimodal and Natural Interaction. HCII 2020. Lecture Notes in Computer Science*. Editor M. Kurosu (Cham: Springer), Vol. 12182, 549–559. doi:10.1007/978-3-030-49062-1_37

Botti, C. (2015). Feminine Virtues or Feminist Virtues? The Debate on Care Ethics Revisited. *Etica Politica/ Ethics Polit.* XVII (2), 107–151.

Botti, D. (2018). *Cura e differenza. Ripensare l'etica*. Milano: LED. doi:10.4000/cybergeo.29495

Butler, J. (2004). *Precarious Life: The powers of Violence and Mourning*. London & New York: Verso.

Cavallo, F., Semeraro, F., Fiorini, L., Magyar, G., Sinčák, P., and Dario, P. (2018). Emotion Modelling for Social Robotics Applications: A Review. *J. Bionic Eng.* 15, 185–203. doi:10.1007/s42235-018-0015-y

Cenci, A., and Cawthorne, D. (2020). Refining Value Sensitive Design: A (Capability-Based) Procedural Ethics Approach to Technological Design for Well-Being. *Sci. Eng. Ethics* 26, 2629–2662.

Coeckelbergh, M. (2020). "Moral Appearances: Emotions, Robots, and Human Morality," in *Machine Ethics and Robot Ethics*. Editors W. Wallach and P. Asaro (England, UK: Routledge), 117–123. doi:10.4324/9781003074991-13

Collins, S. (2015). *The Core of Care Ethics*. London: Palgrave Macmillan.

Driver, J. (2001). *Uneasy Virtue*. Cambridge: Cambridge University Press. doi:10.1017/cbo9780511498770

Dumouchel, P., Damiano, L., and DeBevoise, M. (2017). *Living with Robots*. Boston: Harvard University Press. doi:10.4159/9780674982840

Gadamer, H.-G. (2013). "Truth and Method (1960)," in *Transl. Rev.* J. Weinsheimer and D.G. Marshall, London: Bloomsbury.

Gilligan, C. (1982). *In a Different Voice: Psychological Theory and Women's Development*. Boston: Harvard University Press.

Goodall, N. J. (2014). Ethical Decision Making during Automated Vehicle Crashes. *Transportation Res. Rec.* 2424, 58–65. doi:10.3141/2424-07

Gunkel, D. J. (2012). *The Machine Question: Critical Perspectives on A.I., Robots, and Ethics*. Cambridge: MIT Press. doi:10.7551/mitpress/8975.001.0001

Held, V. (1993). *Feminist Morality: Transforming Culture, Society, and Politics*. Chicago: University of Chicago Press.

Held, V. (2006). *The Ethics of Care. Personal, Political, and Global*. Oxford: Oxford University Press.

Hume, D. (2007). "*A Treatise Of Human Nature (1739-1741)*," in *The Clarendon Edition of the Works of David Hume, 2, Voll*. Eds by D.F. Norton and M. J. Norton (Oxford: Oxford Clarendon Press).

Jacobs, N., and Huldtgren, A. (2018). Why Value Sensitive Design Needs Ethical Commitments. *Ethics Inf. Technol.*, 1–4.

Jecker, N. S. (2020). You've Got a Friend in Me: Sociable Robots for Older Adults in an Age of Global Pandemics. *Ethics Inf. Technol.*, 1–9. doi:10.1007/s10676-020-09546-y

Kant, I. (1998). "*Critique of Pure Reason (1781)*," in *Transl. Ed.* P. Guyer and A.W. Wood, Cambridge: Cambridge University Press.

Kittay, E. F. (1999). *Love's Labor: Essays on Women, equality and Dependency*. England, UK: Routledge.

Lecaldano, E. (2005). *Bioetica. Le Scelte Morali*. Roma-Bari: Laterza.

Leget, C., van Nistelrooij, I., and Visse, M. (2019). Beyond Demarcation: Care Ethics as an Interdisciplinary Field of Inquiry. *Nurs. Ethics* 26 (1), 17–25. doi:10.1177/0969733017707008

Noddings, N. (1984). *Caring. A Feminist Approach to Ethics and Moral Education*. Berkeley: University of California Press.

Noddings, N. (2002). *Starting at Home. Caring and Social Policy*. Berkeley: University of California Press.

Ostrowski, A. K., DiPaola, D., Partridge, E., Park, H. W., and Breazeal, C. (2019). Older Adults Living with Social Robots: Promoting Social Connectedness in Long-Term Communities. *IEEE Robot. Automat. Mag.* 26, 59–70. doi:10.1109/MRA.2019.2905234

Pirni, A., and Carnevale, A. (2014). Technologies Change-Do We Change as Well? on the Link between Technologies, Self, and Society. *Politica Società* 3 (2), 173–184.

Pirni, A., and Carnevale, A. (2013). "The challenge of Regulating Emerging Technologies: a Philosophical Framework," in *Law and Technology. The Challenge of Regulating Technological Development*. Editors E. Palmerini and E. Stradella (Pisa, Italy: Pisa University Press), 59–75.

Pirni, A., Esposito, R., Carnevale, A., and Cavallo, F. (2017). Sostenibilità etica dei personal care robot: linee per un inquadramento preliminare. *Nuova Corrente* LIX (n. 159), 133–151.

Pirni, A. (2013). "Freedom of the Will in Communitarian Perspective," in *Kant und die Philosophie in weltbürgerlicher Absicht* (De Gruyter), 509–520.

Pirni, A. (2006). *Kant Filosofo Della Comunità*. Pisa: Edizioni ETS.

Pirni, A. (2016). "The Place of Sociality: Models of Intersubjectivity According to Kant," in *Kant and Social Policies*. Editor A. Faggion, A. Pinzani, and N. Sanchez Madrid (Cham: Palgrave Macmillan), 65–92. doi:10.1007/978-3-319-42658-7_4

Sharkey, A. (2014). Robots and Human Dignity: a Consideration of the Effects of Robot Care on the Dignity of Older People. *Ethics Inf. Technol.* 16 (1), 63–75. doi:10.1007/s10676-014-9338-5

Slote, M. (2010). *Moral Sentimentalism*. Oxford: Oxford University Press.

Slote, M. (2011). *The Impossibility of Perfection: Aristotle, Feminism and the Complexities of Ethics*. Oxford: Oxford University Press.

Smith, A. (1976). "Theory of Moral Sentiments (1759)," A.A. Macfie and D. D. Raphael, Oxford: Oxford University Press.

Stradella, E., Salvini, P., Pirni, A., Di Carlo, A., Oddo, C. M., Dario, P., and Palmerini, E. (2012). "Subjectivity of Autonomous Agents: Some Philosophical and Legal Remarks," in ECAI Workshop on Rights and Duties of Autonomous Agents (RDA2), Montpellier, France, 24–31.

Tonkens, R. (2015). "Ethics of Robotic Assisted Dying," in *Machine Medical Ethics*. Editors S.P. Van Rysewyk and M. Pontier (Cham, Switzerland: Springer), 207–221. doi:10.1007/978-3-319-08108-3_13

Tronto, J. (2013). *Caring Democracy. Markets, Equality and Justice*. New York: New York University Press.

Tronto, J. (2015). *Who Cares? How to Reshape a Democratic Politics*. Ithaca: Cornell University Press.

Tronto, J. C. (1993). *Moral Boundaries: A Political Argument for an Ethics of Care*. London: Routledge.

Tronto, J. C. (2010). Creating Caring Institutions: Politics, Plurality, and Purpose. *Ethics Soc. Welfare* 4 (2), 158–171. doi:10.1080/17496535.2010.484259

Umbrello, S., Capasso, M., Balistreri, M., Pirni, A., and Merenda, F. (2021). Value Sensitive Design to Achieve the UN SDGs with AI: A Case of Elderly Care Robots. *Minds Mach.*, 1–25.

Van den Hoven, J. (2010). The Use of Normative Theories in Computer Ethics *The Cambridge Handbook of Information and Computer Ethics*, 59–76.

Van den Hoven, J., Miller, S., and Pogge, T. (2017). The Design Turn in Applied Ethics. *Design Ethics*, 11–31.

Van Rysewyk, S. P., and Pontier, M. (2015). "A Hybrid Bottom-Up and Top-Down Approach to Machine Medical Ethics: Theory and Data," in *Machine Medical Ethics*. Editors S.P. Van Rysewyk and M. Pontier (Cham, Switzerland: Springer), 93–110. doi:10.1007/978-3-319-08108-3_7

van Wynsberghe, A. (2013a). A Method for Integrating Ethics into the Design of Robots. *Ind. Robot* 40 (5), 433–440. doi:10.1108/IR-12-2012-451

van Wynsberghe, A. (2013b). Designing Robots for Care: Care Centered Value-Sensitive Design. *Sci. Eng. Ethics* 19 (2), 407–433. doi:10.1007/s11948-011-9343-6

van Wynsberghe, A. (2012). *Designing Robots with Care: Creating an Ethical Framework for the Future Design and Implementation of Care Robots*. Twente, Netherlands: University of Twente. doi:10.3990/1.9789036533911

van Wynsberghe, A. (2015). *Healthcare Robots. Ethics, Design and Implementation*. Franham: Ashgate.

van Wynsberghe, A. (2016). Service Robots, Care Ethics, and Design. *Ethics Inf. Technol.* 18 (4), 311–321. doi:10.1007/s10676-016-9409-x

van Wynsberghe, A. (2021). Social Robots and the Risks to Reciprocity. *AI Soc* [Epub ahead of print]. doi:10.1007/s00146-021-01207-y

Voigt, K. (2020). *Relational Egalitarianism*. Oxford, UK: Oxford Research Encyclopedia of Politics. doi:10.1093/acrefore/9780190228637.013.1387

Wallach, W., and Allen, C. (2015). *Moral Machines. Teaching Robots Right from Wrong*. Oxford: Oxford University Press.

Williams, B. (2011). *Ethics and the Limits Of Philosophy (1985)*. London: Routledge.

Winfield, A. F. T., Blum, C., and Liu, W. (2014). "Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection," in *Advances in Autonomous Robotics Systems*. Editors M. Mistry, A. Leonardis, M. Witkowski, and C. Melhuish (Berlin: Springer International Publishing), 85–96. doi:10.1007/978-3-319-10401-0_8

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Robotics and AI

PERSPECTIVE
published: 17 June 2021
doi: 10.3389/frobt.2021.634297

Onshoring Through Automation; Perpetuating Inequality?

*Matthew Studley**

Bristol Robotics Laboratory, Bristol, United Kingdom

Many analyses of the ethical, legal and societal impacts of robotics are focussed on Europe and the United States. In this article I discuss the impacts of robotics on developing nations in a connected world, and make the case that international equity demands that we extend the scope of our discussions around these impacts. Offshoring has been instrumental in the economic development of a series of nations. As technology advances and wage share increases, less labour is required to achieve the same task, and more job functions move to new areas with lower labour costs. This cascade results in a ladder of economic betterment that is footed in a succession of countries, and has improved standards of living and human flourishing. The recent international crisis precipitated by COVID-19 has underlined the vulnerability of many industries to disruptions in global supply chains. As a response to this, "onshoring" of functions which had been moved to other nations decreases risk, but would increase labour costs if it were not for automation. Robotics, by facilitating onshoring, risks pulling up the ladder, and suppressing the drivers for economic development. The roots of the economic disparities that motivate these international shifts lie in many cases in colonialism and its effects on colonised societies. As we discuss the colonial legacy, and being mindful of the justifications and rationale for distributive justice, we should consider how robotics impacts international development.

Keywords: ethics, onshoring, robotics, inequality, Development

OPEN ACCESS

Edited by:
Martin Magnusson, Örebro University, Sweden

Reviewed by:
Jessica Sorenson, University of Southern Denmark, Denmark
Angelika Zimmermann, Loughborough University, United Kingdom

***Correspondence:**
Matthew Studley
matthew2.studley@uwe.ac.uk

Specialty section:
This article was submitted to Ethics in Robotics and Artificial Intelligence, a section of the journal Frontiers in Robotics and AI

Received: *27 November 2020*
Accepted: *04 June 2021*
Published: *17 June 2021*

Citation:
Studley M (2021) Onshoring Through Automation; Perpetuating Inequality? Front. Robot. AI 8:634297. doi: 10.3389/frobt.2021.634297

INTRODUCTION

The year 2020 was noteworthy in many ways. Firstly, the SARS-CoV-2 virus COVID-19 pandemic had a massive impact on international trade and the global economy and highlighted the vulnerabilities of global supply chains (Free and Hecimovic, 2021), accelerating a drive to onshore, or "bring home", manufacturing that is in turn enabled through robotics. Secondly, even though mass assembly was contrary to public health recommendations, a wave of protests about the lethal consequences of police brutality and racially motivated violence spread from Minneapolis, MN, United States, to cities worldwide (Weine et al., 2020).

In Bristol, United Kingdom, these protests saw a city centre statue of the philanthropist Edward Colston toppled from its pedestal and rolled into the harbour (Nasar, 2020) to protest the trader's pivotal role in the Royal African Company, which "shipped more enslaved African women, men, and children to the Americas than any other single institution" (Pettigrew, 2013). The use of slave labour by the European Colonial powers, and the nations which inherited their colonial possessions, led via untold misery to the racist violence protested by the Black Lives Matter movement, and was also instrumental in many ways in creating the wealth disparities within and between nations which persist today. In large part it is this inequality which underlies the socio-economic imperative that

created global supply chains, and it is against this sweeping global historical backdrop that developments in robotics now find themselves playing a part.

In this article I first discuss offshoring, the process by which organisations in high-wage countries move part of their operation to nations where wages are lower. I briefly examine the impacts, both economic and otherwise, in the nations providing the offshored service, and then present the accelerating process of "onshoring", in which these functions are "brought home" in a process enabled by the replacement of human employees with automation. I argue that in so doing we risk stifling one of the ways in which inequalities are reduced through economic betterment. Thereafter I discuss the disparity in wealth between nations which motivates the ebb and flow of on- and off-shoring. These disparities are in many cases not due to accidents of geography but in a large part to historic depredation and abuse, enslavement and exploitation, and it is this history which is part of the narrative that led to the Black Lives Matter protests. Finally, I argue that we should recognise this wider, international dimension in our discussions of the ethical, societal, and legal impacts of robots, and that it is incumbent upon us as moral agents to act to redress these impacts beyond the national scope within which they are normally considered.

In the following sections I will often focus on the United States and China as exemplars due to their economic predominance, though of course there are many nations involved in these global flows of goods and money.

OFFSHORING

The process of global trade has existed long before the current era; some authors argue that globalization began in the 16th Century (Flynn and Giráldez, 2004), and accelerated massively in the 19th (O'Rourke and Williamson, 2004). However, especially since 1980 there has been a tendency not just to trade, but to structure the world's manufacturing production around global supply chains, in which raw materials and intermediate goods shuttle back and forth across the planet before they are exported from this process to consumers (Free and Hecimovic, 2021). This trend has been accelerated by the neoliberal consensus which conceives of markets, rather than states, as the primary driving force in social organisation (Mudge, 2008). Neoliberal globalisation predicates increased international flows of trade, labour, capital and technology.

As the notion of globalised supply chains developed, there has been an increasing trend for many companies in high-income economies to outsource manual tasks to countries with lower labour costs in a process known as offshoring. This trend of "global labour arbitrage" (Roach, 2004) has been enabled by international IT infrastructure that enables rapid communication and the direct comparison of prices worldwide, and has been driven by a desire to seek efficiencies through cutting costs as companies lose pricing leverage in an era of excess supply. In general, the lower the per-capita income of a United States trading partner, the higher its share of United States "arm's length trade" in which production is entirely subcontracted (Lakatos and Ohnsorge, 2017), indicating that it is the lure of lower labour costs which drives the offshoring trend.

The size of this shift has been enormous. Global annual trade in physical merchandise has grown to $20T USD, with emerging economies accounting for almost half that figure, totalling $8.2T USD in exports (World Trade Statistical Review 2019). However, there are a number of reasons why the desirability of this trend is being re-assessed. The struggle for global hegemony between the US and China has led to a trade war (Kim, 2019) in which the US seeks to enlist its allies. Meanwhile, increasing public outcry about carbon emissions has called into question companies' reliance on long distance supply chains, though assigning emissions within these chains from production to final consumption is not easy (Kagawa et al., 2015).

THE ROLE OF OFFSHORING ON DEVELOPMENT

The impact of offshoring on both the developed and developing partners is nuanced. A naïve expectation might be that there is a disadvantage to the country that exports jobs; in fact some studies show an increase in the demand for skilled labour at both ends of the supply chain (Feenstra et al., 1996), with a concomitant rise in wage inequality within each nation as lower skilled jobs are lost. While the overall effect of the relationship is economically positive, there are a number of reasons why the effect on the developing nation may be ambiguous, but it has been shown that these can be addressed through labour market policies (such as a minimum wage) in the developing country (Bandyopadhyay et al., 2020).

One of the impacts of robotics is claimed to be an increase in high-skill and some middle-skill occupations (Dahlin, 2019), and a similar change in employment patterns is one of the advantages which has traditionally come from offshoring; low skill jobs move overseas, where they cause an upskilling in the local working population (Feenstra et al., 1996).

As skills and infrastructure improve in the developing nation, less labour is required to achieve the same aim. National differences in wage bargaining power, along with average wage levels, can influence the decisions multinationals make about where to locate production (Sly and Soderbery, 2014). Jobs move from nation to nation; nations which were once the source in global supply chains start in turn to source materials and production from other nations (Kizu et al., 2019). Most legislation and institutions which protect workers' rights and wages operate within countries, not between them, and offshoring represents a mechanism through which these institutions can be evaded, which can lead to negative effects on pay and working conditions (Drahokoupil and Fabo, 2017).

Although changing exchange rates and fallible data make comparisons hard, it has been shown that relative unit labor costs (constructed from available compensation, employment, and value added data) have risen in China between 1998 and 2012, although remaining far lower than in the United States (Ceglowski and Golub, 2012). However, much of this rise has

been due to the appreciating value of the yuan; indeed a meta-analysis of the impact of offshoring on wages shows the average effect to be negligible in both the originating and destination countries (Cardoso et al., 2020). Within this period, there has been a fundamental shift in living standards within China; in 2000, only 4% of urban households in China was middle class, increasing to 68% in 2012, and this has been forecast to reach 76% in 2022 (Barton, 2013). This distribution of wealth is expected to have massive macroeconomic impacts, as better healthcare, education and a rising service sector are expected to provide the basis for innovation and technological advancement, enabling Chinese industry to upgrade and climb the value chain. The process of offshoring jobs from the United States (which remains the dominant export destination for Chinese goods (Kizu et al., 2019)) has enabled a revolutionary change in China's economy, for the betterment of its people.

It should be noted that the impact of offshoring is not just economic. In their paper (Ravishankar et al., 2010) on the impacts of offshoring on workers in India, Ravishankar et al. describe feelings of insecurity that arise from recognising that sentiment in the client's nation might turn against the arrangement, or that client organisations might move on to another low-wage economy. Again, workers may feel their ambitions are constrained to routine work with limited opportunity for progression. Throughout, one hears the echoing psychological impact of being in a subordinate role, especially within the postcolonial context and its attendant baggage. A complex interplay of accommodation and resistance is an attendant part of offshoring's role in economic betterment.

ONSHORING

In 2020, the advent of COVID-19 highlighted the fragility of global supply chains (McKibbin and Fernando, 2020). An abrupt drop of 13.5% in China's industrial production in the first 2 months of 2020 reverberated around the world, causing shortages in many manufactured goods (Free and Hecimovic, 2021). This was further amplified by nationalist protectionism, for example, with China banning the export of masks and other medical supplies (Busch, 2020) and the EU restricting the export of PPE (Müller and Terem, 2021), in a wide-spread drive towards self-sufficiency in medical supplies. The COVID-19 crisis has added an urgency to calls for a greater degree of onshoring, and questioned the sustainability of existing patterns of supply (Free and Hecimovic, 2021), and many companies are looking to adopt new supply chains to reduce exposure to global disruptions in trade flows (Helmold et al., 2020; Javorcik, 2020; Shih, 2020).

Companies must place supply chain integrity above the cost savings associated with offshoring; lower labour costs in trading partners can increase profit margins, but must be weighed against the catastrophic spectre of having no goods to sell due to fragile supply chains. One way of increasing supply chain integrity is by reducing exposure to global shocks, restrictive trade practices and transport challenges by "onshoring," or "reshoring"; bringing work activities home and shortening supply chains. This long predates the COVID-19 pandemic, of course. There are numerous drivers for onshoring, with cost motivations only one part of the picture (Barbieri et al., 2018), and reshoring has been gathering pace since before 2012, when it was reported that 14% of United States firms surveyed "definitely planned to reshore" (Gray et al., 2013). Robotics and automation are one of the enablers of this trend (Slaby, 2012; Salazar and Lunsford, 2014; Sayer, 2016; Robey and Bolter, 2020) which claims as one of its potential benefits an increase in sustainability through, e.g. reducing the carbon emissions from travel of raw materials, part work and finished goods (Ashby, 2016).

However, this "onshoring" is unlikely to result in a one-for-one move of jobs into developed nations; for example, the use of technologies such as automation, robotics and additive manufacture allowed Adidas to employ only 160 high-skilled workers in a plant in Germany to replace the 1,000 workers in one of its comparable plants in East Asia (Economist, 2017). The role of robotics and automation in enabling onshoring is clear; there is a reduction in jobs (especially low-skilled jobs) overall, with the additional prospect of raising wage inequality in the home economy (Krenz et al., 2018).

ROBOTICS AS AN INHIBITOR OF INTERNATIONAL EQUALITY

We have seen then that the comparative lack of international regulation might predict that offshoring may drive down wages and degrade workers' rights (Drahokoupil and Fabo, 2017), but there can be a notable positive impact on the lives of working people in developing nations through the redistribution of wealth, upskilling, and education. This ladder of economic betterment finds its feet in different nations, and the cascade repeats.

What then might be the impact of robotics on this international engine of development and equality? It might be argued that through its role in allowing onshoring, automation will reduce this upward pressure, having a negative impact on developing nations. Furthermore, this impact of robotics and automation in developing economies could be exacerbated as automation in the developing nation may rapidly drive down wages by reducing effective labour requirements per unit task (Bandyopadhyay et al., 2020). In China, it has been forecast that automation could remove 20% of manufacturing jobs (12% of the country's total) by 2030, replacing one-fifth of the country's jobs in the manufacturing industry, with the possibility that up to 100 million workers will need to change their field of work (Yiran, 2018).

Robotics and automation, by facilitating the onshoring of supply chains, risk inhibiting the economic betterment of people in developing countries.

INTERNATIONAL WEALTH DISPARITIES AND HISTORIC WRONGS

Offshoring then is a process whereby resources flow between nations, with their role in each offshoring relationship determined by their relative wealth. But why are some nations

rich and others poor? Throughout this discussion I have casually referred to developing economies and nations, which are most often the destinations of offshoring decisions. "Developing" might not be the best term for these nations, as it implies a hierarchy with "Western" economies (to use another loaded term) portrayed as ideal destinations, being as they are "Developed". As an alternative one might invert the implied value relationship between nations and refer instead to "countries that were colonised" (Silver, 2015), setting the distinction between nations in the context of their mutual histories.

However, not all nations which were colonised in the last 500 years developed in the same way; for example, the United States developed along a different economic trajectory to the nations of Latin America and Sub-Saharan Africa although all (except Liberia) were colonised by European powers. North America lacked a large and dense indigenous population which could be exploited, while South America and Africa did not. Irrespective of their nation of origin, where colonisers found densely-settled lands they set up "extractive institutions", to profit from the land and labour of indigenous peoples (Acemoglu and Robinson, 2017). Similarly, the disease environment in different areas affected the likelihood of settlement by Europeans, which in turn affected the institutions which were established there. Acemoglu and Robinson argue that up to 30% of per capita income inequality can be explained by the varying impact of European colonialism on different societies.

In the Americas and Carribean where no suitable population existed (or where a population such as the indigenous peoples of Spanish South America, having been hitherto exploited was no longer available, in large part due to the depredations of the colonists (Meade, 2016)), the colonial powers imported slaves in the second great wave of African slavery. Apart from the immense human suffering visited upon the enslaved and their descendants, this process had a profound impact on the economies of the countries involved. The rich got richer. The nations from which slaves were taken suffered economically, culturally, and politically, and have continued to suffer thus ever since; the nations which received the slaves saw other, profound impacts, especially in the continued disparity in wealth and power between the elites and the slaves. See Bertocchi for a review of the evidence for these causal relationships and their many and varied impacts (Bertocchi, 2016).

Poor countries are poor for many reasons. A history of colonial exploitation plays a large part in this, as may the impact of slavery. Might robotics, in enabling onshoring, help perpetuate this injustice?

DISCUSSION

We have seen the link between robotics and development. Robotics enables the process of onshoring that has been stimulated and given urgency by the COVID-19 crisis, and it is possible that the process of economic betterment for at least some of the majority of the world's population may be slowed or choked by this sea change.

The benefits that people enjoy from their human and natural environment vary in their distribution, both within and between nations. Frameworks and arguments about the ways in which these distributions should function are the subject of International Distributive Justice. Few would argue that gender, disability or ethnicity should be a justification for discrimination in access to these benefits; why then should ones nation of birth be a basis for inequality (Pogge, 1989)? Distributive justice is not only ethically salient, but also important for the maintenance of our shared natural environment where in richer economies, higher income inequality increases per capita emissions (Grunewald et al., 2017).

Apart from the monetary flows from trade, this rebalancing most obviously takes the form of aid and charity. However, accepting that colonialism and historic injustices underpin the wealth differential between the "developed" and the "developing", the language of aid and charity seems disingenuous at best. Poverty is not a given, a fact of the world which has somehow sprung into being despite our best efforts; it is created and is a violation of the natural rights of the poor. It was an understanding of "Natural Law", of the rights of men and women, by nature free and equal, that led Locke to propound his early arguments for Reparation, i.e. the satisfaction due to a victim from the perpetrator of suffered wrongs. Framing international distributive justice within the wider context of reparations (for the descendents of slaves, or for nations that were stripped by slavery or colonialism) is a partial answer, but only partial at best, for not all the world's poor are descended from slaves, not all the countries which profit so greatly today were equally implicated in this ill, and the language of blame and reparation creates hostility. Iris Young argued that we all bear an obligation to address structural injustice, by virtue of being members of society (Young and Nussbaum, 2011). As we recognise our global interconnectedness, through pandemics, failing supply chains and the seismic economic shocks that reverberate through economies, it has never been more clear that society itself is global and our responsibilities are global too.

In this short paper we have seen the following argument;

1. International inequalities underlie the economic rationale for global supply chains (GSC). These inequalities are to some extent addressed through the redistribution of wealth through globalisation.
2. In some large part, these inequalities may be due to a colonial history that enriched many nations today recognised as "developed", at the expense of the "developing".
3. These inequalities are unjust, stemming from historic ills visited on the weak by the strong
4. As moral agents, we bear a collective responsibility to address injustice.
5. Robotics and Automation enable a reduction in GSC through onshoring, and this trend has been accelerated due to the difficulties in international trade experienced due to the COVID-19 pandemic.
6. Since (5) reduces the redistributive effect of GSC, robotics and automation may act contrary to (4).

What then is the impact of this discussion on robotics, and how should we respond?

1. By considering Ethical, Legal and Societal aspects of robotics and automation beyond national boundaries, and to recognise the international impact of our actions. Frameworks such as MEESTAR (Wutzkowsky and Böckmann, 2018) already recognise the societal context in their assessment of the impacts of new technologies; society does not end at a nation's border. I suggest that we need to make the transnational implications of our decision making about robotics and automation an explicit and expected subject of our ethical assessments.
2. By recognising that our actions within the discipline of robotics may have an impact within the current historic context, beyond that which we might normally consider. The tools of AI and robotics stand ready to fundamentally change the world. The great social trends and challenges of our times, the empowerment of the disenfranchised and economically repressed, and the righting of historic wrongs, can be helped or hindered through the ways in which we choose to support and abet the application of these tools.
3. By promoting the concept of a global mechanism which puts the redistribution of wealth generated through robotics in the context not just of national, but of global welfare. This will be difficult; agreement on how to tax robotics is elusive (Kovacev, 2020), and it is hard to regulate international taxation regimes to ensure that companies pay their dues (Ozai, 2018-2019) though there is public and governmental appetite, and a moral case, to address unfair and unethical practices (West, 2018). Despite the latter, the consideration of the international redistribution of wealth as a means of addressing inequality "is almost absent from the international agenda" (Melamed and Smithyes, 2009).

Whether motivated by the exigencies of climate change, addressing entrenched inequity, or claiming back value from multinationals for the benefit of the peoples who make and consume their services and goods, some problems can only be addressed through collective action.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Acemoglu, D., and Robinson, J. A. (2017). The Economic Impact of Colonialism. *Long Econo. Polit. Shadow Histo. Volume I. A Glob. View*, 81.

Ashby, A. (2016). From Global to Local: Reshoring for Sustainability. *Operations Manag. Res.* 9, 75–88. doi:10.1007/s12063-016-0117-9

Bandyopadhyay, S., Mitra, D., Basu, A., and Chau, N. (2020). Consequences of Offshoring to Developing Nations: Labor-Market Outcomes, Welfare, and Corrective Interventions. *Econo. Inquiry* 58, 209–224. doi:10.20955/wp.2016.011

Barbieri, P., Ciabuschi, F., Fratocchi, L., and Vignoli, M. (2018). What Do We Know about Manufacturing Reshoring? *Journal of Global Operations and Strategic Sourcing*. 11, 79. doi:10.1108/jgoss-02-2017-0004

Barton, D. (2013). *The Rise of the Middle Class in China and its Impact on the Chinese and World Economies*. US-China Economic Relations in the Next 10 years: Towards Deeper Engagement and Mutual Benefit, 138–148.

Bertocchi, G. (2016). The Legacies of Slavery in and Out of Africa. *IZA J. Migration*. 5, 24. doi:10.1186/s40176-016-0072-0

Busch, M. L. (2020). *Trade Protectionism Won't Help Fight COVID-19 - Global Trade Magazine*. Available at: https://www.globaltrademag.com/trade-protectionism-wont-help-fight-covid-19/ (Accessed October 29, 2020).

Cardoso, M., Neves, P. C., Afonso, O., and Sochirca, E. (2020). The Effects of Offshoring on Wages: a Meta-Analysis. *Rev. World Econ.* 157, 149–179. doi:10.1007/s10290-020-00385-z

Ceglowski, J., and Golub, S. S. (2012). Does China Still Have a Labor Cost Advantage? *Glob. Econ. J.* 12, 1850270. doi:10.1515/1524-5861.1874

Dahlin, E. (2019). Are Robots Stealing Our Jobs? *Socius*. 5, 2378023119846249. doi:10.1177/2378023119846249

Drahokoupil, J., and Fabo, B. (2017). *"Outsourcing, Offshoring and the Deconstruction of Employment: New and Old Challenges," in The Deconstruction of Employment as a Political Question*.(Palgrave), London, UK.

Economist (2017). *Adidas's High-Tech Factory Brings Production Back to Germany*. Economist, London, UK.

Feenstra, R., Grossman, G., and Irwin, D. (1996). *The Political Economy of Trade Policy: Papers in Honor of Jagdish Bhagwati*. The MIT Press. Cambridge, MA, USA.

Flynn, D. O., and Giráldez, A. (2004). Path Dependence, Time Lags and the Birth of Globalisation: A Critique of O'Rourke and Williamson. *Eur. Rev. Econ. Hist.* 8, 81–108. doi:10.1017/s1361491604001066

Free, C., and Hecimovic, A. (2021). Global Supply Chains after COVID-19: the End of the Road for Neoliberal Globalisation? Accounting. *Auditing Account. J. ahead-of-print*. 34. doi:10.1108/AAAJ-06-2020-4634

Gray, J. V., Skowronski, K., Esenduran, G., and Johnny Rungtusanatham, M. (2013). The Reshoring Phenomenon: What Supply Chain Academics Ought to Know and Should Do. *Supply Chain Manage.: Int. J.* 49, 27–33. doi:10.1111/jscm.12012

Grunewald, N., Klasen, S., Martínez-Zarzoso, I., and Muris, C. (2017). The Trade-Off between Income Inequality and Carbon Dioxide Emissions. *Ecol. Econ.* 142, 249–256. doi:10.1016/j.ecolecon.2017.06.034

Helmold, M., Einmahl, M., Rassmann, K. J., and Carvalho, L. N. (2020). *Lessons from the COVID-19 Situation: Rethinking Global Supply Chain Networks and Strengthening Supply Management in Public Procurement in Germany*. Bad Honnef, Germany: IUBH University of Applied Sciences. Available at: https://ideas.repec.org/p/zbw/iubhtl/42020.html (Accessed March 31, 2021).

Javorcik, B. (2020). *Global Supply Chains Will Not Be the Same in the post-COVID-19 World. COVID-19 and Trade Policy: Why Turning Inward Won't Work 111*. Available at: https://www.svensktnaringsliv.se/bilder_och_dokument/iz8xue_covid-19-and-trade-policy-28-aprilpdf_1005375.html/Covid-19+and+trade+policy+28+april.pdf#page=122.

Kagawa, S., Suh, S., Hubacek, K., Wiedmann, T., Nansai, K., and Minx, J. (2015). CO2 Emission Clusters within Global Supply Chain Networks: Implications for Climate Change Mitigation. *Glob. Environ. Change*. 35, 486–496. doi:10.1016/j.gloenvcha.2015.04.003

Kim, M.-H. (2019). *International Trade, Politics and Development*. Available at: https://www.emerald.com/insight/content/doi/10.1108/ITPD-02-2019-003/full/html.

Kizu, T., Kühn, S., and Viegelahn, C. (2019). Linking Jobs in Global Supply Chains to Demand. *Int. Labour Rev.* 158, 213–244. doi:10.1111/ilr.12142

Kovacev, R. (2020). *A Taxing Dilemma: Robot Taxes and the Challenges of Effective Taxation of AI, Automation and Robotics in the Fourth Industrial Revolution*. Available at: https://papers.ssrn.com/abstract=3570244 (Accessed July 1, 2020).

Krenz, A., Prettner, K., and Strulik, H. (2018). *Robots, Reshoring, and the Lot of Low-Skilled Workers*. Center for European Governance. Göttingen, Germany, doi:10.2139/ssrn.3208886

Lakatos, C., and Ohnsorge, F. (2017). *Arm's-length Trade: A Source of post-crisis Trade Weakness*. The World Bank. Washington, DC, USA.

McKibbin, W., and Fernando, R. (2020). *The Economic Impact of COVID-19. Economics In the Time Of COVID-19 45*. Available at: https://www.incae.edu/sites/default/files/covid-19.pdf#page=52.

Meade, T. (2016). *History of Modern Latin America: 1800 to the Present*. John Wiley & Sons, Hoboken, NJ, USA.

Melamed, C., and Smithyes, C. (2009). *Global Redistribution as a Solution to Poverty*. Brighton: Institute of Development Studies at the University of Sussex. Brighton, UK.

Mudge, S. L. (2008). What Is Neo-Liberalism? *Socioecon Rev.* 6, 703–731. doi:10.1093/ser/mwn016

Müller, V., and Terem, P. (2021). *Globalization of EU Trade Policy in the COVID-19 Era*. Les Ulis: SHS Web Of Conferences; Les Ulis (Les Ulis, France, Les Ulis: EDP Sciences). doi:10.1051/shsconf/20219201034

Nasar, S. (2020). Remembering Edward Colston: Histories of Slavery, Memory, and Black Globality. *Womens Hist. Rev.* 29, 1218–1225. doi:10.1080/09612025.2020.1812815

O'Rourke, K. H., and Williamson, J. G. (2004). Once More: When Did Globalisation Begin? *Eur. Rev. Econ. Hist.* 8, 109–117. doi:10.1017/s1361491604001078

Ozai, I. O. (2018-2019). Tax Competition and the Ethics of Burden Sharing. *Fordham Int'l L.J.* 42, 61–100.

Pettigrew, W. A. (2013). *Freedom's Debt: The Royal African Company and the Politics of the Atlantic Slave Trade*. University of North Carolina Press, Chapel Hill, NC, USA, 1672–1752.

Pogge, T. (1989). *Realizing Rawls*. Cornell University Press, Ithaca, NY, USA.

Ravishankar, M. N., Cohen, L., and El-Sawad, A. (2010). Examining Resistance, Accommodation and the Pursuit of Aspiration in the Indian IT-BPO Space: Reflections on Two Case Studies. *Ind. Relat. J.* 41, 154–167. doi:10.1111/j.1468-2338.2009.00560.x

Roach, S. (2004). *How Global Labour Arbitrage Will Shape the World economyGlobal Agenda*. Available at: http://ecocritique.free.fr/roachglo.pdf.

Robey, K., and Bolter, J. (2020). *Strategic Reshoring: A Literature Review*. Available at: https://research.upjohn.org/cgi/viewcontent.cgi?article=1257&context=reports.

Salazar, F. J., and Lunsford, R. (2014). *Onshoring: An I-Opener for Apple, Inc, International Journal Of Trends In Economics Management & Technology (IJTEMT) 3*. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.684.7646&rep=rep1&type=pdf.

Sayer, L. (2016). *Inequality in an Increasingly Automated World*. World Social Science Report, Paris, 41.

Shih, W. (2020). Is it Time to Rethink Globalized Supply Chains? *Cambridge*. 61, 1–3.

Silver, M. (2015). *If You Shouldn't Call it the Third World, what Should You Call it?* NPR. Available at: https://www.npr.org/sections/goatsandsoda/2015/01/04/372684438/if-you-shouldnt-call-it-the-third-world-what-should-you-call-it (Accessed November 18, 2020).

Slaby, J. R. (2012). Robotic Automation Emerges as a Threat to Traditional Low-Cost Outsourcing. *HfS Research Ltd.* 1, 3.

Sly, N., and Soderbery, A. (2014). Strategic Sourcing and Wage Bargaining. *J. Dev. Econ.* 109, 172–187. doi:10.1016/j.jdeveco.2014.04.005

Weine, S., Kohrt, B. A., Collins, P. Y., Cooper, J., Lewis-Fernandez, R., Okpaku, S., et al. (2020). Justice for George Floyd and a Reckoning for Global Mental Health. *Glob. Ment. Health (Camb).* 7, e22. doi:10.1017/gmh.2020.17

West, A. (2018). Multinational Tax Avoidance: Virtue Ethics and the Role of Accountants. *J. Bus. Ethics*. 153, 1143–1156. doi:10.1007/s10551-016-3428-8

World Trade Statistical Review (2019). Available at: https://www.wto.org/english/res_e/statis_e/wts2019_e/wts19_toc_e.htm (Accessed October 29, 2020).

Wutzkowsky, J., and Böckmann, B. (2018). Using MEESTAR to Identify Ethical and Social Issues Implementing a Digital Patient-Centered Care Platform. *Stud. Health Technol. Inform.* 248, 278–285. doi:10.3233/978-1-61499-858-7-278

Yiran, Z. (2018). *AI Automation to Shake up Labor Market by 2030*. Available at: http://www.chinadaily.com.cn/a/201808/23/WS5b7e2a62a310add14f387582.html (Accessed November 18, 2020).

Young, I. M., and Nussbaum, M. (2011). *Responsibility for Justice (Oxford Political Philosophy)*. Cary NC, USA: OUP.

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Robotics and AI

ORIGINAL RESEARCH
published: 29 June 2021
doi: 10.3389/frobt.2021.644336

Role-Play as Responsible Robotics: The Virtual Witness Testimony Role-Play Interview for Investigating Hazardous Human-Robot Interactions

Helena Webb[1], *Morgan Dumitru*[1], *Anouk van Maris*[2], *Katie Winkle*[3], *Marina Jirotka*[1]* *and Alan Winfield*[2]

[1]Department of Computer Science, University of Oxford, Oxford, United Kingdom, [2]Bristol Robotics Lab, University of West of England, Bristol, United Kingdom, [3]Royal Institute of Technology, Stockholm, Sweden

The development of responsible robotics requires paying attention to responsibility within the research process in addition to responsibility as the outcome of research. This paper describes the preparation and application of a novel method to explore hazardous human-robot interactions. The Virtual Witness Testimony role-play interview is an approach that enables participants to engage with scenarios in which a human being comes to physical harm whilst a robot is present and may have had a malfunction. Participants decide what actions they would take in the scenario and are encouraged to provide their observations and speculations on what happened. Data collection takes place online, a format that provides convenience as well as a safe space for participants to role play a hazardous encounter with minimal risk of suffering discomfort or distress. We provide a detailed account of how our initial set of Virtual Witness Testimony role-play interviews were conducted and describe the ways in which it proved to be an efficient approach that generated useful findings, and upheld our project commitments to Responsible Research and Innovation. We argue that the Virtual Witness Testimony role-play interview is a flexible and fruitful method that can be adapted to benefit research in human robot interaction and advance responsibility in robotics.

Keywords: role-play, interview, responsible research and innovation, methods, human-robot interaction

OPEN ACCESS

Edited by:
Martim Brandão,
King's College London,
United Kingdom

Reviewed by:
Dieter Vanderelst,
University of Cincinnati, United States
Anders Lennart Green,
Södertörn University, Sweden
Jason Millar,
University of Ottawa, Canada

***Correspondence:**
Marina Jirotka
marina.jirotka@cs.ox.ac.uk

Specialty section:
This article was submitted to Ethics in Robotics and Artificial Intelligence, a section of the journal Frontiers in Robotics and AI

Received: *20 December 2020*
Accepted: *01 June 2021*
Published: *29 June 2021*

Citation:
Webb H, Dumitru M, van Maris A, Winkle K, Jirotka M and Winfield A (2021) Role-Play as Responsible Robotics: The Virtual Witness Testimony Role-Play Interview for Investigating Hazardous Human-Robot Interactions. Front. Robot. AI 8:644336. doi: 10.3389/frobt.2021.644336

INTRODUCTION

The COVID-19 pandemic has created massive disruption to research projects relying on data collection involving human participants. At the same time, the need for adaptation has fostered opportunities for creativity in the adoption and application of methods. This paper describes how our research team developed and tested a new approach that enabled us to explore human-robot interaction whilst working from home. We developed a research protocol to investigate accidents involving social robots and humans; this protocol uses an online format and invites human participants to role-play a scenario in which they become observers to the aftermath of an accident and provide witness testimony in relation to it. This research formed part of our ongoing project work on responsible robotics and, despite the constraints, the pandemic proved to be a catalyst for innovation in our approach and enabled us to work through certain logistical and ethical challenges we were facing in our study. The protocol we developed provided a means for us to incorporate principles of Responsible Research and Innovation (RRI) into our work by establishing a

safe and ethical process through which participants can experience hazardous human-robot interactions. In this paper, we briefly outline the overall focus of our project and then catalogue the decision-making that led to the development of our new research protocol. We situate this approach within a discussion of role-play in studies of human-robot interaction (HRI) and human-computer interaction (HCI) more generally. We provide a detailed account of how our initial set of Virtual Witness Testimony (VWT) role-play interviews were conducted and our findings section focuses on what we discovered about the efficiency of the approach, the quality of the results it generated, and its limitations. In the Discussion section we comment on how the VWT role-play interview forms a flexible and fruitful method that can be adapted to benefit others working in HRI and seeking to advance responsibility in robotics.

BACKGROUND

RoboTIPS: Developing Responsible Robotics for the Digital Economy

The ongoing RoboTIPS study (Webb et al., 2019) is a five-year fellowship project that explores opportunities for the development of responsible robotics within the context of the contemporary digital economy. The project is underpinned by a commitment to Responsible Research and Innovation (RRI), an initiative that seeks to ensure that processes of research and innovation benefit society and the environment (Stilgoe et al., 2013; Rome Declaration 2014). In the context of academic research, adopting an RRI approach involves acknowledging that the responsibilities held by researchers, universities and funders broaden out from traditional issues of research integrity and plagiarism etc., (vom Schomberg and Hankins 2019). This broadening brings in further aspects around research processes such as gender equality and stakeholder inclusion, and also requires attending to the social, policy and environmental impacts of work. Within this perspective, responsibilities for the practices and outcomes of research are shared out across the research ecosystem, research communities take on new co-responsibilities, and society (*via* stakeholders) becomes involved in research and innovation across all of its phases Owen et al., 2013.

A key strand of RoboTIPS examines the investigation of incidents and accidents involving social robots in which humans are harmed in some way (Winfield et al., 2021). We focus on social robots that interact with humans as part of their day-to-day function, in particular assistive robots, automated vehicles and robot toys. We take the position that as more and more robots become commercially available, incidents and accidents, whilst hopefully rare, can be expected to occur. Therefore, it is necessary to develop mechanisms to identify the causes of these incidents and take steps to prevent them re-occurring. Our project work includes the design, development and testing of an innovative safety feature for social robots. The Ethical Black Box (EBB) (Winfield and Jirotka, 2017) is a data recorder for social robots, equivalent to the flight data recorders used in aviation. It continuously records sensor and relevant internal status data and can be extended in scope to also capture the AI decision-making processes of the robot and environmental phenomena around it. Just as black boxes are used on in aviation to provide crucial evidence following an accident, so the EBB can be used as a data source following some kind of incident or accident involving a social robot. The information provided by the EBB can help to identify failures in the robot and to understand why it behaved in the way it did. This data is used as part of a wider investigation process. Human witnesses to the incident report their recollections and understandings of the event, and the EBB data provides another form of witness testimony. In addition, various experts provide details about the specific setting and the robots involved. As a result, this investigation process aims to determine the cause of the incident and then produce recommendations–which might take the form of technical changes to the robot and its setup, as well as organisational changes in the setting–to prevent similar events from occurring in future and therefore avoid further harms. In this way, the EBB-informed investigation process serves as an innovation for safety, trust, accountability, and transparency in social robotics.

Incidents and Accidents Involving Social Robots: Investigating Hazardous Human-Robot Interactions

Our RoboTIPS project work to develop and trial the EBB requires us to understand how accidents involving social robots and humans unfold, and how humans at the scene respond to them. This includes understanding how humans might interact with the robot in the aftermath of an incident and how (as well as how much) they recall what they saw afterwards. Deriving this understanding will help us to optimise the accident investigation process, for instance in determining what kind of interactions humans might have with an EBB-enhanced robot in the context of an accident and how EBB data can best supplement testimony provided by human witnesses. We ultimately plan to run a series of laboratory-based simulated scenarios in which we stage an accident involving real robots and human participants and then run an investigation process with expert participants who will work through the human and robot witness data to discern the causes of the accident. This quasi-naturalistic approach will collect highly valuable data but also represents logistical and ethical challenges. It requires a great deal of advance planning and piloting to ensure it is fully workable: care needs to be taken to ensure that the accident scenario is viable and realistic but does not cause any actual harm. Careful organization is needed so that the processes of accident and subsequent investigation run smoothly in the time available, and all participants are required to give at least one full day of their time. In terms of ethical issues, observing a simulated scenario in which a human being appears to be physically harmed and at risk could potentially cause a research participant distress. Whilst participants will be aware that they are taking part in a research exercise and therefore that what they see was staged, it is possible that a realistic looking scenario might lead them to

forget this momentarily and become upset at what they see. At this stage in the project, we do not know how much of a risk this is. In particular, due to the newness of robotic technology, we don't know the extent to which the presence of a robot in a simulated accident scenario, coupled with the potential that it may cause harm to a human, might trigger participant discomfort or distress.

The switch to remote working necessitated by the United Kingdom lockdown in response to the spread of COVID-19 shifted our attention to the use of an online format for fieldwork. We realized that we could draw on this format to continue our work on accident scenarios but do so in a way that limited the logistical and ethical challenges outlined above. Specifically, we saw the benefit in asking our online participants to role-play a scenario in which they were witnesses to the occurrence or aftermath of an accident involving a social robot. Setting up and running the data collection would be relatively quick and non-labour intensive–especially in comparison to simulating the scenario in laboratory conditions. If the accident scenario proved to be unrealistic or the witnessing process unviable, we would have opportunities to make quick alternations and try again. We could use our participants' responses to learn more about the process of witnessing accidents and also use them as testimony in accident investigation exercises in our study. In addition, the distance provided by an online platform, combined with the absence of actual robots, could create a safe space in which participants could experience hazardous interactions with a robot. We would be able to elicit their responses as if they were in the scenario, to learn about their interactions with the robot but with far less risk of making them feel uncomfortable or distressed. As such we would be putting our commitment to RRI into practice. We decided to develop a research protocol based on this online approach and trial it. As we demonstrate in this paper our trials show it to be a highly useful method. In RoboTIPS we plan to use it as a complement to (and preparation for) future laboratory-based simulations, but it can also be used as an alternative to *in situ* human-robot interaction studies. Before we describe the research protocol and its development, we spend some time discussing role-play as method and how it can contribute to the study of human-robot interaction.

Role-Play as Research Method

Broadly speaking, the term "role-play" in research describes a multi-party interaction in which individuals play out a series of actions based on taking a specific role (Lewis-Beck et al., 2004). Individuals may take on the role of an imagined other in a role-play but might also act as themselves. The technique has been widely used a tool for communication skills training in medicine and beyond (Joyner and Young, 2006; Stokoe, 2014) as well as one for language learning (Ladousse, 1995). The aim of the research-focused role-play is to investigate how participants respond to certain activities or stimuli within the interaction. The method can provide a highly effective means to simulate a scenario which is perhaps too complex or risky to stage naturalistically whilst eliciting useful data. It can also be used to elicit participant responses regarding hypothetical futures and emerging technologies, so is therefore of significant potential benefit to fields such as human-robot interaction (HRI). We conducted a literature review of role-play in HRI and found numerous references to the term, alongside references to other adjacent terms. There is an absence of consistent usage across the literature but we can broadly characterize these terms as: :"scenarios" - a combination of physical context and task created to replicate a real life situation in which human participants may or may not be involved; "simulations" which tend to be virtual scenarios or physical role-plays where human participants are optional and, if they do exist, play themselves; "narrative interactions," which tend to be role-plays with a pre-planned narrative arc; and "imaginaries," which tend to be fictional situations that come from the imagination of participants with some prompting by researchers.

This cluster of methods has been used in HRI in a number of ways. Typically, human behavior and responses to a particular HRI scenario are captured and observed through HRI experiments. These are mostly conducted in physically-situated, video-based or virtual reality contexts. Our review of the literature identified role-play (and its associated forms) deployed as a capability of robots, as a method of teaching humans and robots, as method of prototyping, and as method of conducting research. The latter two are the most common forms. Where role-play in HRI has been used as a HRI prototyping method, this work is intended both to test the performance of a specific human-robot-task-context combination and to test the methodology. The results can provide valuable insight into the human experience of interacting with robots. Tonkin et al. (2018) used role-plays to test prototype behaviours of a PAL REEM humanoid-wheeled social robot in preparation for deployment in an airport. The role-plays were conducted in a lab, where visitors interacted with the robot and provided feedback to researchers. The findings helped the team develop their design methodology by providing a mechanism for quick, early-stage feedback. Koay et al. (2020) provided further evidence for the value of narrative-based prototyping for social robots. They used episodic, narrative role-plays to prototype home companion robots. Participants interacted with multiple embodiments of a single agent in a series of 1-h role-plays, held twice a week for a month. Each session began with a narrative introduction, after which participants interacted with the robot exclusively, enabling the authors to examine user acceptance of narratively-connected scenario and user-agent relationships after embodiment migrations.

When used as a research method within HRI, role-plays have been conducted to test out a much wider range of research questions. For instance, existing studies have use role-plays to: gather input during robot design protocols (Vallès-Peris et al., 2018); attempt to reproduce "observed real-world social interactions with a robot" (Liu et al., 2016); test the development of natural language user interfaces for robots with cognitive capabilities (Green et al., 2006); plus to explore opportunities and challenges around the collaboration between humans and robots in industrial settings (Meneweger et al., 2015; Weiss et al., 2016) and identify ways to optimize this collaboration (Wurhofer et al., 2015; Weiss and Huber 2016).

Frequently these role-plays do not involve human participants interacting with actual robots; instead Wizard-Of-Oz style simulations are deployed instead. For instance, in their 2006 paper, "Measuring Up as an Intelligent Robot–On the Use of High-Fidelity Simulations for Human-Robot Interaction Research," Green et al. ran two such simulations in which participants gave a robot a tour of a staged home environment and the robot's actions were tele-operated. This proved a viable method for developing the social robots in the areas of spatial language research, whole-system conceptualization, and user attitude assessment. Staging role-plays without a real robot can help to protect vulnerable research participants. Vallès-Peris et al. (2018) used imaginaries to engage with first-grade aged school children and encourage them to share their needs, feelings and preferences around robots in healthcare contexts.

The literature on role-play in HRI research is informative but still relatively small. We were interested to note that we could not find any examples of role-play being conducted in an online format, with participants communicating over an online platform whilst physically distanced. This is a significant gap that neglects the potential of the remote format. We were also interested to look at other fields that have embraced role-play to see what we could learn about the value of the method as well as the challenges it presents. The use of role-play in areas such as education, entertainment, and design is highly illuminating here. Role-plays have been used in a wide range of formats and for a variety of purposes. For instance, physical role-plays with tokens have been used to teach farmers in Ghana (Villamor and Badmos, 2016) and virtual games have been used to teach cultural awareness to military operatives in the United States (Prasolova-Førland et al., 2013). MMORPGs like World of Warcraft have eclipsed their live-action and pen-and-paper predecessors in the realm of interactive entertainment. Designers have used research-oriented games to explore plausible futures (Coulton et al., 2016) and narrated scenarios to refine communication tool prototypes (Nielsen, 2012).

Examination of the literature reveals the importance of enabling participants to fully engage with the context presented in the role-play, by making it immersive or as realistic as possible, in order to elicit genuine responses from them. This is exemplified in Mariani (2020) paper, "Other Worlds. When Worldbuilding and Roleplay Feed Speculation," which highlights the aspects of games that make them well suited for the exploration of alternative circumstances from an HCI perspective. The paper demonstrated that in allowing people to suspend their disbelief, the worldbuilding aspect of games makes them a valuable aid for "envisioning, speculating, and framing possibilities and alternatives." Similarly, Ortiz and Harrell (2018) used their findings to argue that narrative role-plays were more effective than earlier HCI methodologies at facilitating engagement in the form of human self-reflection. Participants completed a virtual, single-party role-play, Chimeria:Grayscale via an online game the researchers had developed, and then answered a survey with system usability and game experience questions. The data from this survey suggested that Chimeria: Grayscale enabled self-reflection in participants and provided evidence for the authors' ongoing research on computer-supported self-reflection.

A final set of work that proved very illuminating in informing our work also comes from HCI and concerns the use of role-play (and other naturalistic techniques) to facilitate cultural experiences. Benford et al. (2012), Benford et al. (2015) have conducted various studies to expose participants to unusual and often uncomfortable interactions. These take many forms but have included allowing participants to first watch a breath-powered swing ride and then take on the role of controller, determining another's experiences, as well as participation in a large-scale community alternate reality game (ARG) in which participants observe a "kidnap" and sign up to be players in the game to investigate that crime with some of them ultimately being kidnapped themselves and being interrogated about what they knew. Although the scenarios are carefully designed to be physically safe, they are also designed to prompt feelings of thrill and excitement in the participant, which may tip over into discomfort or fearfulness. Benford and his collaborators (2015) argue that it is possible to conduct such work in ways that is immersive to participants in order to elicit genuine and spontaneous responses from them but that is also ethical and manages the risks involved. In some cases, this may involve the provision of consent from the participant through a process of negotiation across the encounter, rather than in an informed consent phase at the very beginning of it. This upholds the participants' rights to determine their own experiences and to withdraw if they choose but also avoids the full nature of the experience being revealed early on and so preserves opportunities for more spontaneous responses as the situation unfolds.

Insights from these various literatures helped to inform our own study design. As described further in the next section, we adopted the use of a narrator-led role-play to facilitate participant interaction in a setting involving a social robot. This would occur online rather than in person - a necessity due to COVID-19 social distancing constraints but also an arrangement that was highly time efficient and placed minimal demand on participants. In this arrangement we needed to find ways to encourage our remote participants to engage with the setting presented to them. Since our project involves a focus on incidents and accidents, we were further interested in how we could explore hazardous or uncomfortable interactions between humans and robots and do so in a way that was safe and ethical. In the next section we our novel research approach, the Virtual Witness Testimony role-play interview, in detail.

METHODS: THE VIRTUAL WITNESS TESTIMONY ROLE-PLAY INTERVIEW

As described above, we decided to conduct online interviews in which human participants role-played a response to an accident scenario. Specifically, we wanted our participants to witness the aftermath of an accident involving a social robot and a human so that we could then elicit their responses to it as if they were in that scenario and then elicit their recollections of it afterward. The work was designed to benefit our RoboTIPS study by providing

insights into processes of human witnessing and thereby help us optimize the conduct of EBB-informed accident investigation processes. As a safe and efficient means to expose participants to hazardous human-robot interactions the approach can also be used on its own or–as we intend to do in RoboTIPS–as a complement to further, more naturalistic, methods.

Development of the VWT Role-Play Interview

We began with an accident scenario. In this a human is harmed whilst a social robot is present, and the social robot perhaps caused the accident in some way:

In a supported living community for older people, assistive robots supplement human staff to provide support to residents. For instance, they can prepare drinks, carry small items, set up audio-visual entertainment, conduct basic conversation, set up telephone calls, detect falls and raise an alert when a fall has been detected. One day a neighbor of a resident named Rose, enters Rose's flat and finds her lying on the floor in need of medical attention. Rose's personal robot is nearby and is moving backwards and forwards close to Rose. It has not raised a fall alert and does not seem to be able to connect to the internet. Rose also has bruising on her legs, consistent with a robot making impact with them.

The scenario is deliberately set in an imagined future so that the robot's functional capacity is more advanced than the current state of the art allows. Our study participants would play the role of Rose's neighbor who comes into the flat and finds her on the floor. The role-play interview would elicit information about how the neighbor responds–for instance, do they call for help, do they attempt to interact with the robot etc. With the scenario in place, we needed to determine how our participants should witness the scene. We rejected the use of a video animation or interactive illustrations as too time consuming to produce and the use of robot simulation software such as Gazebo or Webot as not able to generate the contextual detail we wanted. We decided that illustrations would be suitable to depict the scenario.

At the same time as discussing the visual prompts to detail the scenario we also discussed how best to elicit role-play responses from the participants who would witness it. An initial plan to use a closed-question survey in combination with one or more illustrations was rejected in favor of an approach that could elicit more detailed responses from participants. We chose to apply a variation on the "game master" style role-play. In this, a narrator verbally introduces the participant to an imagined scenario and provides them with opportunities to explore and interact with it, with any actions having an impact on how the scenario unfolds. We could use a series of illustrations to help the participant engage with the scenario as it goes on. Whilst there was a specific setting (supported living complex) and core action point (finding Rose on the floor of her home) for the scenario, there was no fixed outcome to the narrative. Depending on the decisions made by the participant, the action could unfold in a number of ways. The participants would be asked to make decisions about what they wanted to do at certain moments in the scenario and the narrator/interviewer would respond to their choices. This particularly appealed us to as an opportunity to simulate genuine actions and interactions relating to a robot within the scenario context. To help us develop this, we took advice from an expert game master on how best to set up the scenario to make it understandable, believable, and immersive for participants. We also prepared a decision tree to establish the various possible outcomes of the scenario–for instance participants might call for medical support for Rose, they might attempt to talk to Rose's robot to find out why it had not called for help, etc.—and how the narrator/researcher would respond to them.

We iterated our study documents multiple times. We worked with professional illustrators to create images that depicted sequential moments in the scenario. To further prevent any potential for participant distress at seeing the depiction of a human coming to harm, we ensured the illustrations did not look life-like or like photographs, instead they were clearly illustrations. We also requested that Rose, the woman in our scenario who has a fall on the floor, does not appear to have any overtly visible or "gory" injuries. We turned our scenario narrative into a script for the researcher to read out and refined it to include details relevant to the core action points whilst ensuring they were embedded within wider detail and didn't stand out as too obvious. We tested out the role-play on each other and then later piloted it with research students in our institutions, making improvements based on our observations of the process. Once we were happy, and had secured appropriate Research Ethics Committee clearance, we launched the data collection with real participants. We next describe the process of recruitment and detail the exact content of our Virtual Witness Testimony (VWT) role-play interviews. The decision tree and interview script are included as **Supplementary Material** to this paper.

The Role-Play Process

Recruitment

A message was placed on a popular participant research recruitment website stating:

We are conducting a project called RoboTIPS, which explores the use of social robots such as driverless cars and robots for assistive living. As part of this we are conducting short online interviews in which we show an interviewee illustrated scenarios involving robots and humans, and the ways that robots might go wrong. We will ask the interviewee questions about the scenario. Interviews last for 30–40 min and take place on Zoom.

Participant Requirements

- Age 18 or over
- Good internet connection essential for online interview
- Those with a degree in robotics and/or medicine are not eligible to take part but participants from all other backgrounds are welcome.

Instructions

Participants who express an interest will be sent our study information sheet with further details of what taking part involves. They will be contacted to check their eligibility and

availability. Then we will set up a time for an online interview and send a Zoom link. Participants will be interviewed individually and the interviews will be recorded. Participants do not need to do any preparation ahead of the interview.

We invited only those over the age of 18 to take part so that they we could be sure participants were able to give informed consent for themselves. For the purposes of informed consent, we also needed to give some indication of what participants would see during the role-play, and in particular provide those who might be anxious about witnessing details of harm etc., an opportunity to self-select not to volunteer to take part. For that reason, we referred to robots going "wrong" in our recruitment description and participant information sheets. At the same time, we wanted to ensure that the precise scenario was unknown to participants ahead of seeing it, in order to prompt a spontaneous response from them. For that reason, we did not provide detailed information about the scenario to participants ahead of their role-play interview. In addition to age, we set exclusion criteria to make those with a high level of medical and/or robotics knowledge ineligible to take part as we wanted to focus on the responses of a lay population.

Participants responded through the website to indicate their interest, and the researcher emailed them to give a choice of dates and times for the interview along with a participant information sheet that gave fuller details of what was involved in the study and how data would be collected and handled. Once a date and time had been arranged, participants were emailed a link for the online meeting and a consent form, which they were asked to sign and return by email ahead of the call.

The VWT Role-Play Interview

Each online call took place between one researcher and one participant and involved a number of short phases. First the researcher welcomed the participant, gave a brief run through of what to expect and checked the participant was happy for the call to be recorded. Next the researcher shared their screen so that the participant could also view it and asked some simple background questions–two closed questions and one open question:

(1) Which of these best describes your age? (participants select their answer from a list shown on the screen)
(2) Which of these best describes your highest level of formal education? (participants select their answer from a list shown on the screen)
(3) Our study is about social robots–those that interact with humans as part of their day-to-day function (participants directed to look at images of social robots on screen). Have you heard of these kinds of robot and do you think you would ever consider having one in your own home?

The primary aim of this phase was to help the participant ease into the call by answering some straightforward questions and expressing their own personal viewpoint. It also presented an opportunity to build rapport between the researcher and participant as the researcher asked follow up questions about the participants' employment/topic of study etc.

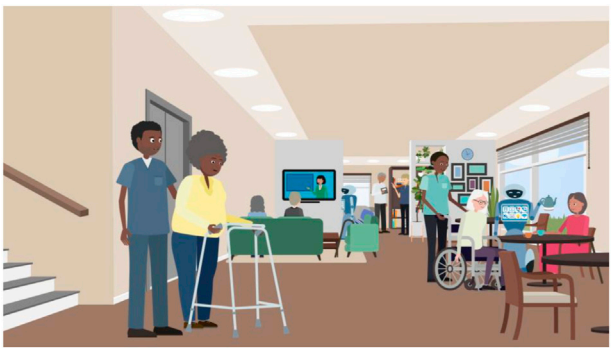

FIGURE 1 | Illustration used in VWT role-play interview showing the communal area of the supported living complex.

The researcher then moved into the role-play scenario phase by telling the participant:

We are going to talk through a scenario. You will not be you in the scenario, you will imagine you are a different person. The scenario is not real or one that occurs at the moment; it is setting in a hypothetical future. I am going to show you some pictures to help you imagine that scenario and there will be some times that you will have a chance to make decisions about what you would like to do. It will be very straightforward. You can also ask as many questions as you want to help you understand the scenario and make decisions about what you want to do in it. There is no right or wrong thing to do–it is entirely up to you.

The researcher then narrated the scenario script and gave opportunities for the participant to give responses or ask questions at certain points. To begin with the participant saw only a white screen and then images were shown at relevant points in the description. The narration began:

> The year is 2025—so a little way into the future. You are not you; you are 70 years old and you have just retired after a long career.
>
> You are in pretty good health but you're a little bit less mobile than in previous years, you move a bit slowly, your knees hurt a bit and you need to take a lot of naps. Sometimes you forget things too. Nothing very serious but you'd like to have a bit more support in daily life.
>
> You have recently moved into a retirement community that is also a supported living complex. This is in the United Kingdom. It has a communal area and you also have your own flat.
>
> First let's go into the communal area

At this point the researcher showed the participant **Figure 1** and asked: What can you see here; what kinds of activities do you think are going on?

At this point we wanted the participant to become familiar with the detail of the setting in the scenario and also note their observations of the presence of robots in it. The researcher then continued:

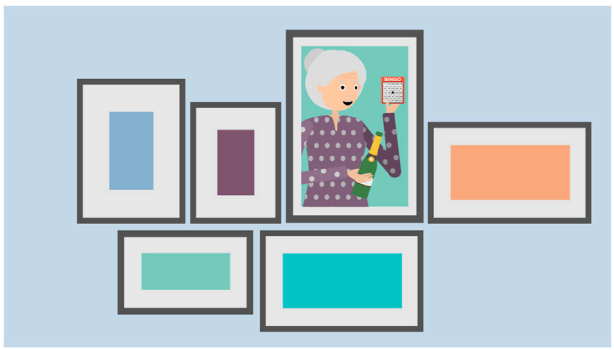

FIGURE 2 | Illustration used in VWT role-play interview showing Rose.

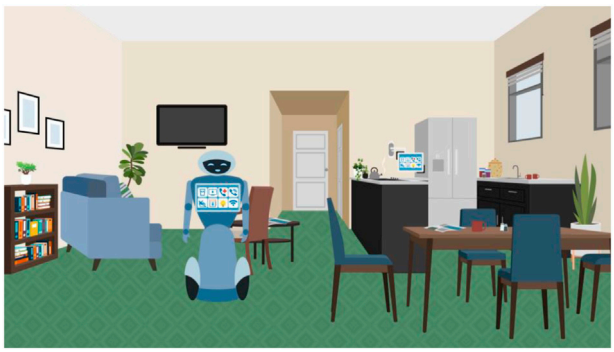

FIGURE 3 | Illustration used in VWT role-play interview showing the interior of the neighbour's flat.

> You've been here for about 2 months and you really like it here. It's very good to have staff on and when you need a bit of help—because you can't move around as much anymore, it's helpful to have staff to do some jobs for you. You like being around other people too. There are about 20 residents here and they all seem pretty nice. In particular, you have become friends with your next-door neighbour Rose. Here is a picture of Rose. (**Figure 2**)
>
> She helped you a lot when you first moved in—she helped you get to know the local area and you often go out for fun outings. You go for slow walks together as you both have slightly bad knees and you also enjoy playing games. Last week Rose won top prize in a bingo competition—as you can see from the photo, she was very happy about it! You both had such a good that you decided to go out to bingo later today. In fact, you are going to meet Rose again in a little while but before you do that, let's go inside your own flat. (**Figure 3**)

At this point, the participant was asked What can you see here? Following their response, they were asked further questions to help them become familiar with the robot in the image and its various functions.

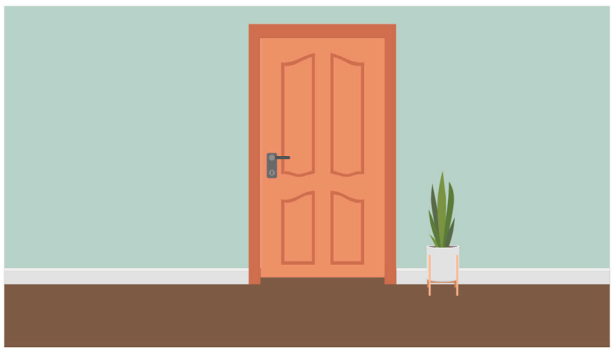

FIGURE 4 | Illustration used in VWT role-play interview showing the exterior of Rose's flat.

> Your flat has technological features that can assist you and make you more comfortable in your day-to-day life. The main feature is your own robot which links up to a smart home system. What name have you given your robot? The robot can do lots of things, some of them are shown on the screen here. The first one is to get a drink. What do you think the others are for?

The researcher and participant talked through each of the function icons shown on the robot's torso (and control panel toward the back of the room) in **Figure 3**. The top row icons indicate (left to right) providing drinks, setting up entertainment on the television, making a call for help and making a general telephone call. On the bottom row (right to left) Wi-Fi connection, turning on/off the lights, opening/closing curtains and fall detection. The key aim here was to make the participant aware of the robot's fall detection function but to do so in a way that was embedded in other detail and did not draw overt attention to it. After this the participant was told:

> You can ask your robot to do a task by pressing the button on its screen or the console in the kitchen. You can also talk to the robot to ask it do tasks or to ask it a question. Let's practise doing this now.

The participant was then encouraged to give a simple voice command to the robot, by asking it the time–with the researcher providing the reply as if they were the robot. This served to introduce the participant to the ways in which they could interact with the robot and also led in to an opportunity for the robot to "remind" the participant that it was time for them to go and meet Rose. So, the researcher then continued:

> OK so it's time to go to Rose's flat. You put your shoes and coat on and walk over to Rose's rather slowly. You are looking forward to seeing her and playing bingo and you know she is excited too as she won last time. Here you are outside the door. (**Figure 4**) You knock on her door. There is no answer. What do you do?

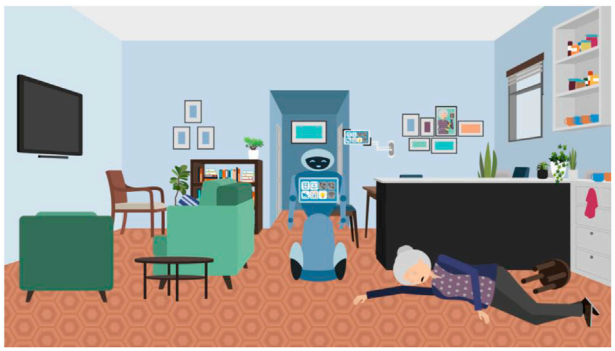

FIGURE 5 | Illustration used in VWT role-play interview showing the scene inside Rose's flat.

At this point participants had a free choice about what action to take. It was expected that most of them would attempt to get inside Rose's flat, and our decision tree mapped out various ways in which that could occur including: opening the door; asking their own robot to open the door; asking a staff member to go inside; calling for emergency support etc. If participants appeared hesitant about what to do or inclined to go somewhere else (e.g., their home, the bingo hall, etc.), they were given further detail emphasizing how unusual it was for Rose to not answer her door and the worry that something might have happened to her. However, the decision remained their own.

Once inside Rose's flat participants were shown **Figure 5** and asked What do you see here? What do you do?

Once again participants had a free choice of what to do. Our decision tree mapped out various options and how the scenario would then unfold subsequently. It was expected that they would attempt to find medical support for Rose in some way by calling an ambulance etc. It was also expected that they might interact with Rose's robot in some capacity. For instance, they might ask the robot what had happened to Rose and/or instruct it to make an emergency call. The decision tree determined that whenever they did so, the only response the robot would give would be "Can you help? Cannot connect to Internet."

After making their first response to the question the participant was then asked "What do you do now?" or "is there anything else you do now?" as an opportunity for them to take further actions. For instance, if the participant had called for an ambulance, this question would prompt them to take further actions whist waiting for it to arrive. If the participant had not yet attempted to interact with the robot, the researcher would find a suitable moment to narrate:

> The robot moves closer to you and says "Can you help? Cannot connect to Internet."

Assuming that the scenario led to the calling of an ambulance, the researcher eventually said:

> After a while the ambulance arrives and take Rose to hospital. She has broken her hip and needs to stay in for a few weeks but will make full recovery. She can't remember what happened and is not sure how she came to be on the floor.

This marked the end of the role-play phase. Alternative endings were provided in the decision tree in case the participant did not initiate an emergency response. In all cases the researcher then took away **Figure 5** and replaced it with a white screen. The researcher then said:

> The next day the manager of the supported living complex comes to you and asks some questions to try to work out what happened to Rose. Can you tell me everything you did and saw after you knocked on Rose's door? What do you think might have happened to Rose? Did you notice anything about Rose's robot—what was it saying or doing?

These final questions moved the scenario into the recollection phase. We were interested to see how and how much participants recalled the scenario and the actions they had taken within it. We were also interested to hear any speculations about what had happened to Rose and the robot's connection to the incident–for instance, why the robot might have failed to detect her on the floor and why it might not be able to connect to the Internet.

Once this phase was complete the researcher provided a short debrief to explain a bit more about the aims of the study. The participant was thanked for their time, invited to ask questions or make comments and then the call was ended. Immediately afterward the researcher emailed the participant a small shopping voucher to thank them for taking part.

Conduct of the VWT Role-Play Interviews

In November 2020 we conducted a first set of 22 VWT role-play interviews using the study design described above. All the interviews were conducted by the same researcher. They were conducted over an online teleconferencing platform (Zoom) and recorded using its embedded recording function. This produced approximately 12 h of video data, capturing all phases of the interviews. The participant responses to the researcher questions were treated as data for analysis. As this was the first time running this new approach, we primarily wanted to assess the usefulness of the method itself. Our analysis focused on determining how successful the method was in eliciting information from participants that could help us to in our work on witnessing and human-robot interactions in accident scenarios.

Our key research questions for this first round analysis were:

- Are participants able to understand the scenario and express decisions about what actions they will take when prompted by the interviewer?
- Are we able to place participants in a situation where they are part of a problematic encounter involving a social robot without causing them to experience distress or discomfort?
- Are participants able to engage with the scenario sufficiently that they provide authentic and spontaneous responses to it?

- Are we able to use this format to successfully observe participants' recollections following the scenario incident?

Our findings are discussed in the next section. As this is a methods paper, we therefore primarily focus our comments on assessing the value of the method itself. In relation to the research questions above, we discuss how the application of the approach worked in practical and ethical terms as well as the extent to which we found that the results it generated could advance the aims of our RoboTIPS study.

FINDINGS

The findings are discussed in relation to the efficiency of the method and the quality of the data it generated. Overall, these findings show that the VWT role-play interview is a promising format. It enables the quick collection of detailed data and can successfully elicit role-play responses from participants. Analysis of the data can reveal produce insights into participant perspectives on social robots, their imagined interactions with robots and their recollections following the scenario they have observed. There are some limitations, including the extent to which participants are fully immersed in the scenario and give a truly genuine response to the situation they are presented with. These limitations are familiar to role-play methods in general (Lewis-Beck et al., 2004) and in the Discussion section we go on to highlight the trade-offs that exist between practicality and immersion when conducting work in this area.

Practical Issues

Efficiency of Method

As described above, the initial drafting and piloting phase took some time to complete due to the cycles of decision making, reflection and iteration involved. However, once we were able to begin, the recruitment phase and conduct of the role-play interviews was very quick. We received a large number of immediate responses to the online recruitment request and were able to schedule calls with participants very rapidly. We aimed to fill an initial sample of 20 participants. Four days after publishing the request we had a sample of 25 signed up for interview slots (a larger number than required allowing for the likelihood of some participants dropping out) and could have set up further calls due to the many individuals who expressed an interest.

Running the study was also very time efficient. Set-up for the VWT role-play interviews was minimal for the researcher. It was necessary to keep track of the schedule of calls and be online at the right times. A standard copy of the illustrations and scenario script was used for each call; these were simply kept on file and opened in preparation for each interview. Most calls took around 25–45 min to complete, with the longest taking 55 min. Following the end of the call, the researcher filed the completed consent form, allocated a participant identification number to the participant for the purposes of anonymization and saved the recording to a secure disk. The researcher also made notes about the key points of the interview regarding the participant's spoken responses and observations. In all, the work involved in each VWT role-play interview took around 1 h. Multiple interviews could be scheduled per day and the total round of role-play interviews was conducted within a period of 10 days. Only three of the 25 individuals who signed up to participate failed to attend, representing a very manageable drop-out rate. For the participants, involvement in the study also presented a very small time burden. There was no need for them to travel to attend and there was no preparation involved–beyond completing the consent form. They simply needed to be online at the allocated time and follow the link they had been sent to join the call. As the calls were relatively quick to complete, involvement took little away from participants' day-to-day commitments and they also had a broad choice of time slots to choose from so could select one that fitted best with their own schedules. Participants were also able to decide whether to use their mobile phone, tablet, or laptop etc., when joining the call.

Ethical Considerations

We secured University Research Ethics Committee clearance for our workplan before beginning the data collection and used information sheets and consent forms for our informed consent processes. We took a number of steps to achieve a balance between maintaining some element of surprise when participants saw the scenario, in order to best elicit a genuine response from them, and providing enough detail about what participation entailed so that their consent was appropriately informed, and in particular to prevent against participants becoming distressed when witnessing a scenario that involved a human being harmed. We further ameliorated the possibility of participant distress through the use of non-photographic style illustrations and comments in the narrator script to note that the scenario participants were about to see was set in a hypothetical future and not something happening now. Across the 22 role-play interviews there were signs that participants were engaged with the scenario (see below) but no indications that they were suffering distress (becoming highly emotional and/or unable to speak or respond to questions etc.) when witnessing it.

The use of the online format also brought a specific set of ethical concerns to be considered and attended to. We chose to conduct all the calls on the platform Zoom, using an institutional license rather than a private one and ensuring we met all the University of Oxford's guidelines for best practice in terms of data collection and storage etc. Given that people are sometimes wary about the data collected by platforms and the purposes to which they are put, all participants were told they could join the call by using Zoom in their browser rather than downloading it as an app. We used Zoom's own recording function to capture the calls. All data were saved on a local external hard drive rather than to the cloud to ensure nothing was saved on third party servers (Bokhove and Downey, 2018). We also did not make use of Zoom's automatic transcription service, as this would have required us to store the recording on the cloud and potentially risked the recordings being used as further training data for the algorithmic transcription system. As Zoom captures video by default in its recording function, we told participants ahead of and at the beginning of each call that they were welcome to join

via audio only if they did not want to have their image captured. A number of participants chose to use the audio only option for the entire call and some others elected to turn off their video when the prompt was given during the call. The recording function was only turned on when the participant stated they were happy for it to begin, and participants were also alerted to the icon showing on their screen that indicated recordings were taking place.

Quality of Findings

Demographic Characteristics of Participants

The participants who took part in the study were from a relatively wide range of demographic backgrounds. Most were based in the United Kingdom–due to the recruitment site used–but were a mix of nationalities from across Europe and beyond. They also ranged in age and level of education achieved. From the answers given to the two opening questions: six were aged 18–25; 10 were aged 26–35; three were aged 36–45; one was aged 46–55 and two were aged 56–55. The educational level was a little less spread out, perhaps due to the demographic of people who access these kinds of research recruitment platforms: two participants were currently studying for an undergraduate degree; 11 had completed an undergraduate degree and nine had completed some form of postgraduate training. Although at this early stage of using the method we were not concerned with achieving a representative or proportionate spread of the overall population, these responses indicated the relative ease with which this could be achieved.

Perspectives on Social Robots

There was clear value in asking the preliminary question to participants to elicit their feelings about the development of social robots and whether they might ever consider having one themselves. In addition to "warming up" the participants to the interview process by giving them something relatively easy to answer, it also prompted some interesting perspectives. Most participants said they had heard about these kinds of robots before–either in news articles or films. Two talked about them without giving any clear personal assessment for them, seven referred to them entirely positively and four referred to them in only negative terms. Positive assessments related to how social robots could be useful for the conduct of tasks and helping older citizens or those who are isolated. Negative ones referred to robots being "creepy" (in particular when they have a humanoid form), "dangerous," or offering no value to society. The remaining 11 participants were equivocal in their assessments or referred to social robots in both positive and negative terms, for instance stating that they themselves would not like to have a robot but could see why others might find them helpful.

In addition to eliciting interesting data on participants' subjective viewpoints, the placement of this question ahead of the role-play scenario proved very fruitful. It gave the researcher the opportunity when moving on to the scenario phase to state an extra reminder that "you are not you in this scenario so you might have different feelings about robots"—particularly so for those participants who were entirely negative about social robots since engagement with the scenario required some level of acceptance of their use in a supported living context. The researcher could also bring in another reminder when the illustrations were shown to point out that in the scenario "you are happy in the accommodation where you live and have positive feelings about the robot." This guarded against the participants refusing to engage with the role-play at all (on the basis of their disapproval of social robots) whilst also helping them to feel that they had had an opportunity to put forward their actual feelings at an earlier point in the interview. For the same reasons, the decision to ask the participants to role-play as someone other than themselves was highly fruitful since it enabled them to engage and interact with the robot within the parameters of the scenario even if they feel they would not do so in "real life."

Engagement With the Role-Play Task and human-robot interaction

Analyzing the video data collected demonstrates that the participants engaged with the scenario and role-play task to a productive level and that they also engaged with the element of human-robot interaction. All participants completed the role-play task. Sometimes they needed to check what they were being asked to do or required some prompts to work out what kinds of action they could take; in particular when outside Rose's flat they were sometimes hesitant and benefitted from a prompt that they might want to find a way to get inside to check she was okay. But everyone selected an action to take at each point they were asked to and also made observations about what they could see when asked. There was substantial variety in the actions they chose, in terms of how they attempted to get inside Rose's flat and what they then did when they found her on the floor. This demonstrated to us that the role-play element was working; our participants were being provided with an opportunity to make decisions as if they were in the scenario and they were exercising that opportunity. They were making decisions on the basis of their own understandings and feelings about the scenario. A task for our analysis of the participant responses is to identify reasons or associations for the different kinds of choices our participants made. In the current data we have we can see that this is sometimes sequential in that a choice made at one point shapes a later decision. For instance, where participants elected to ask a staff member to open and go into Rose's flat, this determined that later on the staff member would take charge of checking on Rose and calling for medical help. An opportunity for future iterations of this study would be to compare the choices made by different kinds of participant; for instance, to compare individuals with experience in first aid/healthcare provision etc., with individuals without it.

In addition, all participants displayed indications of engagement with the scenario in that they drew on details of the information they had been given about it and appeared committed to selecting appropriately when asked to make a decision. For instance, when inside Rose's flat all participants attempted to get medical help for her. They frequently did so by drawing on details they had learned in the earlier part of the scenario such as finding a staff member to help, trying to ask Rose's robot to make an emergency call and/or going back to their own flat to instruct their own robot to do so when it was apparent that Rose's robot was not functioning properly. Some participants uttered expressions of sympathy such as "Oh no!" or "Poor Rose" indicating a level of affective engagement with the scenario even

though it was hypothetical. They also referred to what was and was not possible within the context of the scenario, for instance stating that "maybe my bad knees mean I can't get onto the floor to check on Rose properly" or "the staff member will be able to do first aid so I will keep out of the way."

The majority of the participants noticed that some of the function icons on Rose's robot were greyed out (see **Figure 5**) and deduced (sometimes with prompting) that this meant the robot could not perform those functions and was not connected to the internet. Several of them spent time within the scenario (typically whilst waiting for the ambulance to arrive) attempting to work out why the robot was not functioning properly and/or attempting to reconnect it to the Internet–for instance looking for a reboot switch or asking it further questions. In another sign of their engagement with the scenario, some participants continued to offer observations or suggestions for what might have happened (the robot might have tipped over the stool Rose was standing on, an obstacle might have been in the way to prevent the robot detecting Rose was on the floor etc.) during the debrief phase of the call or even by email afterward.

As the role-play was designed to stimulate human-robot interactions, we were particularly keen to assess to what extent the format worked in encouraging our participants to interact with the robot in the scenario. Again, this was largely successful. All participants bar one gave "their" robot a name (with one participant stating they would rather not do anything that humanizes a machine) and all took part in the practice questioning/instruction giving to their robot. Once inside Rose's flat, all but three of the participants noted the presence of the robot when describing what they could see and incorporated it into their decision making about what to do without any prompting from the researcher. Most tried to use Rose's robot to make an emergency call or take other actions such as pick Rose up, open a window to help her feel more comfortable etc. Two participants also tried to ask the robot what had happened to Rose. Six noted immediately that the greyed-out icons on the robot meant it was not possible for it to make a call it and looked for an alternative means to raise an alert. Seven participants decided to go back to their own flat that use their own robot to make a call for help–either after trying to use Rose's robot or as an alternative to this. All participants heard the robot stating it could not connect to the Internet–either as a consequence of their unprompted interactions with it or due to the researcher adding it in to the scenario. Seven participants took steps to identify the cause of the problem and help the robot reconnect; for instance, by trying to find a reset button on the robot or by calling a staff member. Six explicitly stated that they would not try to help the robot reconnect, either because they did not have enough knowledge to know what to do or because they wanted to focus on looking after Rose. The others gave no direct response to hearing the robot speak. Participants were also able to recall their interactions with the robot in the final phase of the interview, as discussed next.

Data Regarding Witnessing

All the participants were able to complete the final task in the VWT role-play interview. When prompted they recounted what they had done and seen, acting as witnesses providing testimony after the event. They also produced, sometimes without prompting, speculations of what might have caused Rose to fall and what problems might have occurred with the robot. This indicated to us that we could use participant recollections to help in our RoboTIPS work. For instance, they could form part of the witness testimony to be use in our mock accident investigations, and we could also analyze them further to determine what kinds of information the robot's EBB should collect in order to best complement the evidence provided by human witnesses at the scene of an accident.

We were interested to observe certain differences in between what participants said in the role-play phase of the call and then in the recollection phase. On several occasions, participants omitted details in their recollections–about what they had said to Rose, how they called for help, what they saw in Rose's room etc. This may have been because they seemed too mundane to need stating or because they had forgotten them. In particular several participants did not recollect the robot talking to them or moving closer to them; even when asked a question prompt they did not recount this information. These differences between the role-play and the recollection phase are very interesting; even though they took place immediately after one another, memory recall and/or certain other dynamics appear to play a role in the witness testimony witnesses produce. In future iterations of the study, we would like to leave a longer period of time between the role-play and recollection phases to further inspect these dynamics. Having a period of time between the two would also better reflect the circumstances witness testimony would be collected in a genuine case.

Limitations of Findings

Overall, we assessed the conduct of our first set of VWT role-play interviews very strongly. We were happy that the approach was workable and that it produced the kinds of findings that would benefit our research aims. We also see it as a highly versatile approach; the VWT role-play interview format can be used with different kinds of participants (age, occupation, familiarity with robots etc.) and enable them to observe different kinds of incidents or accidents where a robot is involved. We presented the aftermath of an accident in our role-play interview but participants could also be asked to witness and respond to the occurrence of the incident itself. These incidents can include robots of various kinds and take place in all manner of contexts that can be represented via illustrations. As we discuss later, the format can therefore contribute to work in HRI very widely. However, in our initial study we also observed a number of limitations relating to the set-up of the role-play and its conduct with our participants.

In terms of set-up, the use of static images meant we could not convey the more dynamic parts of the scenario effectively. For instance, in our original description of the scenario, Rose's robot is moving backwards and forwards in a way that suggested she might have collided with Rose's legs. However, it proved difficult to convey this in our remote format and it was largely dropped–beyond the visual of the robot standing close to Rose and a narrator statement about the robot moving closer to the

participant. As a result, only two participants mentioned the robot colliding with Rose as a possibility. Another issue with the set-up was that it was very important to spend time in the preparation and iteration of the narrative and illustrations to ensure that potential areas of non-understanding can be identified and corrected. Although we spent a great deal of time in our preparations, some problematic details did slip through. A number of participants assumed that their robot immediately went with them to Rose's flat. This could easily be corrected by the interview/narrator during the interview itself with a statement that "you have left your robot at home" and later by inserting into the script "as you leave your flat, you say goodbye to your robot." It also turned out that the reference playing "bingo" was highly culturally specific and several of our international participants were unclear what it meant. This was harder to resolve as bingo was embedded into our second image so could not be dropped from the narrative and it was harder to give an explanation about what bingo is and why people go out to play it entailed stepping outside of the scenario narrative so was not optimum at this point in the call.

Regarding the conduct of the VWT role-play interview, it was not possible to provide participants with entirely the same version of the script and experience of the scenario. The ways in which they asked questions or produced responses to what they saw meant that there were always slight differences even before participants were asked to make decisions about their actions in the scenario. This is not necessarily a limitation (and is a feature of all but the most structured role-play methods) as the capacity for participants to intervene in the unfolding of the scenario enabled their engagement with it. But it does reduce the potential for close comparison in the analysis. Another issue we observed was that participants might at times be looking to provide what they felt was the "correct" answer for the research study rather than an authentic one. Research participants attempting to perform "well" or "correctly" is a well-known phenomenon (Orne and Washington, 2000) and here it potentially manifested itself in the decisions we asked them to make. In several cases they chose more complex actions involving robots ahead of more straightforward ones without them–for instance only two participants said they would try to open Rose's door to see if it was unlocked and several of them chose to go back to their own flat to collect their robot without attempting this first. As our participants knew the project was about robots, they may have felt that we were looking for them to give answers that demonstrated their awareness of the robots in the scenario. They therefore turned to details in the scenario that had been presented to them–e.g., the use of the robot to conduct certain tasks. A possible means to limit this might be to spend longer in the set-up phase of the scenario by providing more detail of the setting and the kinds of activities going on there. This could have the effect of helping the participants to immerse themselves in the scenario further (see below) so that they were less likely to be conscious of the study details and less likely to feel an obligation to center their answers around robots. It would however, make running the exercise longer and more time-consuming for researcher and participants.

The largest limitation relates to the extent to which participants could or could not fully immerse themselves in the scenario and give responses that took the scenario fully seriously. As noted, our participants did display significant indications of engagement with the scenario; however, at the same time the remote format limited the extent to which they could become fully immersed in it. Our participants were calling in from home or work etc., and surrounded by features that might draw their attention–such as people walking into the room, mobile phone notifications, glitches with internet connections etc.—and distract them from the scenario. The less immersed they were in the scenario the less concentration or effort they may have put into treating it seriously. There were a couple of initial jokey responses made when we asked participants would they would do on seeing Rose fallen on the floor, such as "I'd find someone else to go to bingo with!" and it is possible that lack of detail in later recollections may stem from not concentrating on the scenario fully in the role-play phase. We would expect that greater immersion in the scenario would lead to more careful consideration of the responses participants produce and therefore it is important to reflect on how we might be able to increase this element of the process whilst also maintaining its convenience for participants. For instance, the inclusion of sound effects, more illustrations or greater detail in the set-up phase. Ultimately however, we cannot guarantee that our participants will definitely respond in a genuine manner. But this is true of all role-play since there is no guarantee that any simulation elicits people's "real" reactions - whether role-playing as themselves or taking on the role of another. This is acknowledged as a limitation of the role-play method in general (Lewis-Beck et al., 2004) but does not mean that the approach cannot provide useful findings. In particular, a role-play can provide a safe means to test out an interaction that would be impractical or risky to set up in a fully naturalistic way. Therefore, it is very fruitful for the exploration of potentially hazardous interactions between robots and humans. As our study demonstrates, virtual role play studies can also be highly time efficient and convenient in their conduct, allowing a large amount of data to be collected in a short time. We therefore feel that whilst it is important to acknowledge this limitation, we can still recognize value of using the VWT-role play format as providing a reasonable trade-off between immersion and practical/safety issues. In the following discussion we highlight the value of our VWT role-play interview approach to the study of hazardous human-robot interactions.

DISCUSSION

In this methods paper we have described the development and first use of an online role-play method within our research project on responsible robotics. The development of the Virtual Witness Testimony (VWT) role-play interview emerged as a creative response to the requirement for socially distanced fieldwork during the COVID-19 pandemic as well as the commitment of the RoboTIPS study to ethical best practice and the principles of Responsible Research and Innovation (RRI). We wanted to engage our participants in a scenario that included witnessing a human being coming to physical harm whilst a social robot was also present and had apparently had a malfunction of some kind. We needed to make sure that participants were physically safe

and not at risk of coming to emotional harm whilst they witnessed this scenario. Taking up this approach enabled us to conduct responsible fieldwork whilst researching the development of responsible robotics. As the above findings have shown, the VWT role-play interview method is a very promising one. It allows for the efficient collection of data involving a wide range of participants and can elicit useful information from them. It can engage participants to deliver considered responses about what they would do in the scenario they are presented with and then to give their recollections on what they observed in the scenario. In particular it can elicit imagined interactions between humans and social robots. The findings of our first VWT role-play interview study are highly useful to our RoboTIPS study. They will inform our ongoing work on accidents, for instance by helping us to understand how individuals might respond when witnessing an accident scenario and what kinds of testimony they produce following it. This will also assist our work on accident investigations, helping us to identify what kind of witness testimony humans can produce following an accident and what kind of data a robot Ethical Black Box can provide to best complement this. We plan to run further VWT role-play interviews based on different accident scenarios and involving different kinds of participants. These will collect valuable data that we can use for analysis and will also help us to test out the practical and ethical viability of simulating the same scenarios in face-to-face laboratory conditions. In the rest of this discussion we describe opportunities for the broader use of the method, to benefit HRI and associated fields.

The VWT role-play method is highly flexible and its format can be adapted in a number of ways. Various different scenarios can be used within it and narratives created to guide participants through the direct witnessing of an accident involving a social robot or its aftermath. Since the method works well at eliciting responses, perspectives and recollections, participants can be required both to respond to the scenario and then give their recollections and comments on it afterward. This can illuminate research into human interaction with social robots in a wide range of contexts, including those that are set in an imagined future or are too hazardous to be observed *in situ*. Further data can also be collected on human perspectives and attitudes around social robots. The recruitment and data collection process is very quick, so it is possible to conduct VWT role-play interviews with a large number of participants and aim for a demographically representative sample. Researchers can then look for any systematic differences in response according to occupation, age, gender, nationality etc., of participants. In our RoboTIPS study we are interested to compare the responses given by participants to the background question about their views on social robots and their responses to the scenario itself. In particular, we are interested to observe whether those with more negative views about social robots are more likely to perceive the robot in the scenario as hazardous and to "blame" it for the harm caused to Rose. In future uses of the method we will widen the time gap between the role-play and recollection phase, to better test the effects of memory on recall about the scenarios and speculations around the cause of the problem with the robot. We also plan to run the same scenario but with a new narrative that requires participants to take on the role of the first medical responder on the scene who attends to Rose. The participants we recruit for this will all have medical or first aid training to ensure they have the background knowledge necessary for this role. We can then compare their responses to those of the lay population, to see if there are differences in the ways they interact with Rose, interact with the robot and speculate on the causes of Rose's fall and the robot malfunction. Similarly, when other scenarios are used for VWT role-play interviews, different participants with relevant characteristics can take up different roles within them, enabling comparison across the groups. The potential to include a large number and diverse range of participants in VWT interview role-play studies is a benefit that brings analytic rigor to the approach. It is also a further way to uphold principles of Responsible Research and Innovation by bringing multiple stakeholders into the processes of research.

The VWT role-play interview has merit as standalone method but another benefit is that it can be combined with other research methods to consolidate the value of both. In RoboTIPS, the opportunity to role-play a planned scenario online in this format before running it face-to-face is hugely beneficial. We are able to test out the logic and believability of the scenario in this less time-intensive format before committing to a much more laborious face-to-face version. We were also able to check the extent to which placing participants in a scenario that involves a potentially hazardous interaction with a robot might risk causing them emotional distress in some way. The online format and use of illustrations provide a safe space to trial this scenario and creates much less risk of distress than when exposing participants to the scenario face-to-face. We can draw on what we learn online iterations of the scenario and make accommodations to ensure that when it is run face-to-face it similarly avoids causing participant harm. In addition to informing the design of our later work, the VWT role-play interview findings we gathered can be used for analytic comparison. We will compare the observations and witness testimony provided by participants in the face-to-face scenarios with the data collected in the online version. This will help us to assess the generalisability of our findings and also prompt us to identify the reasons for any differences between the two. Looking more broadly, other research using this format can reap the same benefits. Creating a scenario and beginning with a streamlined online role-play of it is a highly efficient way to test out its robustness in terms of believability and capacity for participants to engage with it, before running it in a more time and resource intensive face-to-face format. It can also be an extremely important step in testing out the extent to which the role-play might cause participants to encounter distress or discomfort allowing adaptions to be made before running the riskier face-to-face format. This is particularly important for research in the field of social robotics where–as our own findings show–many citizens are uncomfortable, even fearful, about the idea of robots being part of everyday life. Finally, collecting data from both online and face-to-face versions of a role-play provides opportunities to compare results and triangulate findings. The combination of the two methods can work to address the limitations of each. The online version is more efficient, which allows for a larger number of

participants to be involved at the loss of some level of engagement. The face-to-face version immerses the participant more fully in the scenario and may therefore elicit more reasoned and genuine responses; however, it is far more time intensive and likely allows a significantly smaller number of participants.

We view Responsible Robotics as including the application of Responsible Research and Innovation to the field of robotics. Responsible Robotics therefore requires careful consideration across issues in the design, manufacture, operation, repair and end-of-life recycling of robots to identify practices that seek the most benefit to individuals and society, and the least harm to the environment (Winfield et al., 2021). In order to be responsible, researchers in robotics need to consider the potential positive and negative impacts of the robotic systems they develop, as well as the processes through which they conduct their work. This relates to the inclusion of stakeholders, treatment of human participants, attention to environmental concerns, and communication of findings to lay audiences. As researchers in this field, we have a responsibility to consider what happens when things go wrong in human-robot interaction scenarios. This could be in the context of technical malfunctions which disrupt a robot's intended function, but also, for example, the intentional (mis)use and abuse of robotic systems to cause harm. We need to test out these scenarios and involve human participants as stakeholders. However, we need to do this in a way that is both analytically efficient and avoids causing participants distress or any other kind of harm. We have set out the Virtual Witness Testimony role-play interview as an approach that can achieve this–either on its own or used in combination with other methods. We have demonstrated the value of the method in a study conducted as part of our own research study and propose that it can be adapted to investigate other ethically-hazardous HRI scenarios, offering a safe and practical alternative to be considered alongside physically situated or virtual reality HRI experiments.

DATA AVAILABILITY STATEMENT

Restrictions apply to the datasets: The datasets presented in this article are not readily available due to ethical constraints. The data collected consist of video recorded interviews with participants and therefore constitute personal data. Participants were assured that their taking part would be kept confidential and that the data would not be shared with others. Participants were given the option to consent to their anonymised data being placed in a repository at the end of the project in keeping with the requirements of our funding body. Where they have consented, these data will be made available on request under strict ethical conditions.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the University of Oxford, Computer Science Departmental Research Ethics Committee. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors contributed to the development and refinement of the study approach described in the paper. They also all contributed to the original conception of the paper and/or its planning and drafting. HW took responsibility for the initial drafts and MD conducted the literature review. AVM added content to the Methods and Discussion sections. MD, AVM, and MJ provided refinements and redrafts to the text.

FUNDING

RoboTIPS: Developing Responsible Robots for the Digital Economy is an EPSRC Established Career Fellowship awarded to Professor Marina Jirotka. Award reference EP/S005099/1.

ACKNOWLEDGMENTS

We would like to thank all the individuals who participated in the research study described in this paper. We are also very grateful to the reviewers for their comments on this paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frobt.2021.644336/full#supplementary-material

REFERENCES

Benford, S., Greenhalgh, C., Anderson, B., Jacobs, R., Golembewski, M., Jirotka, M., et al. (2015). The Ethical Implications of HCI's Turn to the Cultural. *ACM Trans. Comput.-Hum. Interact.* 22 (5), 1–37. doi:10.1145/2775107

Benford, S., Greenhalgh, C., Giannachi, G., Walker, B., Marshall, J., and Rodden, T. (2012). *"Uncomfortable Interactions" Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'12)*. New York, NY, USA: Association for Computing Machinery, 2005–2013. doi:10.1145/2207676.2208347

Bokhove, C., and Downey, C. (2018). Automated Generation of 'good Enough' Transcripts as a First Step to Transcription of Audio-Recorded Data. *Methodological Innov.* 11, 205979911879074. doi:10.1177/2059799118790743

Coulton, P., Burnett, D., and Gradinar, A. (2016). "Games as Speculative Design: Allowing Players to Consider Alternate Presents and Plausible Features," in DRS Conference Papers. Available at https://dl.designresearchsociety.org/drs-conference-papers/drs2016/researchpapers/2/. doi:10.21606/drs.2016.15

Green, A., Topp, E. A., and Hüttenrauch, H. (2006). "Measuring up as an Intelligent Robot – on the Use of High-Fidelity Simulations for human-robot interaction Research.," in Performance Metrics for Intelligent Systems workshop, Gaithersburg, Maryland USA (Gaithersburg: National Institute of Standards and Technology), 21–23.

Joyner, B., and Young, L. (2006). Teaching Medical Students Using Role Play: Twelve Tips for Successful Role Plays. *Med. Teach.* 28 (3), 225–229. doi:10.1080/01421590600711252

Koay, K. L., Syrdal, D. S., Dautenhahn, K., and Walters, M. L. (2020). A Narrative Approach to Human-Robot Interaction Prototyping for Companion Robots, *Paladyn, J. Behav. Robotics*, 11(1), 66–85. doi:10.1515/pjbr-2020-0003

Ladousse, G. (1995). *Role Play*. New York: Oxford University Press.

Lewis-Beck, M., Bryman, A., and Futing Liao, T. (2004). *The SAGE Encyclopedia of Social Science Research Methods*. Thousand Oaks, CA: Sage Publications, Inc. doi:10.4135/9781412950589

Liu, P., Glas, D. F., Kanda, T., Ishiguro, H., et al. "Data-Driven HRI: Learning Social Behaviors by Example From Human–Human Interaction," in *IEEE Transactions on Robotics* IEEE Xplore, vol. 32 (4), 988–1008. doi:10.1109/TRO.2016.2588880

Mariani, I. (2020). "Other Worlds. When Worldbuilding and Roleplay Feed Speculation.," in *Design, User Experience, and Usability. Design for Contemporary Interactive Environments*. Editors A. Marcus and E. Rosenzweig (Springer International Publishing), 482–495. doi:10.1007/978-3-030-49760-6_34

Meneweger, T., Daniela, W., Verena, F., and Manfred, T. (2015). "Working Together with Industrial Robots: Experiencing Robots in a Production Environment," in *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (IEEE Xplore), 833–838. doi:10.1109/ROMAN.2015.7333641

Nielsen, L. (2012). "Acting as Someone like Me - Personas in Participatory Innovation," in Participatory Innovation Conference 2012, Melbourne Australia. Available at https://pure.itu.dk/ws/files/38903074/PIN_C_Acting_as_me_final.pdf. doi:10.21236/ada571843

Orne, M. T., and Whitehouse, W. G. (2000). "Demand Characteristics," in *Encyclopedia of Psychology*. Editor A. E. Kazdin (Washington, D.C.: American Psychological Association and Oxford Press), 469–470.

Ortiz, P., and Harrell, D. F. (2018). "Enabling Critical Self-Reflection through Roleplay with Chimeria," " in Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play, Association for Computing Machinery (New York: ACM Digital Library), 353–364. doi:10.1145/3242671.3242687

Owen, R., Stilgoe, J., Macnaghten, P., Gorman, M., Fisher, E., and Guston, D. (2013). A Framework for Responsible Innovation. *Responsible innovation: managing responsible emergence Sci. innovation Soc.* 31, 27–50. doi:10.1002/9781118551424.ch2

Prasolova-Forland, E., Fominykh, M., Darisiro, R., and Morch, A. I. (2013). "Training Cultural Awareness in Military Operations in a Virtual Afghan Village: A Methodology for Scenario Development.," in 46th Hawaii International Conference on System Sciences, 2013, 903–912. doi:10.1109/HICSS.2013.571

Rome Declaration on Responsible Research and Innovation in Europe (2014). *Rome Declaration on Responsible Research and Innovation in Europe*. Brussels: European Commission https://ec.europa.eu/research/swafs/pdf/rome_declaration_RRI_final_21_November.pdf.

Schomberg, R. V., and Hankins, J. (2019). *International Handbook on Responsible*. Cheltenham: Innovation Edward Elgar Publishing. doi:10.4337/9781784718862

Stilgoe, J., Owen, R., and Macnaghten, P. (2013). Developing a Framework for Responsible Innovation. *Res. Pol.* 42 (9), 1568–1580. doi:10.1016/j.respol.2013.05.008

Stokoe, E. (2014). The Conversation Analytic Role-Play Method (CARM): A Method for Training Communication Skills as an Alternative to Simulated Role-Play. *Res. Lang. Soc. Interaction* 47 (3), 255–265. doi:10.1080/08351813.2014.925663

Tonkin, M., Vitale, J., Herse, S., Williams, M.-A., Judge, W., and Wang, X. (2018). "Design Methodology for the UX of HRI," in Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (New York: Association for Computing Machinery), 407–415. doi:10.1145/3171221.3171270

Vallès-Peris, N., Angulo, C., and Domènech, M. (2018). Children's Imaginaries of Human-Robot Interaction in Healthcare. *Ijerph* 15 (5), 970. doi:10.3390/ijerph15050970

Villamor, G. B., and Badmos, B. K. (2016). Grazing Game: A Learning Tool for Adaptive Management in Response to Climate Variability in Semiarid Areas of Ghana. *Ecol. Soc.* 21 (1). https://www.jstor.org/stable/26270341. doi:10.5751/ES-08139-210139

Webb, H., Jirotka, M., F.T. Winfield, A., and Winkle, K. (2019). Human-robot Relationships and the Development of Responsible Social Robots. *ACM Proc. Halfway Future Symp.* 1–7. doi:10.1145/3363384.3363396

Weiss, A., and Huber, A. (2016). User Experience of a Smart Factory Robot: Assembly Line Workers Demand Adaptive Robots. 1606.03846.

Weiss, A., Huber, A., Minichberger, J., and Ikeda, M. (2016). First Application of Robot Teaching in an Existing Industry 4.0 Environment: Does It Really Work? *Societies* 6 (3), 20. doi:10.3390/soc6030020

Wurhofer, D., Meneweger, T., Fuchsberger, V., and Tscheligi, M. (2015). "Deploying Robots in a Production Environment: A Study on Temporal Transitions of Workers' Experiences," in *Human-Computer Interaction – INTERACT 2015*, Editors A. Julio, et al. (Springer International Publishing), 203–220. doi:10.1007/978-3-319-22698-9_14

Winfield, A. F. T., and Jirotka, M. (2017). "The Case for an Ethical Black Box," in *Towards Autonomous Robotic Systems. TAROS 2017. Lecture Notes in Computer Science*. Editors Y. Gao, S. Fallah, Y. Jin, and C. Lekakou (Cham: Springer), 10454.

Winfield, A. F. T., Winkle, K., Webb, H., Lyngs, U., Jirotka, M., MacraeC Jirotka, M., et al. (2021). "Robot Accident Investigation: a Case Study in Responsible Robotics," in *Software Engineering for Robotics*. Editors (Cham: Springer). Available at: arXiv:2005.07474v1.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling Editor declared a past co-authorship with with the authors HW and MJ.

frontiers in Robotics and AI

ORIGINAL RESEARCH
published: 26 July 2021
doi: 10.3389/frobt.2021.665729

IEEE P7001: A Proposed Standard on Transparency

Alan F. T. Winfield[1], Serena Booth[2], Louise A. Dennis[3], Takashi Egawa[4], Helen Hastie[5], Naomi Jacobs[6], Roderick I. Muttram[7], Joanna I. Olszewska[8], Fahimeh Rajabiyazdi[9], Andreas Theodorou[10], Mark A. Underwood[11], Robert H. Wortham[12] and Eleanor Watson[13]*

[1]Bristol Robotics Laboratory, UWE Bristol, Bristol, United Kingdom, [2]Computer Science and AI Laboratory (CSAIL), MIT, Cambridge, MA, United States, [3]Department of Computer Science, University of Manchester, Manchester, United Kingdom, [4]NEC Corporation, Tokyo, Japan, [5]Department of Computer Science, Heriot-Watt University, Edinburgh, United Kingdom, [6]ImaginationLancaster, Lancaster Institute for Contemporary Arts, University of Lancaster, Lancaster, United Kingdom, [7]Fourth Insight Ltd, Ewhurst, United Kingdom, [8]School of Computing and Engineering, University of the West of Scotland, Paisley, United Kingdom, [9]Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada, [10]Department of Computing Science, Umeå University, Umeå, Sweden, [11]Synchrony Financial, Stamford, CT, United States, [12]Department of Electronic and Electrical Engineering, University of Bath, Bath, United Kingdom, [13]Nell Watson Ltd., Carrickfergus, United Kingdom

This paper describes IEEE P7001, a new draft standard on transparency of autonomous systems[1]. In the paper, we outline the development and structure of the draft standard. We present the rationale for transparency as a measurable, testable property. We outline five stakeholder groups: users, the general public and bystanders, safety certification agencies, incident/accident investigators and lawyers/expert witnesses, and explain the thinking behind the normative definitions of "levels" of transparency for each stakeholder group in P7001. The paper illustrates the application of P7001 through worked examples of both specification and assessment of fictional autonomous systems.

Keywords: transparency, explainability, autonomous systems, robot ethics, AI ethics

OPEN ACCESS

Edited by:
Masoumeh Mansouri, University of Birmingham, United Kingdom

Reviewed by:
Pablo Jiménez-Schlegl, Consejo Superior de Investigaciones Científicas (CSIC), Spain
Benjamin Kuipers, University of Michigan, United States

****Correspondence:***
Alan F. T. Winfield
alan.winfield@brl.ac.uk

Specialty section:
This article was submitted to Ethics in Robotics and Artificial Intelligence, a section of the journal Frontiers in Robotics and AI

Received: *08 February 2021*
Accepted: *05 July 2021*
Published: *26 July 2021*

Citation:
Winfield AFT, Booth S, Dennis LA, Egawa T, Hastie H, Jacobs N, Muttram RI, Olszewska JI, Rajabiyazdi F, Theodorou A, Underwood MA, Wortham RH and Watson E (2021) IEEE P7001: A Proposed Standard on Transparency. Front. Robot. AI 8:665729. doi: 10.3389/frobt.2021.665729

1 INTRODUCTION

There is broad agreement in the AI and robot ethics community about the need for autonomous and intelligent systems to be transparent; a survey of ethical guidelines in AI (Jobin et al., 2019) reveals that transparency is the most frequently included ethical principle, appearing in 73 of the 84 (87%) sets of guidelines surveyed. It is clear that transparency is important for at least three reasons: 1) autonomous and intelligent systems (AIS) can, and do, go wrong, and transparency is necessary to discover how and why; 2) AIS need to be understandable by users, and 3) without adequate transparency, accountability is impossible.

It is important to note that transparency does not come for free. Transparency and explainability are properties that AIS may have more or less of, but these properties are not hardwired–they must be included by design. However, sometimes transparency might be very difficult to design in, for instance in "black box" systems such as those based on Artificial Neural Networks (including Deep Machine Learning systems), or systems that are continually learning.

This paper describes IEEE P7001, a new draft standard on transparency (IEEE, 2020). P7001 is one of the P70XX series of "human standards" emerging from the IEEE Standards Association global

[1]This paper solely represents the views of the authors and does not necessarily represent a position of either the IEEE P7001 Working Group, IEEE or IEEE Standards Association.

initiative on the ethics of autonomous and intelligent systems (IEEE, 2019b). For an overview, see Winfield (2019).

In this paper, we outline the development and structure of P7001. We present the rationale for both transparency and explainability as measurable, testable properties of autonomous systems. We introduce the five stakeholder groups in P7001: users, the general public and bystanders, safety certification agencies, incident/accident investigators and lawyers/expert witnesses. For each of these stakeholders, we outline the structure of the normative definitions of "levels" of transparency.

We will show how P7001 can be applied to either assess the transparency of an existing system–a process of System Transparency Assessment (STA)–or to specify transparency requirements for a system prior to its implementation–a process of System Transparency Specification (STS). We will illustrate the application of P7001 through worked examples of both the specification (STS) and assessment (STA) of fictional autonomous systems.

This paper proceeds as follows. In **Section 2**, we briefly survey the literature on transparency and explainability as a prelude to, in **Section 3**, introducing and justifying the definitions for transparency and explainability in P7001. **Section 3** also describes the scope and structure of P7001 including each of the five stakeholder groups, and the way P7001 approaches the challenge of setting out testable, measurable levels of transparency for each of these stakeholder groups. In **Section 4**, we describe how P7001 may be used through the two processes of System Transparency Assessment (STA) and System Transparency Specification (STS), then outline case studies for each, in order to illustrate the application of P7001. **Section 5** concludes the paper with a discussion of both the value and the limits of P7001.

2 RELATED WORK

The term transparency emerged in the 1990s in the context of information management (Ball, 2009). Nowadays, transparency has become of prime importance in the design and development of autonomous systems (Alonso and de la Puente, 2018), intelligent systems (Olszewska, 2019) as well as human-machine teaming (Tulli et al., 2019; Vorm and Miller, 2020) and human-robot interactions (HRI) (Cantucci and Falcone, 2020).

Transparency can be defined as the extent to which the system discloses the processes or parameters that relate to its functioning (Spagnolli et al., 2016). Transparency can also be considered as the property that makes it possible to discover how and why the system made a particular decision or acted the way it did (Chatila et al., 2017), taking into account its environment (Lakhmani et al., 2016). Indeed, at the moment, there is no single definition of transparency in the literature (Theodorou et al., 2017; Larsson and Heintz, 2020), as it varies depending on its application domain (Weller, 2019) and its dimensions (Bertino et al., 2019). The notion of transparency is also often interwoven with other related concepts such as fairness (Olhede and Rodrigues, 2017), trustworthiness (Wortham, 2020; Nesset et al., 2021), interpretability (Gilpin et al., 2018), accountability (Koene et al., 2019), dependability (TaheriNejad et al., 2020), reliability (Wright et al., 2020), and/or safety (Burton et al., 2020).

The closely related study of explainability has become popular in recent years with the rise of Artificial Intelligence (AI) and AI-based systems (Adadi and Berrada, 2018; Baum et al., 2018; Gunning et al., 2019). This has led to the new field of explainable AI (XAI) (Barredo Arrieta et al., 2020; Confalonieri et al., 2021), which is concerned with the ability to provide explanations about the mechanisms and decisions of AI systems (Doshi-Velez and Kim, 2017; Lipton, 2018).

Current research in XAI focuses on the development of methods and techniques to understand and verify AI-based autonomous and/or intelligent systems (Páez, 2019; Dennis and Fisher, 2020). Explaining AI applications, especially those involving Machine Learning (ML) (Holzinger, 2018), and Deep Neural Networks (DNN) (Angelov and Soares, 2020; Booth et al., 2021), is howbeit still an ongoing effort, due to the high complexity and sophistication of the processes in place (e.g., data handling, algorithm tuning, etc.) as well as the wide range of AI systems such as recommendation systems (Zhang and Chen, 2020), human-agent systems (Rosenfeld and Richardson, 2019), planning systems (Chakraborti et al., 2020), multi-agent systems (Alzetta et al., 2020), autonomous systems (Langley et al., 2017), or robotic systems (Anjomshoae et al., 2019; Rotsidis et al., 2019).

3 P7001 SCOPE AND STRUCTURE

The aim of P7001 is to provide a standard that sets out "measurable, testable levels of transparency, so that autonomous systems can be objectively assessed and levels of compliance determined" (IEEE, 2020). An autonomous system is defined in P7001 as "a system that has the capacity to make decisions itself, in response to some input data or stimulus, with a varying degree of human intervention depending on the system's level of autonomy".

The intended users of P7001 are specifiers, designers, manufacturers, operators and maintainers of autonomous systems. Furthermore P7001 is generic; it is intended to apply to all autonomous systems including robots (autonomous vehicles, assisted living robots, drones, robot toys, etc.), as well as software-only AI systems, such as medical diagnosis AIs, chatbots, loan recommendation systems, facial recognition systems, etc. It follows that P7001 is written as an "umbrella" standard, with definitions of transparency that are generic and thus applicable to a wide range of applications regardless of whether they are based on algorithmic control approaches or machine learning.

3.1 Defining Transparency in P7001

The UK's Engineering and Physical Science Research Council (EPSRC) Principles of Robotics–the first national-level policy on AI–states, as principle four: "Robots are manufactured artifacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent" (Boden et al., 2017). The EPSRC definition of

transparency emphasises, through contrast, that transparency in robotics means that the end user is well aware of the manufactured and thus artificial nature of the robot.

Since the release of the EPSRC Principles of Robotics, numerous guidelines and other soft policy declarations have been released by governmental, intergovernmental, non-governmental, and private organisations where transparency is one of the most mentioned ethical principles (Jobin et al., 2019). Yet, each provides its own–vague–definition. For example, the European Commission's High-Level Expert Group (HLEG) on their "Guidelines for Trustworthy AI" considers transparency to be one of its seven key principles and defines it as a combination of three elements: traceability, explainability, and communication (EC, 2018). Another prominent intergovernmental organisation, the OECD, in its AI ethics guidelines considers transparency as the means of understanding and challenging the outcomes of decisions made by intelligent systems (OECD, 2019). As we saw in the previous section, a similar disagreement exists in the academic literature, where each scholar in transparency-related research has their own definition.

Arguably there can be no universally accepted definition for any given ethical value (Theodorou and Dignum, 2020). Instead, as P7001 is self-contained, an actionable and explicit definition of transparency is required. Thus P7001 defines transparency as "the transfer of information from an autonomous system or its designers to a stakeholder, which is honest, contains information relevant to the causes of some action, decision or behavior and is presented at a level of abstraction and in a form meaningful to the stakeholder."

P7001 recognises that AI technology cannot be separated from the larger socio-technical system of which it is a component, hence the explicit reference to the designers of the system as responsible agents in providing relevant information. That information, depending on the stakeholder to whom it is targeted, can be anything from records of development decisions to interactive manuals. Further, the keyword *honest* emphasises that only information that is neither false or deceptive can be considered as compliant to the standard.

Furthermore "to consider an autonomous system transparent to inspection, the stakeholder should have the ability to request meaningful explanations of the system's status either at a specific moment or over a specific period or of the general principles by which decisions are made (as appropriate to the stakeholder)" (Theodorou et al., 2017). This allows the consideration of transparency not only as a *real-time* property, but also as the means of ensuring *traceability* for past events to aid incident investigators (Winfield et al., 2021) and when necessary ensure accountability (Bryson and Theodorou, 2019).

As with transparency, there are multiple definitions for explainability in the literature (Barredo Arrieta et al., 2020). P7001 defines explainability as "the extent to which the internal state and decision-making processes of an autonomous system are accessible to non-expert stakeholders". Again, this is not an attempt to provide a universally-accepted definition, but rather a *workable* one. The relationship between transparency and explainability in P7001 is that the latter is transparency that is *accessible* to non-experts. In P7001 explainability is a subset of transparency. P7001 defines explainability to stay close to existing literature, while also taking into consideration the multi-stakeholder approach and the wide spectrum of autonomous systems the standard is meant to cover. Thus its normative requirements aim to satisfy both definitions of transparency and explainability. It is also important to note that providing an explanation does not necessarily make a system's actions completely transparent (De Graaf and Malle, 2017).

3.2 Transparency Is Not the Same for Everyone

Transparency is not a singular property of systems that would meet the needs of all stakeholders. In this regard, transparency is like any other ethical or socio-legal value (Theodorou et al., 2017). Clearly a naive user does not require the same level of understanding of a robot as the engineer who repairs it. By the same reasoning, a naive user may require explanations for aspects of reasoning and behaviour that would be obvious and transparent to developers and engineers.

P7001 defines five distinct groups of stakeholders, and AIS must be transparent to each group, in different ways and for different reasons. These stakeholders split into two groups: non-expert end users of autonomous systems (and wider society), and experts including safety certification engineers or agencies, accident investigators, and lawyers or expert witnesses. Stakeholders are beneficiaries of the standard, as distinct from users of the standard: designers, developers, builders and operators of autonomous systems.

Let us now look at the transparency needs of each of these five groups.

3.2.1 Transparency for End Users

For users, transparency (or explainability as defined in P7001) is important because it both builds and calibrates confidence in the system, by providing a simple way for the user to understand what the system is doing and why.

Taking a care robot as an example, transparency means the user can begin to predict what the robot might do in different circumstances. A vulnerable person might feel very unsure about robots, so it is important that the robot is helpful, predictable—never does anything that frightens them—and above all safe. It should be easy to learn what the robot does and why, in different circumstances.

An explainer system that allows the user to ask the robot "why did you do that?" (Sheh, 2017; Chiyah Garcia et al., 2018; Winfield, 2018; Koeman et al., 2020) and receive a simple natural language explanation could be very helpful in providing this kind of transparency[2]. A higher level of explainability might be the ability to respond to questions such as "Robot: what would you do if I fell down?" or "Robot: what would you do if I forget to take my medicine?" The robot's

[2]Noting that Winograd's SHRDLU Natural Language Processing system demonstrated this capability in 1972 (Winograd, 1972)

responses would allow the user to build a mental model of how the robot will behave in different situations.

3.2.2 Transparency for the Wider Public and Bystanders

Robots and AIs are disruptive technologies likely to have significant societal impact (EC, 2018; Wortham, 2020). It is very important therefore that the whole of society has a basic level of understanding of how these systems work, so we can confidently share work or public spaces with them. That understanding is also needed to inform public debates—and hence policy—on which robots/AIs are acceptable, which are not, and how they should be regulated.

This kind of transparency needs public engagement, for example through panel debates and science cafés, supported by high quality documentaries targeted at distribution by mass media (e.g., YouTube and TV), which present emerging robotics and AI technologies and how they work in an interesting and understandable way. Balanced science journalism—avoiding hype and sensationalism—is also needed.

For this stakeholder group, P7001 defines levels of transparency starting with a requirement that follows a proposed *Turing Red Flag* law: "An autonomous system should be designed so that it is unlikely to be mistaken for anything besides an autonomous system, and should identify itself at the start of any interaction with another agent." (Walsh, 2016). Successive levels build upon this by requiring that systems provide warnings and information about data collected or recorded, since data on bystanders may well be captured.

3.2.3 Transparency for Safety Certifiers

For safety certification of an AIS, transparency is important because it exposes the system's decision making processes for assurance and independent certification.

The type and level of evidence required to satisfy a certification agency or regulator that a system is safe and fit for purpose depends on how critical the system is. An autonomous vehicle autopilot requires a much higher standard of safety certification than, say, a music recommendation AI, since a fault in the latter is unlikely to endanger life. Safe and correct behaviour can be tested by verification, and fitness for purpose tested by validation. Put simply, verification asks "is this system right?" and validation asks "is this the right system?".

At the lowest level of transparency, certification agencies or regulators need to see evidence (i.e., documentation) showing how the designer or manufacturer of an AIS has verified and validated that system. This includes as a minimum a technical specification for the system. Higher levels of transparency may need access to source code and all materials needed (such as test metrics or benchmarks) to reproduce the verification and validation processes. For learning systems, this includes details of the composition and provenance of training data sets.

3.2.4 Transparency for Incident/Accident Investigators

Robots and other AI systems can and do act in unexpected or undesired ways. When they do it is important that we can find out why. Autonomous vehicles provide us with a topical example of why transparency for accident investigation is so important. Discovering why an accident happened through investigation requires details of the situational events leading up to and during the accident and, ideally, details of the internal decision making process in the robot or AI prior to the accident (Winfield et al., 2021).

Established and trusted processes of air accident investigation provide an excellent model of good practice for AIS–processes, which have without doubt contributed to the outstanding safety record of modern commercial air travel (Macrae, 2014). One example of best practice is the aircraft Flight Data Recorder, or "black box"; a functionality we consider essential in autonomous systems (Winfield and Jirotka, 2017).

3.2.5 Transparency for Lawyers and Expert Witnesses

Following an accident, lawyers or other expert witnesses who have been obliged to give evidence in an inquiry or court case or to determine insurance settlements, require transparency to inform their evidence. Both need to draw upon information available to the other stakeholder groups: safety certification agencies, accident investigators and users. They especially need to be able to interpret the findings of accident investigations.

In addition, lawyers and expert witnesses may well draw upon additional information relating to the general quality management processes of the company that designed and/or manufactured the robot or AI system. Does that company, for instance, have ISO 9001 certification for its quality management systems? A higher level of transparency might require that a designer or manufacturer provides evidence that it has undertaken an ethical risk assessment of a robot or AI system using, for instance, BS 8611 *Guide to the ethical design of robots and robotic systems* (BSI, 2016).

3.3 Measurable and Testable Transparency

Standards generally belong to one of two categories: those that offer guidelines or those that set out requirements. P7001 falls into the latter category. P7001 describes a set of normative requirements, which must be met in order for a given system, its documentation, and the processes used to design and test it, to be labeled as "compliant".

A major challenge in drafting P7001 was how to express transparency as something measurable and testable. At first this might seem impossible given that transparency is not a singular physical property of systems, like energy consumption. However, when one considers that the degree to which an end user can understand how a system operates will depend a great deal on the way that user documentation is presented and accessed; or the extent to which an accident investigator can discover the factors that led up to an accident can vary from impossible (to discover) to a very detailed timeline of events, it becomes clear that transparency can be expressed as a set of testable thresholds.

It was on this basis that early in the development of P7001 a scale of transparency from 0 (no transparency) to 5 (the maximum achievable level of transparency) was decided upon, for each of the five stakeholder groups outlined above. At the heart of P7001 are five sets–one set for each stakeholder group–of normative definitions of transparency, for each of the levels 1 to 5.

TABLE 1 | Transparency levels for end users.

Transparency levels (Non-cumulative)	Definition
0	None
1	A user manual must be provided, which sets out how a robot will behave in different circumstances
2	The user manual should be presented as an interactive visualisation or simulation
3	The robot should be equipped with a "why did you just do that?" function which, when activated, provides the user with an explanation of its previous action, either as displayed or spoken text koeman et al. (2020)
4	The robot should be equipped with a "what would you do if … ?" function
5	Not defined

TABLE 2 | Transparency levels for accident investigators.

Transparency levels (Cumulative)	Definition
0	None
1	The robot should be fitted with a recording device to allow capture and playback of the situation around it, leading up to and during an accident
2	The robot should be equipped with a data logging system capable of recording a date and time stamped record of robot sensor inputs, user commands, and actuator outputs
3	As level 2, except that the data logging system should conform to an existing open or industry standard, and additionally log high level decisions
4	As level 3, except that the data logging system should also log the reasons for the robot's high level decisions
5	In addition to level 4, the robot's designers should provide accident investigators with tools to help visualise the robot's data log

Each definition is a requirement, expressed as a qualitative property of the system. In each case the test is simply to determine whether the transparency property required by a given level for a given stakeholder group is demonstrably present or it is not. The choice of five levels was determined as a compromise between a reasonable level of granularity while allowing for discernible differences between successive levels.

Having established a general approach to measurable, testable levels of transparency, the P7001 working group then faced the key question "should those discrete transparency levels be written to reflect the transparency properties found in present day autonomous systems, or should they instead go beyond the present state of the art?" Phrased in another way, should P7001 be written such that most well-designed present day autonomous systems achieve a high level of compliance, or instead as a standard that stretches designers beyond current good practice? Taking a cue from the IEEE P70XX series of standards-in-development, which–in expressing human (ethical) concerns as standards for guidance or compliance–go well beyond the scope of traditional standards, it was determined that P7001 should similarly aim to challenge and extend the practice of transparency.

Given increasingly rapid advances in the capabilities of AI systems, it was also felt that P7001 should consider likely near, and to some extent, medium term advances in the state of the art (for instance, explainable AI for machine learning). However, the working group did not take account of possible long term advances, such as artificial general intelligence or machine consciousness. As and when it becomes necessary, the standard can be updated to meet with advances in the state of the art.

The general principle was, therefore, established that transparency levels should start (on level 1) with transparency measures that we might generally expect to find in well-designed present day systems, or that could be easily provided. Levels 2 and up should be successively more demanding, going beyond what one would presently expect in most well designed systems, and in some cases require solutions that are—at the time of writing—the subject of ongoing research.

The approach outlined above is illustrated below in **Tables 1, 2** for the stakeholder groups "end users" and "accident investigators", respectively, (The illustrations in **Tables 1, 2** are abbreviated versions of the transparency definitions for end users and accident investigators in P7001 for robots only, rather than autonomous systems in general).

In **Tables 1, 2**, we see that each level n describes a successively greater degree of transparency than the previous level $n - 1$. For most stakeholder groups each level builds upon previous levels, so if a system meets level n, then it also meets levels $n - 1$, etc. Thus transparency levels are cumulative for accident investigators in **Table 2**, but not in **Table 1** for end users so, for instance, a designer may choose to provide an interactive visualisation of level 2 instead of the user manual of level 1 (or they may choose to provide both).

Level 1 in **Table 1**–a user manual–will typically be present for all present day robots. Similarly, the recording device required by level 1 in **Table 2** will be easy to provide, if not already present.

Consider now levels 2–4 in **Table 1** for end users. Level 2–an interactive visualisation–is more demanding than level 1 but perfectly feasible with current simulation and visualisation technology. Levels 3 and 4 do, however, go beyond the current state of the art in robotics, but methods for implementing this kind of explainability in robots are emerging (Theodorou et al.,

2017; Winfield, 2018; Rosenfeld and Richardson, 2019; Koeman et al., 2020; Dennis and Oren, 2021).

Consider also levels 2–5 in **Table 2**, for accident investigators. Level 2–a bespoke data logging system, while not currently present in many robots, would not be technically challenging to implement. Winfield and Jirotka (2017) provide a general outline of what is required. Level 3–a data logging system conforming to an existing standard is more challenging since, for robots in general, such standards do not yet exist[3]. For autonomous vehicles, however, a closed standard for automotive Event Data Recorders does exist in (SAE J1698_2017), with another in development (IEEE P1616). Level 4 goes further in requiring the data logger to record the reasons for high level decisions; something that would require access to internal processes of the robot's control system, which would not normally be accessible via, for instance, the robot's API.

3.4 Compliance

A system would be compliant with P7001 if it meets at least level 1 transparency for at least one stakeholder group. Note, however, that a simple statement that "system x is compliant with P7001" would be misleading. The correct way to describe P7001 compliance is through the multi-element description of the STA, outlined in **Section 4**.

Consider a system that is assessed as providing level 1 transparency for one stakeholder group only: the absolute minimum level of compliance. In some bounded and benign use cases, such a level might still be regarded as adequate. However, what constitutes sufficient or appropriate levels of transparency will vary a great deal from one system and its intended use to another.

It is important also to recognise that stakeholder groups and their transparency requirements are independent of each other, thus there is no expectation that if a system meets a particular level in one stakeholder group, it should also meet the same level in other groups.

In practice, the decision over which transparency level is needed in each stakeholder group should be guided by an ethical risk assessment. BS 8611 sets out a method for ethical risk assessment of robots or robotic systems (BSI, 2016), and an example of ethical risk assessment for a child's toy robot can be found in Winfield and Winkle (2020). Example scenarios will be outlined in **Section 4** below.

It is clear that 1) compliance with P7001 will vary a great deal between systems, and between stakeholder groups for a particular system, and 2) whether the level of compliance for a given system is adequate or not will depend on the possible risk of (ethical) harms should the system fail or be compromised. So we might expect that, in general, safety-critical autonomous systems would require higher levels of transparency than non-critical systems. One thing we can be reasonably sure of is that a system that fails to score even level 1 for any stakeholder group is unlikely to have adequate transparency.

[3]Although at least one open standard for robots is known to be in draft.

4 P7001 PROCESSES

P7001 is a process standard; it does not specify *how* the transparency measures defined in it must be implemented, only the kind of transparency each measure affords and how to determine whether it is present or not. Some transparency measures will require designers to include well understood features; transparency for accident investigators, for instance, requires that systems incorporate event data recorders (EDRs)–the functional equivalent of aircraft flight data recorders–without which it would be impossible to investigate accidents. The draft standard does not, however, specify required functionality of the EDR, except at a very generic level.

As mentioned above, P7001 has two primary functions. The first is as a tool for assessing the transparency of existing systems, called a System Transparency Assessment (STA), and the second as a guide for creating a transparency specification for a given system prior to, or during, its design: this is a System Transparency Specification (STS). Each of these will now be illustrated with a case study.

4.1 System Transparency Assessment for a Robot Toy

In Winfield and Winkle (2020), we describe an ethical risk assessment for a fictional intelligent robot teddy bear we called RoboTED. Let us now assess the transparency of the same robot. In summary, RoboTED is an Internet (WiFi) connected device with cloud-based speech recognition and conversational AI (chatbot) with local speech synthesis; RoboTED's eyes are functional cameras allowing the robot to recognise faces; RoboTED has touch sensors, and motorised arms and legs to provide it with limited baby-like movement and locomotion—not walking but shuffling and crawling.

Our ethical risk assessment (ERA) exposed two physical (safety) hazards including tripping over the robot and batteries overheating. Psychological hazards include addiction to the robot by the child, deception (the child coming to believe the robot cares for them), over-trusting of the robot by the child, and over-trusting of the robot by the child's parents. Privacy and security hazards include weak security (allowing hackers to gain access to the robot), weak privacy of personal data especially images and voice clips, and no event data logging making any investigation of accidents all but impossible[4].

The ERA leads to a number of recommendations for design changes. One of those is particularly relevant to the present paper: the inclusion of an event data recorder, so our outline transparency assessment, given below in **Table 3**, will assume this change has been made.

[4]The ERA also considers environmental risks, including sustainability, repairability and recyclability, but these have no bearing on transparency and do not need to be considered here.

TABLE 3 | Outline system transparency assessment (STA) for RoboTED.

Stakeholder Group	Transparency level(s)	Evidenced by
[i] users	1, 2	A user manual is provided for parents. As well as detailing how parents can show children how best to use RoboTED, the manual explains the risks (addiction, deception and over-trusting) and how to minimise these. The manual also shows how to guard against hacking and check personal data has been deleted (level 1). An interactive online visual guide is also provided, for both parents and children (level 2)
[ii] general public	1	P7001 level 1 requires that a robot identifies itself as an autonomous system, following Walsh (2016). When powered up, or on waking from sleep mode, RoboTED announces itself as a robot
[iii] certification agencies	2	RoboTED has been certified as safe against standard EU EN 62115 (2020) *Safety of Electric Toys*, and descriptions of the system and how it has been validated are available for safety certifiers. This meets P7001 level 2
[iv] accident Investigators	2	The robot is equipped with a data logging system as outlined in **Table 2**
[v] lawyers and expert witnesses	2	P7001 level 2 requires that a system has been subjected to an ethical risk assessment, which can be made available to lawyers or expert witnesses. This is the case for RoboTED

TABLE 4 | Outline system transparency specification (STS) for nextVac.

Stakeholder Group	Transparency level(s) Required	Rationale
[i] users	1, 2 (see **Table 1**)	A comprehensive user manual is required, covering both use and maintenance. The manual should be written in compliance with standard IEC/IEEE std 82,079 *Preparation of information for use*, as recommended by P7001 (level 1). An interactive online visual guide is also required, for both operators of the cleaning robot and facilities managers (level 2). Levels 3 and 4 are not required as the robot is not expected to need a complex human robot interface. The robot will only require a limited number of behaviours and these will be indicated by warning lights and sounds, see group [ii] below
[ii] general public	1, 2	The robot's design will ensure that its machine nature is apparent; lights and sounds will provide simple audio-visual indications of what the robot is doing at any time (level 1). The robot will provide physical cues showing the location of sensors, and publicly available information will explain what data is stored and why (see [iv] accident Investigators in this table), and that this data will not include any personal data (level 2)
[iii] certification agencies	3	The robot will be certified as safe against relevant standards, such as ISO 10218 (2011) (noting that ISO 10218 is a generic standard for the safety of industrial robots). Descriptions of the system and how it has been validated will be made available to safety certifiers (level 2). In addition, a high level model (simulation) of the robot will be developed and made available (level 3)
[iv] accident Investigators	3 (see **Table 2**)	The robot will be equipped with a data logging system, which records high level decisions (as outlined in **Table 2**). Noting that the data logging system will not record any personal data. Levels 4 and 5 are not considered essential, as the cleaning robot will only require a limited number of behaviours, nor will it learn
[v] lawyers and expert witnesses	4	nextVac already has certification of quality management (QM) to standard ISO 9001 (level 1). Ethical risk assessment (ERA) against BS8611 will be undertaken (level 2). nextVac has in place processes of ethical governance (level 3). nextVac also maintains complete audit trails for QM, ERA and ethical governance processes (level 4)

4.2 System Transparency Specification for a Vacuum Cleaner Robot

Consider now a fictional company that designs and manufactures robot vacuum cleaners for domestic use. Let us call this company nextVac. Let us assume that nextVac is well established in the domestic market and has a reputation both for the quality of its products and responsible approach to design and manufacture. nextVac now wishes to develop a new line of robot vacuum cleaners for use in healthcare settings: including hospitals, clinics and elder care homes.

nextVac begins the design process with a scoping study in which they visit healthcare facilities and discuss cleaning needs with healthcare staff, facilities managers and cleaning contractors. Mindful of the additional safety, operational and regulatory requirements of the healthcare sector (over and above their domestic market), nextVac decides to capture the transparency needs of the new product–while also reflecting the findings of the scoping study–in a System Transparency Specification (STS), guided by IEEE P7001. Their intention is to follow the STS with an initial product design specification. In turn this specification will be subjected to an Ethical Risk Assessment (ERA), guided by BS8611. Depending on the findings of the ERA, the company will iterate this process until a product specification emerges that is technically feasible, tailored to customer needs, and addresses both ethical risks and transparency needs.

Capturing the full process of drafting an STS for this scenario is beyond the scope of this paper, so instead we outline the key requirement in **Table 4**.

The outline STS for nextVac's proposed new vacuum cleaning robot for healthcare, leads to a number of clear technical design requirements, especially for stakeholder groups [i], [ii], and [iv], alongside process requirements for groups [iii] and [v]. The STS will thus feed into and form part of the product design specification.

Note also that the outline STS in **Table 4** illustrates–for groups [i] and [iv]–the value of also asking the question, and therefore

seeking explicit justification, for why certain higher levels of transparency are *not* required.

5 CONCLUDING DISCUSSION

In this concluding section, we first discuss security, privacy and transparency before then outlining and discussing the challenges faced when drafting P7001, and its limitations.

5.1 Security, Privacy and Transparency

Security and privacy practices are generally embedded within the fabric of autonomous systems. Security standards, especially for regulated industries such as transportation, utilities and finance, receive particular attention by system architects and auditors, but transparency within these mature frameworks tends to be addressed indirectly. To adequately consider transparency for security and privacy, STA and STS statements must be tied closely to prevailing information security standards.

The STA equivalent in security standards such as ISO 27001 and NIST 800-53 tends to be framed as governance or assurance tasks (NIST, 2020). These tasks, both automated and manual, verify the presence of a security control. For instance, in P7001 example scenario B.6 (Medical Decision Support), an assurance task verifies that patient information is encrypted in transit and in rest and is not exposed beyond a circumscribed list of providers. An autonomous system whose security and privacy protections are transparent will disclose the methods being used to protect sensitive information. In some cases, users can perform assurance tasks themselves.

Autonomous system architects can fashion STAs following recommendations of the NIST Big Data Reference Architecture (SP 1500-r2) wherein higher security and privacy safety levels provide additional disclosures–i.e., transparency–via multiple techniques including a System Communicator. NIST 1500-r2 addresses three voluntary levels of system transparency, each of which can be integrated into an STS (Chang et al., 2019, Sect. 2.4.8). Big data plays an increasingly prominent role in autonomous systems and presents particularly challenging security and privacy risks.

Newer autonomous systems constructed using DevOps principles offer additional opportunities to embed STS requirements. IEEE 2675–2021 cites benefits for DevOps communities: "Transparency prioritizes ease of visibility, availability, reachability, and accessibility of information and actions between entities, people, or systems" (IEEE, 2021, Sect. 5.3.3).

Some facets of security and privacy are global and human, affecting well-being in ways that require different and novel metrics. IEEE 7010–2019 directly cites the relevance of P7001 and further recommends that autonomous data collection plans address ". . .issues related to collection and use of data, such as ethics, *transparency*, data privacy, data governance, security, protection of data, nudging, coercion, algorithmic bias, asymmetry, and redundancy . . . " (IEEE, 2019a, Sect. 5.3.1, Table 6, italics added).

5.2 Challenges and Limitations

P7001 is, to the best of our knowledge, the first attempt to write a standard on transparency; this alone would make development of the standard challenging. In particular:

(1) The comparative youth of the field makes it difficult to assess what it is practical to require now in terms of transparency, let alone what might be practical within the lifetime of the standard. This is acute in the case of Deep Neural Nets (DNNs), which many people wish to use but also present a challenge to explainability (at least), if not necessarily to transparency in general.
(2) The heterogeneous nature of transparency is a problem. Is the simple provision of information (e.g., a log) sufficient, or must the information be in a contextualised form (e.g., an explanation)? Across and within the stakeholder groups, there was discussion over whether contextualisation was desirable since it necessarily creates a system-generated interpretation of what is happening, which could introduce biases or errors in reporting. Is something transparent if we can inspect all parts of it but not understand the emergent behaviour, as may be the case for a DNN?
(3) What is the best medium for the presentation of such information? There is a tendency to assume it should be written or verbal but diagrams and other visual mechanisms can also be important. A range of possible outputs increases accessibility, and some outputs may be better suited to certain situations, for example, where privacy is a factor, or an incident where all people nearby must be immediately notified through an alarm.
(4) Within P7001's various stakeholder groups, it was sometimes difficult to foresee what transparency might be wanted for, and without knowing the purpose of transparency it was hard to determine what should be required and how compliance might be measured.
(5) When might transparency lead to over-confidence? In a recent paper, Kaur et al. (2020) showed that the provision of explainability mechanisms led to over-confidence in a model. This may also contribute to automation bias, a tendency to place unwarranted trust in the accuracy and infallibility of automated systems.
(6) Transparency exists in tension with a number of other ethical principles, most notably security (where lack of transparency is often a first line of defence) and privacy (for instance, in our RoboTED example, some potential explanations might reveal personal information about the child who owned the toy). This highlights the need for determinations about appropriate levels of transparency to be informed by both ethical risk assessment and the practices outlined in **Section 5.1**.

The challenges mentioned above were further compounded by the demands of writing normative definitions of transparency that are at the same time sufficiently generic to apply to all autonomous systems, while also specific enough to be implemented and expressed with enough precision to allow

the question "is this transparency measure present in this system or not" to be answered. P7001 has been drafted as an "umbrella" standard, and an indicator of its success would not only be its application to real world autonomous systems, including both robots and AIs, but also the subsequent development of domain specific variants. Each branching standard, 7001.1, 7001.2, etc., would inherit the generic definitions of 7001 but elaborate these more precisely as, for instance, standards on transparency in autonomous vehicles, transparency of AIS in healthcare, and so on.

To what extent did the difficulties articulated here lead to limitations in P7001? One clear limitation is that P7001 does not offer detailed advice on how to implement the various kinds of transparency described in it. However, we would argue that a strength of P7001 is the clear articulation of the two processes of systems transparency assessment (STA) and specification (STS). Another related limitation is that several definitions of higher levels of transparency require techniques that have not yet been developed–to the extent that they can be readily applied. One example is the requirement for systems to provide non-expert users with answers to "why" and "what if" questions, in levels 3 and 4 of transparency for users. Another example would be higher levels of verification and validation for systems that learn, within the stakeholder group of certification agencies, given that verification of autonomous systems is challenging–especially for machine learning systems–and remains the subject of current research.

These limitations may suggest that there would be no value in assessment of the transparency of autonomous systems that can learn (either offline or online). However, we would argue that there is value, even–and especially–if assessment exposes transparency gaps in machine learning systems. Just as transparency is vitally important, so is honest appraisal of the levels of transparency of a given system. P7001 will, for the first time, allow us to be rigorously transparent about transparency.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

AW was lead author. AW, SB, LD, TE, HH, NJ, RM, JO, FR, AT, MU, RW, and EW all contributed ideas and material for the paper; final editing was led by AW, supported by SB, LD, HH, NJ, RM, JO, FR, AT, MU, RW, and EW.

FUNDING

The work described in this paper was in part supported by EPSRC grants EP/L024845/1 (Verifiable Autonomy) (LD and AW), EP/V026801/1 and EP/V026682/1 (UKRI Trustworthy Autonomous Systems: Verifiability Node (LD) and Node on Trust (HH)), EP/S005099/1 (RoboTIPS: Developing Responsible Robots for the Digital Economy) (AW) and ESRC grant ES/S001832/1 (Driverless Futures) (AW). AT is funded by the Knut and Alice Wallenberg Foundation, grant agreement 2020.0221, and by the European Union's Horizon 2020 research and innovation program under grant agreement No 825619.

ACKNOWLEDGMENTS

The authors are deeply grateful to fellow volunteer members, both past and present, of IEEE standards Working Group P7001, for their support and advice. The P7001 working group is sponsored by IEEE Standards Committee VT/ITS–Intelligent Transportation Systems. We are also grateful to the reviewers for their insightful comments and suggestions.

REFERENCES

Adadi, A., and Berrada, M. (2018). Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6, 52138–52160. doi:10.1109/ACCESS.2018.2870052

Alonso, V., and de la Puente, P. (2018). System Transparency in Shared Autonomy: A Mini Review. *Front. Neurorobot.* 12, 83. doi:10.3389/fnbot.2018.00083

Alzetta, F., Giorgini, P., Najjar, A., Schumacher, M. I., and Calvaresi, D. (2020). "In-time Explainability in Multi-Agent Systems: Challenges, Opportunities, and Roadmap," in *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*. Editors D. Calvaresi, A. Najjar, M. Winikoff, and K. Främling (Cham: Springer International Publishing), 39–53. doi:10.1007/978-3-030-51924-7_3

Angelov, P., and Soares, E. (2020). Towards Explainable Deep Neural Networks (xDNN). *Neural Networks* 130, 185–194. doi:10.1016/j.neunet.2020.07.010

Anjomshoae, S., Najjar, A., Calvaresi, D., and Främling, K. (2019). Explainable Agents and Robots: Results from a Systematic Literature Review. Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, '19. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems AAMAS, 1078–1088.

Ball, C. (2009). What Is Transparency? *Public Integrity* 11, 293–308. doi:10.2753/PIN1099-9922110400

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible Ai. *Inf. Fusion* 58, 82–115. doi:10.1016/j.inffus.2019.12.012

Baum, K., Hermanns, H., and Speith, T. (2018). "From Machine Ethics to Machine Explainability and Back," in International Symposium on Artificial Intelligence and Mathematics, Fort Lauderdale, FL, 1–8. Available at: https://www.powver.org/publications/TechRepRep/ERC-POWVER-TechRep-2018-02.pdf.

Bertino, E., Kundu, A., and Sura, Z. (2019). Data Transparency with Blockchain and AI Ethics. *J. Data Inf. Qual.* 11, 1–8. doi:10.1145/3312750

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., et al. (2017). Principles of Robotics: Regulating Robots in the Real World. *Connect. Sci.* 29, 124–129. doi:10.1080/09540091.2016.1271400

Booth, S., Zhou, Y., Shah, A., and Shah, J. (2021). "Bayes-TrEx: a Bayesian Sampling Approach to Model Transparency by Example," Proceedings of the AAAI Conference on Artificial Intelligence, (Palo Alto, CA: AAAI Press), 11423–11432.

Bryson, J. J., and Theodorou, A. (2019). "How Society Can Maintain Human-Centric Artificial Intelligence," in *Human-Centered Digitalization and Services*.

Editors M. Toivonen-Noro, E. Saari, H. Melkas, and M. Hasu (Singapore: Springer), 305–323. doi:10.1007/978-981-13-7725-9-16

BSI (2016). *BS8611:2016 Robots and Robotic Devices, Guide to the Ethical Design and Application of Robots and Robotic Systems*. British Standards Institute.

Burton, S., Habli, I., Lawton, T., McDermid, J., Morgan, P., and Porter, Z. (2020). Mind the Gaps: Assuring the Safety of Autonomous Systems from an Engineering, Ethical, and Legal Perspective. *Artif. Intelligence* 279, 103201. doi:10.1016/j.artint.2019.103201

Cantucci, F., and Falcone, R. (2020). "Towards Trustworthiness and Transparency in Social Human-Robot Interaction," In 2020 IEEE International Conference on Human-Machine Systems (ICHMS), (Rome: IEEE). 1–6. doi:10.1109/ICHMS49158.2020.9209397

Chakraborti, T., Sreedharan, S., and Kambhampati, S. (2020). "The Emerging Landscape of Explainable Automated Planning & Decision Making," in Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20. Editor C. Bessiere (International Joint Conferences on Artificial Intelligence Organization), 4803–4811. doi:10.24963/ijcai.2020/669.Survey.track

Chang, W., Roy, A., and Underwood, M. (2019). "NIST Big Data Interoperability Framework: Volume 4, Big Data Security and Privacy [Version 3], Special Publication (NIST SP)," in *Tech. Rep* (Gaithersburg: National Institute of Standards and Technology).

Chatila, R., Firth-Butterflied, K., Havens, J. C., and Karachalios, K. (2017). The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems [standards]. *IEEE Robot. Automat. Mag.* 24, 110. doi:10.1109/MRA.2017.2670225

Chiyah Garcia, F. J., Robb, D. A., Liu, X., Laskov, A., Patron, P., and Hastie, H. (2018). Explainable Autonomy: A Study of Explanation Styles for Building clear Mental Models. Proceedings of the 11th International Conference on Natural Language Generation. Netherlands: Tilburg UniversityAssociation for Computational Linguistics, 99–108. doi:10.18653/v1/W18-6511

Confalonieri, R., Coba, L., Wagner, B., and Besold, T. R. (2021). A Historical Perspective of Explainable Artificial Intelligence. *Wires Data Mining Knowl Discov.* 11, e1391. doi:10.1002/widm.1391

De Graaf, M. M. A., and Malle, B. F. (2017). "How People Explain Action (And Autonomous Intelligent Systems Should Too)," In AAAI Fall Symposium Series 2017, Palo Alto, CA: AAAI Press.

Dennis, L. A., and Fisher, M. (2020). Verifiable Self-Aware Agent-Based Autonomous Systems. *Proc. IEEE* 108, 1011–1026. doi:10.1109/JPROC.2020.2991262

Dennis, L., and Oren, N. (2021). "Explaining BDI Agent Behaviour through Dialogue," in Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021). Editors U. Endriss, A. Nowe, F. Dignum, and A. Lomuscio (Richland SC: International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS)), 429–437.

Doshi-Velez, F., and Kim, B. (2017). *Towards a Rigorous Science of Interpretable Machine Learning*. arXiv. Available at: https://arxiv.org/abs/1702.08608.

EC (2018). *European Commission's High-Level Expert Group (HLEG) Guidelines for Trustworthy AI*. Brussels: European Commission. Tech. rep.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. In 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), Turin, (IEEE). 80–89. doi:10.1109/DSAA.2018.00018

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., and Yang, G.-Z. (2019). XAI-explainable Artificial Intelligence. *Sci. Robot.* 4, eaay7120. doi:10.1126/scirobotics.aay7120

Holzinger, A. (2018). "From Machine Learning to Explainable AI," In 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA), Košice, Slovakia, (IEEE). 55–66. doi:10.1109/DISA.2018.8490530

IEEE (2021). *IEEE 2675-2021 - IEEE Standard For DevOps: Building Reliable And Secure Systems Including Application Build, Package, and Deployment*. Piscataway, NJ: IEEE Standards Association. Tech. rep.

IEEE (2019a). *IEEE 7010-2020, IEEE Recommended Practice For Assessing The Impact Of Autonomous And Intelligent Systems On Human Well-Being*. Piscataway, NJ: IEEE Standards Association. Tech. Rep.

IEEE (2020). "IEEE Draft Standard for Transparency of Autonomous Systems," in IEEE P7001/D1, June 2020. (Piscataway, NJ: IEEE), 1–76. Tech. rep.

IEEE (2019b). *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems*. First Edition. Piscataway, NJ: IEEE Standards Association. Tech. rep.

Jobin, A., Ienca, M., and Vayena, E. (2019). The Global Landscape of AI Ethics Guidelines. *Nat. Mach Intell.* 1, 389–399. doi:10.1038/s42256-019-0088-2

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., and Wortman Vaughan, J. (2020). Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery), 1–14. doi:10.1145/3313831.3376219CHI '20

Koeman, V. J., Dennis, L. A., Webster, M., Fisher, M., and Hindriks, K. (2020). "The "Why Did You Do that?" Button: Answering Why-Questions for End Users of Robotic Systems," in *Engineering Multi-Agent Systems*. Editors L. A. Dennis, R. H. Bordini, and Y. Lespérance (Cham: Springer International Publishing)), 152–172. doi:10.1007/978-3-030-51417-4_8

Koene, A., Clifton, C., Hatada, Y., Webb, H., and Richardson, R. (2019). *A Governance Framework for Algorithmic Accountability and Transparency*. Brussels: European Parliamentary Research Service (EPRS). Tech. Rep.

Lakhmani, S., Abich, J., Barber, D., and Chen, J. (2016). "A Proposed Approach for Determining the Influence of Multimodal Robot-Of-Human Transparency Information on Human-Agent Teams," in *Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience*. Editors D. D. Schmorrow and C. M. Fidopiastis (Cham: Springer International Publishing)), 296–307. doi:10.1007/978-3-319-39952-2_29

Langley, P., Meadows, B., Sridharan, M., and Choi, D. (2017). "Explainable agency for Intelligent Autonomous Systems," In AAAI International Conference on Artificial Intelligence, (Palo Alto, CA: AAAI Press). 17, 4762–4763.

Larsson, S., and Heintz, F. (2020). Transparency and the Future of Artificial Intelligence, *Transparency in Artificial Intelligence*, 9, 1–16. doi:10.1287/lytx.2020.04.01

Lipton, Z. C. (2018). The Mythos of Model Interpretability. *Queue* 16, 31–57. doi:10.1145/3236386.3241340

Macrae, C. (2014). *Close Calls: Managing Risk and Resilience in Airline Flight Safety*. London: Palgrave).

Nesset, B., Robb, D. A., Lopes, J., and Hastie, H. (2021). "Transparency in HRI: Trust and Decision Making in the Face of Robot Errors," in Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. New York, NY, USA: Association for Computing Machinery), HRI '21 Companion, 313–317. doi:10.1145/3434074.3447183

NIST (2020). *Joint Task Force Transformation Initiative Interagency Working Group (2020) Security And Privacy Controls For Federal Information Systems And Organizations, NIST Special Publication (SP) 800-53, Rev. 5*. Gaithersburg, MD: National Institute of Standards and Technology. Tech. Rep.

OECD (2019). *Recommendation of the Council on Artificial Intelligence*. Paris: Organisation for Economic Co-operation and Development, Tech. Rep.

Olhede, S., and Rodrigues, R. (2017). Fairness and Transparency in the Age of the Algorithm. *Significance* 14, 8–9. doi:10.1111/j.1740-9713.2017.01012.x

Olszewska, J. (2019). "Designing Transparent and Autonomous Intelligent Vision Systems," in Proceedings of the 11th International Conference on Agents and Artificial Intelligence, Prague. Scitepress Digital Library, 2. 850–856. doi:10.5220/0007585208500856

Páez, A. (2019). The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds & Machines* 29, 441–459. doi:10.1007/s11023-019-09502-w

Rosenfeld, A., and Richardson, A. (2019). Explainability in Human-Agent Systems. *Auton. Agent Multi-agent Syst.* 33, 673–705. doi:10.1007/s10458-019-09408-y

Rotsidis, A., Theodorou, A., Bryson, J. J., and Wortham, R. H. (2019). "Improving Robot Transparency: An Investigation with Mobile Augmented Reality," in 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). New Delhi, India: IEEE, 1–8. doi:10.1109/RO-MAN46459.2019.8956390

Sheh, R. (2017). "Why Did You Just Do that? Explainable Intelligent Robots," in *AAAI Workshop On Human-Aware Artificial Intelligence*. (Palo Alto CA: AAAI Press).

Spagnolli, A., Frank, L., Haselager, P., and Kirsh, D. (2016). "Transparency as an Ethical Safeguard," in *International Workshop on Symbiotic Interaction* Lecture Notes In Computer Science (Cham: Springer), 10727, 1–6. doi:10.1007/978-3-319-91593-7-1

TaheriNejad, N., Herkersdorf, A., and Jantsch, A. (2020). Autonomous Systems, Trust and Guarantees. *IEEE Des. Test.*, 1. doi:10.1109/MDAT.2020.3024145

Theodorou, A., and Dignum, V. (2020). Towards Ethical and Socio-Legal Governance in AI. *Nat. Mach Intell.* 2, 10–12. doi:10.1038/s42256-019-0136-y

Theodorou, A., Wortham, R. H., and Bryson, J. J. (2017). Designing and Implementing Transparency for Real Time Inspection of Autonomous Robots. *Connect. Sci.* 29, 230–241. doi:10.1080/09540091.2017.1310182

Tulli, S., Correia, F., Mascarenhas, S., Gomes, S., Melo, F., and A., P. (2019). "Effects of Agents' Transparency on Teamwork," in *Explainable, Transparent Autonomous Agents and Multi-Agent Systems. EXTRAAMAS 2019 of Lecture Notes In Computer Science*. Editors D. Calvaresi, A. Najjar, M. Schumacher, and K. Främling (Cham: Springer), 11763. doi:10.1007/978-3-030-30391-4-2

Vorm, E., and Miller, A. (2020). "Modeling User Information Needs to Enable Successful Human-Machine Teams: Designing Transparency for Autonomous Systems," in *Augmented Cognition. Human Cognition and Behavior. HCII 2020 of* Lecture Notes In Computer Science. Editors D. Schmorrow and C. Fidopiastis (Cham: Springer), 12197. doi:10.1007/978-3-030-50439-7-31

Walsh, T. (2016). Turing's Red Flag. *Commun. ACM* 59, 34–37. doi:10.1145/2838729

Weller, A. (2019). "Transparency: Motivations and Challenges," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning of* Lecture Notes In Computer Science. Editors W. Samek, G. Montavon, A. Vedaldi, L. Hansen, and K. Muller (Cham: Springer), 11700. doi:10.1007/978-3-030-28954-6-2

Winfield, A. (2019). Ethical Standards in Robotics and AI. *Nat. Electron.* 2, 46–48. doi:10.1038/s41928-019-0213-6

Winfield, A. F. T. (2018). Experiments in Artificial Theory of Mind: From Safety to story-telling. *Front. Robot. AI* 5, 75. doi:10.3389/frobt.2018.00075

Winfield, A. F. T., and Jirotka, M. (2017). "The Case for an Ethical Black Box," in *Towards Autonomous Robotic Systems (TAROS 2017) Lecture Notes in Computer Science*. Editors Y. Gao, S. Fallah, Y. Jin, and C. Lekakou (Cham: Springer), 10454, 262–273. doi:10.1007/978-3-319-64107-2_21

Winfield, A. F. T., Winkle, K., Webb, H., Lyngs, U., Jirotka, M., and Macrae, C. (2021). "Robot Accident Investigation: A Case Study in Responsible Robotics," in *Software Engineering for Robotics*. Editors A. Cavalcanti, B. Dongol, R. Hierons, J. Timmis, and J. Woodcock (Cham): Springer). doi:10.1007/978-3-030-66494-7_6

Winfield, A. F., and Winkle, K. (2020). RoboTED: a Case Study in Ethical Risk Assessment. in 5th International Conference On Robot Ethics And Standards (ICRES 2020), Taipei, arXiv. Available at: https://arxiv.org/abs/2007.15864v2.

Winograd, T. (1972). Understanding Natural Language. *Cogn. Psychol.* 3, 1–191. doi:10.1016/0010-0285(72)90002-3

Wortham, R. (2020). *Transparency for Robots and Autonomous Systems*. London: IET Press.

Wright, J. L., Chen, J. Y. C., and Lakhmani, S. G. (2020). Agent Transparency and Reliability in Human-Robot Interaction: The Influence on User Confidence and Perceived Reliability. *IEEE Trans. Human-mach. Syst.* 50, 254–263. doi:10.1109/THMS.2019.2925717

Zhang, Y., and Chen, X. (2020). Explainable Recommendation: A Survey and New Perspectives. *FNT Inf. Retrieval* 14, 1–101. doi:10.1561/1500000066

Conflict of Interest: Author TE was employed by company NEC Corporation. Author RM was employed by company Fourth Insight Ltd. Author MU was employed by company Synchrony Financial. Author EW was employed by company Nell Watson Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Robotics and AI

ORIGINAL RESEARCH
published: 19 August 2021
doi: 10.3389/frobt.2021.643160

Ethical Design of a Robot Platform for Disabled Employees: Some Practical Methodological Considerations

Tommaso Colombino[1]*, *Danilo Gallo*[1], *Shreepriya Shreepriya*[1], *Yesook Im*[2] and *Seijin Cha*[2]

[1]Naver Labs Europe, Grenoble, France, [2]Naver Labs, Seoul, South Korea

This paper explains the process of developing a scenario involving the use of a robotic platform to enhance the work experience of disabled employees. We outline the challenges involved in revealing the potential unintended consequences of introducing elements of Artificial Intelligence, automation, and robotics into a socially and ethically complex and potentially fragile scenario, and the practical challenges involved in giving a voice to vulnerable users throughout the design process. While an ideal case scenario would involve the disabled employees as much as possible directly in the design process, this can, realistically, be a challenge. In this paper, we detail a methodological and analytic approach that is centered around ethnography and design fictions. It is designed to provide a deeper understanding of all the stakeholders involved in the scenario while encouraging ethical reflection. Based on our findings, we argue that, while it is relatively easy to adopt an *a priori* ethical stance through notions such as inclusivity and accessibility, there are risks involved in making such *a priori* prescriptions with respect to the perspectives of different stakeholders in an applied research project. More specifically, we highlight the importance of understanding the broad organizational and bureaucratic characteristics of a business or workplace when devising HRI scenarios and tasks, and of considering elements such as business models, operating philosophy, and organizational hierarchies in the design process.

Keywords: robotic platforms, workplace studies, ethnography, user experience design, assistive technology

OPEN ACCESS

Edited by:
Martim Brandão, King's College London, United Kingdom

Reviewed by:
Jainendra Shukla, Indraprastha Institute of Information Technology Delhi, India
Thorsten Kolling, University of Siegen, Germany

***Correspondence:**
Tommaso Colombino tommaso.colombino@naverlabs.com

Specialty section:
This article was submitted to Ethics in Robotics and Artificial Intelligence, a section of the journal Frontiers in Robotics and AI

Received: *17 December 2020*
Accepted: *12 July 2021*
Published: *19 August 2021*

Citation:
Colombino T, Gallo D, Shreepriya S, Im Y and Cha S (2021) Ethical Design of a Robot Platform for Disabled Employees: Some Practical Methodological Considerations. Front. Robot. AI 8:643160. doi: 10.3389/frobt.2021.643160

INTRODUCTION

Extensive research has been conducted on designing robots that successfully support people with disabilities. A great deal of research is dedicated to the use of robots as therapeutic aids in controlled experimental or clinical environments. These studies leverage the fact that robots lend themselves well to repetitive tasks and can be used in training scenarios to teach specific skills (Javed et al., 2018).

HRI scenarios used to test therapy protocols can also be used to investigate and test cognitive, social and intellectual abilities and characteristics of specific disabilities (Hautop Lund, 2009). As in therapeutic scenarios, researchers leverage the suitability of robots for repetitive tasks, and their potentially non-threatening nature. Anthropomorphic robots or robots with facial features are used as proxies for humans to practice emotion recognition skills, under the assumption that they are "easier" to interact with and may boost engagement.

Outside of clinical scenarios, HRI offers the potential for robots to be used in the care and assistance of people with disabilities. These are people with physical disabilities and the elderly as well as people with developmental disabilities (DDs). Moreover, it is necessary to include the needs of

many actors (Yamazaki et al., 2016), including those of caregivers. Robots may be used to assist caregivers or directly replace them, which may be desirable because the elderly and disabled may value their independence (Shiomi et al., 2014).

While experimental and clinical scenarios are of interest to the HCI community, we consider the workplace an equally important setting. Work integration is one of the biggest challenges faced by people with DDs (Dibia et al., 2015). While many countries have legislation mandating companies to employ a quota of disabled workers, the categories are broad and people with DDs may find it difficult to find gainful and interesting employment opportunities (Gaudion, 2016; The National Autistic Society, 2016). There is a lack of support, both when finding employment and during employment itself. It is also sometimes difficult for prospective employers to evaluate the true skill level and potential of employees with disabilities and to provide an environment that is adapted to their needs.

Previous research has studied the use of robotic support to enable employees with DDs to perform specific tasks (Baxter et al., 2018; Kidal et al., 2018; Stöhr et al., 2018; Kidal et al., 2019). However, to the best of our knowledge there is less focus on understanding the impact of organizational roles and characteristics on the definition of the robotic support. We believe that this is a critical aspect to consider in the design process in order to propose solutions that can realistically be implemented.

The research at the heart of this paper was conducted in collaboration with a Korean business that employs people with cognitive and developmental disabilities across a variety of business-to-business service operations. The goal of the study was to contribute to the development of scenarios involving the use of a robotic platform to enhance the work experience of the disabled employees. We used design fictions to elicit future scenarios and better understand the impact of using robotic technology from different stakeholders' perspectives. While we were not looking to promote overly advanced visions of what robots might achieve, these futuristic visions helped us understand the expectations of our stakeholders. The aim of our project was to manage these expectations and work within the current state of the art to deliver a design proposal that could be realistically integrated into a workplace within 1 or 2 years.

We were quite conscious of ethical concerns and risks related to forcing technological innovation onto a potentially vulnerable population (disabled employees). Furthermore, as representatives of a research organization involved in AI and the design of robotic platforms and services, we knew the project would be characterized by a strong technology push. In particular, we were conscious of the fact that the push would involve not only the desire to put our own specific technology at the center of the "solution" to whatever design challenge we might identify, but also to view the introduction of a robotic platform in a work environment as an inherently positive intervention. We also knew that we would be managing more than one organizational configuration of disability: our own as HCI researchers, that of the organization that employs the disabled workers, that of the customers on the receiving end of the provided service, and that of the employees themselves, with the latter having potentially the weakest direct representation in the design process (Mankoff et al., 2010). While we did have some access to the employees through observational work, the language barrier and the reluctance of their managers to engage them directly in a participatory design experiment meant that we would have to make design decisions on their behalf. It should be observed that the managers at the outset of the project had no reason to give us unfettered access to employees whose emotional and professional wellbeing they are responsible for, and that they would consider representing their interests in the research activity as their professional responsibility.

In this paper, we outline how we approached the challenge of bringing together the perspectives and concerns of a variety of different stakeholders around future design scenarios, and how throughout that process we tried to uncover and address the risk of the unintended consequences of introducing elements of AI, automation and robotics into a socially and ethically complex and potentially fragile scenario. We do not presume in doing so to bring any kind of privileged understanding of ethical principles that would apply in general to the deployment of autonomous agents in the wild. Standards for ethical principles in AI are by and large agreed upon and documented (IEEE, 2019) which we do not focus on improving here. What we do want to focus on are some of the practicalities of ethical design in collaborative projects. Principles like the ones defined by the IEEE Global Initiative are sound, but are also created within a specific community of practice which many of the stakeholders in our project (and arguably in most applied research projects) are not a part of. When it comes to a general principle like well-being, and particularly in this case the well-being of disabled employees, we question who has the moral and practical authority to decide which actual features, what degree of automation and which changes to an existing workflow will best embody it?

In this paper, we attempt to document the process we underwent in order to try to bring ethical considerations within the practice of a collaborative, multi-stakeholder project. We discuss specific methodological and analytic approaches that we used in two phases of the project. In the first instance, we sought to gain a broad understanding of the organizational principles of our commercial partner. This includes what they consider from their own point of view to be their mission, their responsibilities towards their employees and their well-being, and how they practically set out to accomplish them. In a second phase of the project, we engaged the stakeholders in a speculative design exercise with a goal of bringing their respective ethical considerations and points of view out in open discussion, and attempt to define a broad scenario which could be agreed upon by all. The specific methodologies we used (ethnography and futuristic auto-biographies) are less important than the overall intent, even though these particular methods were well-suited to our purpose and we will therefore describe them in some detail.

The scenario that was ultimately agreed upon and that is described here has not, at the time of writing, been implemented, so we do not have evidence to present of an ultimately successful outcome. One of the reasons we document the process of trying to practically incorporate ethical principles in the preliminary study

and participatory design phases of the project is that the relationship between researchers and the technologies they develop often ends once the technology is released in the wild (Colombino et al., 2019). A longitudinal study of the impact of a new technology in a workplace could prove our ethical considerations to have been right or wrong, and provide valuable insights for future projects. But once ownership of a technology is transferred it may also be impossible to intervene further, which makes it essential that we try to anticipate problems and understand how a technology will be appropriated before it is implemented.

RELATED WORK

In this section, we first discuss previous studies on robots, their roles, and perceptions in the context of workplace collaboration. We also discuss previous research on the use of robots around people with cognitive disabilities. Second, we discuss research on values, and design activities used in HCI and HRI.

Robots in Workplaces

Robots can be found in many workplaces: from order-picking robots in warehouses, delivery robots on university campuses, to bomb disposal robots working alongside teams of soldiers (Royakkers and van Est, 2015). Robots have been used to guide visitors in public places such as museums and airports (Burgard et al., 1999; Fong et al., 2003; Kuno et al., 2007; Kuzuoka et al., 2008). Robots in workplaces can be divided into three broad categories of pre-programmed (e.g., industrial), tele-operated (e.g., drones) and autonomous robots. Autonomous robots are able to sense their environment and act with purpose. Examples include delivery robots in hospitals that distribute and register patients' medicines (Smids et al., 2019).

In previous research on robots being introduced among people in workplaces (Mutlu and Forlizzi, 2008; Dietsch, 2010; Smids et al., 2019), it was found that robots may affect social settings and be experienced as displaying social behavior simply by being and acting among people. How a robot is perceived affects its adoption. Research has found that people unconsciously give the robot human characteristics (Forlizzi and DiSalvo, 2006) or even pet characteristics (Sirkin et al., 2016). The role and perception of autonomous robots have been studied extensively in the context of hospitals. In studies of hospital delivery robots, it is found that a range of factors influence people's perception. The same robots are perceived differently in different hospital units such as postpartum and medical units (Mutlu and Forlizzi, 2008). In one study (Ljungblad et al., 2012), people described the robot as an alien worker or work partner. In other studies, some employees anthropomorphized the robot, whereas others regarded the robot as a machine (Morkes et al., 1999; Siino and Hinds, 2005).

Robots also help to reduce costs and alleviate the complexity of workflows, for example by reducing physical distance, through the deployment of nursing assistant robots (Chen and Kemp, 2010), and courier robots (Evans et al., 1992; Mutlu and Forlizzi, 2008). Researchers conclude that organizational factors such as workflow, political, social, emotional and environment perspectives are related to perceptions (Crawford et al., 1998). Overall, studies suggest that robots in work environments should be designed to respect organizational constraints, facilitate collaboration and willingness to work, while integrating social aspects.

In the context of people with cognitive disabilities, a great deal of research in HRI is dedicated to the use of robots as therapeutic aids in controlled experimental or clinical environments (Javed et al., 2018; Hautop Lund, 2009). Researchers have studied the use of robotic support to enable employees with DDs to perform specific tasks (Baxter et al., 2018; Kidal et al., 2018; Stöhr et al., 2018; Kidal et al., 2019) and propose task-sharing approaches with collaborative robots in the context of an industrial assembly line job in production facilities. Such research shows that often human workers maintain control over the flow of actions and decision making in the face of unexpected situations, while robots execute repetitive tasks. This sharing of tasks is difficult when workers have cognitive disabilities, so this research inverts the traditional role of task-sharing between humans and robots and proposes the concept of the robot as supervisor. However, this research does not take into account the views of stakeholders.

While existing work studied robots in the areas of health and care for people with DDs (Shukla et al., 2019), these contexts usually present controlled environments in which the well being and development of the person with DDs is usually prioritized over organizational constraints. At the same time, there is limited work studying the introduction of robots in workplaces employing people with DDs, and this work is focused on specific aspects of the collaboration during a task. In our case, we see the practical need to go beyond this and understand the organizational properties as well as the perspective of every stakeholder in the company to recognize the potential impact of the introduction of our robots, to then guide our design decisions in a complex setting with vulnerable users.

Value and Human Robot Interaction

The perceptions and values that designers or roboticists have about technology affect their view of "human," "machine" and "robots" (Suchman, 2007; Wallach and Allen, 2009; Suchman, 2011; Richardson, 2015). Even though users experience robots and attribute qualities to them, designers and developers aim to include specific kinds of experience or quality in their design. This inherent bias—relying on their own preference—has been highlighted in research (Oudshoorn et al., 2004).

Design activities (Brandt and Messeter, 2004; Belman et al., 2011; Friedman and Hendry, 2012) and interviews have been often used as a tool to elicit values from the users with few exceptions (Fleischmann and Wallace, 2009; Fleischmann and Wallace, 2010), which were conducted with developers. Design fiction (Bleecker, 2009; Tanenbaum et al., 2012; Blythe, 2014) has been used in HCI as a speculative space (Blythe et al., 2016) that allows researchers to understand the societal impact of future technology (Blythe, 2017) and the values related to it (Dourish and Bell, 2014; Schulte, 2016; Muller and Liao, 2017; Wong et al., 2017). According to Fitzpatrick (2015), for designers to become "responsible," they need to be reflective practitioners, aware of

their power in inscribing futures. This is possible when designers' and roboticists' viewpoints are also considered and made explicit before design. HRI has mostly used narratives, scenarios (Sung et al., 2009) or stories for feedback from intended users on design concepts and prototypes (Robinette et al., 2016; Lichtenthäler et al., 2013). Surveys (Fong et al., 2003), scenario-focused workshops (Caleb-Solly et al., 2014) and sketching of future scenarios (Sung et al., 2009) have been used to study the perception and needs of the users and to evaluate robots.

Robotics researchers involved in previous studies have documented their opinions as robot experts (Scherer, 2014), when evaluating robot prototypes (Sauppé and Mutlu, 2014) or by participating in design sessions (Lee et al., 2014). Cheon and Su (2016)'s study shows that roboticists' engineering background influences their views on the design of robots. Futuristic stories (Cheon and Su, 2017) and futuristic autobiographies (FABs) (Cheon and Su, 2018) have been used to understand the values of roboticists. Futuristic autobiographies, inspired by design fiction, help to elicit values and perspectives from participants such as prospective users, designers, and researchers (Cheon and Su, 2018). They have been proven to indirectly help us understand the values by examining how participants see the proposed situation and map out possible actions. In our research, we used design stories inspired by FABs and this contributes to the existing body of work using design fiction as a research method in HRI. Through our work, we want to understand how stakeholders perceive the future with respect to people with DDs using robots. Though design that reflects on values and ethics has been stressed (Holmquist and Forlizzi, 2014), there is a lack of guidelines for ethical or responsible robot design (Cheon and Su, 2016). We want to go beyond solving the problem of designing current technologies to explore the social and ethical implications of these technologies, incorporating the views of all stakeholders.

UNDERSTANDING THE SETTING

The foundations of design are often best built on a clear understanding of the people, settings, and purposes you are designing for—this reduces mistaken understandings and beliefs and often provides better insight and orients you to the needs of the people you are designing for. In particular we wanted to qualitatively evaluate the social, organizational and technical operation of our partner organization, and consider what sort of problems design could address, how people doing particular activities with particular needs might be supported, or how an innovative concept might either mesh with or disrupt particular work or activities.

In the first phase of our project, two researchers from our team undertook an ethnographic study consisting of three days of observation of the activities of the company (which was engaged in managing a coffee shop, a printshop, a flower shop, a bakery, and the local delivery of in-house products) and semi-structured interviews with the CEO, the educational team (equivalent to HR) and managers from each area.

Ethnography (Martin and Sommerville, 2004) is specifically designed to provide a rich understanding of social phenomena as they occur in everyday settings (Randall et al., 2007). It is qualitative in nature and involves interviews, observation, and participation in natural settings with the specific groups of people you want to study. Our orientation to ethnography is ethnomethodological (Garfinkel, 1967; Garfinkel, 2002) which means that it is not theory-driven but rather focuses on revealing and describing the way in which the people we study organize their activities and their understandings, closely related to how they ordinarily do things themselves, and aiming to minimize the use of technical language. This means that the work stays close to the lived reality of the natural phenomenon itself and that the products of the research are easy to understand across disciplines, which make this approach particularly useful in multi-disciplinary research. While three days is a short period of time to do a full study of all the activities, it was in this case sufficient to provide a good sense of the overall organizational structure and the relationship between its parts. This approach is close to what (Millen, 2000) describes as "rapid ethnography", which sacrifices depth of understanding for a more focused assessment targeted at key individuals and functions.

Ethnographic data can take different forms: general descriptions of behaviours, descriptions of physical layouts, close descriptions of conversation, thoughts and feelings about what is going on, tentative hypotheses, examples, repeated occurrences, responses to questions, etc. While it can be possible to generalize learning beyond the specific context we are looking at, a more essential analytic choice when engaged in a collaborative design activity is to reach a representation of the activity (and all its elements, including technology) at the heart of our specific scenario that is shared by and recognizable to all the stakeholders.

To understand the organizational and socio-technical properties of the setting or scenario we are looking at, we infer motives, purposes and rules of conduct, and give meaning to the activities we observe. We take these elements to be normative, not causal. They do not exist independently of context and are bound up with the cultures, traditions, plans, etc. of the setting we are dealing with. So analytically what we attempt to do is explain them such as they are adopted, observed, recognized and understood, enforced, broken, etc. by the people in that setting.

Through our ethnographic study and semi-structured interviews, we understood that the self-described goal of the organization we were collaborating with is to show the value of disabled workers and demonstrate that it is possible to provide them with gainful employment, given the appropriate organization of the workplace. Indeed, the CEO told us that if other companies were to learn from them how to manage employees with DDs, her organization would no longer need to exist. They have over two hundred employees with varying degrees of DDs in different business units and run most of their operations at a profit. They go to great lengths to deliver products which are indistinguishable from what might be provided by any other print-shop, florist, or bakery, and with a very short turnaround period. They achieve this by breaking down their workflows into basic tasks and implementing a strict division of labor. This means that many of their employees are engaged in

repetitive activities requiring limited initiative or creativity, basic coordination of tasks (as found on a production line), and little need to deal with unexpected occurrences. Tasks requiring more complex social interactions or creative choices and responsibilities are, with few exceptions, handled or closely supervised by non-disabled managers.

An example of this can be found in the flower shop business line developed by our partner organization. This unit is not a brick-and-mortar flower shop with a customer facing physical location but takes business-to-business orders (via phone and e-mail) for flower arrangements and fruit baskets. As mentioned above, the shop operates somewhat like a factory line, where the activities the employees with DDs undertake are broken down into basic tasks. More technically complex tasks such as back-end ordering and invoicing are handled by non-disabled employees. Interestingly from our point of view, more creative but not technically complex tasks requiring, for example, aesthetic judgment, were also mainly handled by the manager. This left most employees performing mundane tasks such as folding ribbons.

A similar scenario can be found in the printing business unit. Like the flower shop, this unit takes mainly business-to-business orders for a variety of print jobs, with a large proportion of these being for business cards. With the print shop having what appears to be a larger variety of jobs, we also observed a broader variety of tasks involving specialized machinery, such as cutting and binding. But the fundamental principle of breaking jobs down into basic tasks and implementing a division of labor remains. As each step of the process is relatively simple on its own, the likelihood that mistakes would be made that might compromise the quality of the final product is minimized. Furthermore, this kind of division of labor creates a collaborative and social environment without creating a need for complex and potentially stressful communication and coordination of dependent activities.

We did observe some exceptions to the way work is organized (as described above). The print shop had one person who managed the print server for one of their digital production presses. This is technically complex work, and the manager of the service explained to us that the technical literacy of the individual combined with his curiosity led him to that role, but that they wouldn't otherwise try to encourage employees to take on more complex tasks. However, pre-defined employment criteria mandate that all employees have the skills to independently navigate to and from work and be able to use standard technology such as phones, TV, etc. In the flower shop, some employees are encouraged to fulfill orders for certain types of potted plants on their own. This is certainly more complex than folding a ribbon, as it involves several steps, and a degree of aesthetic judgment to decide that the final result is good enough. But bear in mind that even the aesthetic judgment involved here is "reduced" to repeatable instructions, such as measuring the distance between different parts of the composition to ensure consistency and balance or symmetry.

The question as to whether more or even most of the organization's employees would be able to learn to adequately perform more complex and creative tasks, given time and attention, was not, we were told, seriously considered by the organization's managers. This was not due to indifference toward the employees' personal development, or (as the examples above demonstrate) a lack of ability to recognize talent where it exists. What was clear was that the viability of their commercial operations had to take precedence over individual learning and development. Consequently, the assessment that employees with DDs cannot handle uncertainty and exceptions is not a clinical judgment or even a character assessment but is appropriate to the requirements of an efficient workflow.

Further evidence of this can be seen in the handling of the one operation that was openly handled at a financial loss and was therefore not subject to the same operational constraints: delivery. The organization has employees personally deliver some of its products (like business cards) to its customers in the metropolitan area. Employees currently assigned to the delivery service are given at the beginning of their shift a backpack, the name and address of the recipient, a receipt form to be signed by the recipient, and a cellular tracking/communication device that they can use to call for help should the need arise. They are then essentially left to their own devices to find their way to the delivery address and back.

We were given the opportunity to accompany one of the employees on a delivery run and were able to make some, to us, revealing observations. Most notably, the employee in question had developed a very nuanced understanding of the vagaries of the underground transport system, and was able to determine, for example, which carriage would allow him to descend closer to the escalator at a connecting station, and to memorize the complex layout of different stations across the network. This demonstrated initiative and a degree of creativity or inventiveness on his part. This was not a new insight for the managers of the service. After all, being able to make their way independently to and from work is required to be hired by the organization. The service also offers employees an opportunity to interact with people independently and with purpose and seems to foster a sense of pride and accomplishment. In spite of it being a mostly individual task, the delivery service also creates opportunities for socialization, as the employees often leave the office together and may travel together for a distance, and even help each other in the case of new or less confident employees.

This is to say that we do not intend to overstate the somewhat Taylorist character of the work of our partner organization, and that we are not implying judgment of their ethics and their overall mission. Their agenda is to demonstrate that gainful employment and financial independence are possible for employees with developmental disabilities, and on their own terms they appear to be quite successful. We can also observe, although this was not central to our project or our analysis, that the Korean work culture (and perhaps Korean society at large) is quite hierarchical and can ask individuals to subordinate their role in the workplace to shared goals and outcomes. From that point of view, what is experienced by the disabled employees we observed could be considered relevant training for what they could expect to experience in other workplaces, i.e., to contribute to the organization rather than lay stress on individual development.

We are nevertheless aware that the operational concerns we witnessed could compete with the clinical and educational configurations of the worker, and that more flexible assessments of the employee's ability to handle uncertainty and develop skills might conflict with concerns about disrupting existing workflows. As researchers, we bring concerns and biases of our own to a future scenario. The most obvious is that being part of an organization that prototypes modular robotic platforms, there is a technology push toward making our platform fit the scenario. This for us is not just a matter of persuading stakeholders that our technology is good or desirable. The introduction of AI and automation to a workplace (and a robotic platform potentially embodies both) is not an ethically neutral action, and how you design the technology is inevitably driven by a vision of what you believe the role of the people and of the technology involved should be.

As HCI researchers, we were particularly struck by the limitations in terms of creativity and the potential for personal development that the organization of work we witnessed can engender. We wondered therefore whether there was an opportunity for robots to enhance the work experience of employees while maintaining the efficiency of the service. Robots could assist the employees to perform their tasks more efficiently but always respecting their role in the process, prioritizing their social and professional skills above process optimization. Robots with an appropriately designed information management system or interface could also enable new types of tasks by providing a structure that standardizes activities not currently performed by the employees due to their flexible nature or a higher level of complexity.

The subsequent step, which is detailed in the next section, was to try and bring as many of the stakeholders as possible together and try to tease out ethical considerations and perspectives around what role and responsibilities robots might have in this workplace, with the aim of converging and agreeing on a potential, concrete scenario.

DESIGN APPROACH

Our proposed robotic platform can independently navigate complex, crowded environments and could be used to transport and deliver objects. The challenge for us was to identify the need for robot collaboration within the service and to propose effective solutions by adapting the functionalities of the robotic platform.

In prior research on robots being introduced among people in workplaces (Mutlu and Forlizzi, 2008; Dietsch, 2010; Smids et al., 2019), it was found that robots may affect social settings and be interpreted as displaying social behavior simply by being there and acting among people. The envisioned future of robots working alongside DD employees requires careful consideration of the organizational, ethical, and societal consequences and values related to robots.

We conducted participatory design sessions with two managers from the company we studied, four engineers in charge of developing the robots, and four designers whose task it was to define and shape the interactions between humans and robots. During the session, we introduced the capabilities of our robotic platform (it has a touchscreen for interaction and is able to navigate autonomously, detect obstacles, carry items) and the participants assessed our technology's feasibility in various services. They selected the service for which the introduction of the robotic platform would be considered most beneficial. The participants outlined the service's challenges and proposed concepts to solve the challenges thus identified. These proposed concepts were assessed relative to the capabilities and limitations of our robot.

We acknowledge that just eliciting design requirements is not enough for a use case that is ethically complex. For the design of our service, it was important to understand the perceptions and values that designers or roboticists have about technology, which affect their view of "human," "machine" and "robots" (Richardson, 2015). The technology stakeholders have different values, which they feel very strongly (Knobel and Bowker, 2011). They aim to create a specific kind of experience or quality in their design. Practically speaking we are facing the challenge of finding a robotic deployment scenario which balances the technological ambition of our own organization with the business model of the recipient one. The disabled employees themselves are potentially caught in the middle of these ambitions.

Bell and Olick (1989) stated that society is re-created each day as people act, calling on both their memories and anticipation. "Arguably, our job as the futurists designing the narratives, is to make the process of re-creation or re-imagining of the society more conscious." During the workshop, we also conducted a value elicitation exercise with the participants. We used futuristic stories, inspired by futuristic autobiographies (FABs) that allow us to understand the societal impact of future technology and help elicit values and perspectives from participants such as prospective users, designers and researchers (Cheon and Su, 2018). By using this method we aimed to restore the future users, people with DDs, to a central position in the minds of our participants when anticipating, designing, and evaluating the future of robotics.

We wanted the participants to go beyond passive imagination and own the futuristic fictions we created. Futuristic autobiographies have been shown to be an effective means of eliciting rich narratives that incorporate participants' experiences, practices, and viewpoints. While design fiction has been defined as the deliberate use of diegetic prototypes to suspend belief about change (Bosch, 2012), in FABs the participant becomes "diegetic" (Cheon and Su, 2018). Instead of having the focus on or around a prototype, the focus of FABs is on the participant itself. We preferred the autobiographical style where these fictions are not perceived as "too abstract" and could be given new meanings through each individual's experience. Unlike previous research, which uses this method only on roboticists, the FABs we created were crafted with the intention of using them on different stakeholders (executives, designers and roboticists).

We conducted the FABs with eight Korean participants, six male and two females, between the age of 24 and 50. Two (P1, P2) participants were managers from the Korean organization employing the people with DDs, who had no prior experience

with robots. The other participants were from the robotics organization. Two (P3, P4) were User Experience Designers responsible for ergonomics of the robot and its interaction with people and the remaining four were robotic engineers (P5–P8). Our participants were selected to represent the stakeholder groups involved in our project.

Each participant was presented with three FABs that were specifically designed according to their stakeholder group. We used the guidelines and cautions presented in Cheon and Su (2018) to create our FABs. The authors researched each participant's background (prior observation of their tasks, their portfolio of work and research interests) to create the FABs for the stakeholder group. These were less than 80 words long, with interesting and plausible scenarios which facilitated open-ended discussions on multiple themes around work collaboration of robots with developmentally disabled people. Some FABs overlapped between the stakeholder groups as they had aspects of information pertaining to both groups. It was also interesting to analyze differing viewpoints about the same scenario. An example of the FAB presented to managers and designers is: "*Recent declarations have caused concern among the executives of the organization. Many employees have stated that they prefer to collaborate with robots as managers rather than other human managers. Who operates the robots? Is it the managers? If yes, what kind of control was given to the managers to determine robot actions? Why would employees prefer robots over humans?*."

The FABs were conducted by four facilitators (authors of the paper) on the premises of our organization. Seven FABs interviews were conducted in English and one was conducted in Korean and concurrently translated by a facilitator to English. The interviews were held in small meeting rooms and were audio recorded. Each session had one participant and one facilitator and lasted 20–30 min.

The interview data was open coded in turns by two researchers to generate themes. For example, they were coded into categories of "perception about robot," "giving human attributes," "role of technology," "safety," "privacy," etc. These themes were reiterated through discussions with other researchers. Our analysis focused on how the participants responded to ethical and social questions regarding the role of robots, its users and their behavior.

FINDINGS

Participants responded to futuristic stories where workplace collaboration between robots and people with DDs is an everyday task. They imagined the type of robots, their intentions for building or deploying them, the tasks performed by them and their impact. The findings cover the emerging themes of human-robot collaboration and the potential positive and negative consequences of introducing robots.

Roles of the Robot

The Robot as an Assistant

Participants imagined very specific tasks: "robots that carry heavy stuff, guidance robots, surveil-lance robots" (P8 - Engineer), "cleaning robot" (P7 - Engineer), "delivery robot" (P6 - Engineer), to more generic tasks like helping in everyday activities in the workplace. Robots will help in enhancing the capabilities of employees and supporting them in doing more. They discussed examples of robots helping them in their current tasks and undertaking new tasks, such as sharing meeting notes, etc. (a job only done by the management).

> "*The robot will help people in their capabilities. . . to increase their capabilities. For example: processing information and maybe provide navigation*" (P6 - Engineer)
>
> "*They can also arrange another meeting and share meeting minutes with other people. Like just ordering the robot like please, send some meeting minutes to someone*" (P7 - Engineer)

The Robot as a Collaborator

Participants also saw the robots as a potential team member who complements the job of the employees with DDs. It supports their job like a partner, being more collaborative and going beyond just enhancing abilities. Participants imagined a positive relationship with collaborators as they will help the employees to do more. The robots and the employees will complement each other and overcome their weaknesses, such as picking up a screw for the robot and forgetting the correct placement for materials for the employees.

> "*For example, there is a robot that screws small things ten times. Maybe the robot can turn the screw exactly ten times, but grabbing the screw is not possible with the current technology. So maybe the employees can help with those things so that final job is done by the robot, but it is sequential, and they complement each other with their different abilities.*" (P2 - Manager)
>
> "*This would ultimately lead to the employees becoming confident. They get feedback from anywhere, including from robots, managers, and other sources. So, there are more things that the employees could do voluntarily.*" (P1 - Manager)

Robot as a Supervisor

Some participants believed that robots will be successful only if they are more intelligent or if they behaved more intelligently than the people with DDs. They imagined the robots to be like a manager, either replacing or helping them in their existing tasks such as counseling and logging. In these situations, the robots were perceived to be superior because replacing or collaborating with the managers placed robots on a similar work-hierarchical scale of the managers. While robots were perceived as positive for the employees as an assistant and a collaborator, some negative consequences of the robot working as a superior were imagined in the supervisory role.

> "*In the (. . ..) meeting they (people with DDs) can talk, and the robot can write down everything they are saying.*

Then the counselor (educational team member) can analyze the meeting minutes or the logs." (P7 - Engineer)

"*They could be surveillance robots. And as surveillance robots, they could be seen as enemies by the employees because they will be watching them and will report to the managers what they did wrong and stuff.*" (P8 - Engineer)

The Human-Robot Relationship

The conversation about the roles of the robots was often complemented by how participants perceived the robot. It went beyond a technological artifact to referring to it as an individual. Participants described robots having an identity higher than a tool. They highlighted a need for establishing good and bad behavior of humans with the robots.

"*I want them to be equal. And sometimes we have to think about robot rights like human rights. I want the robots to evolve to that level of (humans in) cognitive power and physical abilities.*" (P5 - Engineer)

"*So, in this case (employees hitting robots) it's not about robot or human, it is about someone who cannot hit back. I think it's a problem of human behavior so we have to control their interaction, the people and not the robot.*" (P2 - Manager)

Participants discussed the robot's likeness to humans in terms of physical appearance, cognitive skills and actions. This is in line with the theory of regulatory fit [13] which states that an agent (in this case, a robot) that adapts to people's orientation might elicit more cooperation than someone who doesn't. For successful future collaboration, the robot has to have a developed cognitive power capable of "understanding human intentions and emotions" (P7 - Engineer), be "self-learning and updating" (P2 - Manager) and give "more human-like feedbacks" (P1 - Manager).

"*If it is following the human employees, I think that it's a kind of pet or it could be a machine or a human; even though it's a machine, they react as if it's an live object. Then they could act toward the product like a semi-human being. They try to speak.*" (P4 - Designer)

"*I want the robot to behave like a person. When you pass by other people, they communicate with gestures, faces. I want to imagine that the robot communicates with the people the same way as humans do.*" (P6 - Engineer)

While several participants voiced a concern about robots replacing the job of humans, others did not see robots as a threat. They believe robots will prove to be useful and hence can be seen as friends. Others stated that the robots will bring anxiety and fear in humans because of their power. This might create a larger divide between robotic experts and other people, including people with DDs.

"*Robots are smart enough to recognize that if people are not following their instruction, then they can give them, not a star but a negative of a star and you can keep track of it.*" (P8 - Engineer)

This was said in the context of how to make people respect the authority of robots, by building rewards and performance management strategies into the human-robot relationship.

Autonomy and Control

Participants stated that the efficiency of robots is equivalent to its autonomy in navigation, decision making and achieving self-diagnosis. Robots were deemed useful when they bring automation into the process.

"*At first we should make the robot survive in this world without any special help (from the developers). Engineers have to be there always so it cannot survive by itself. You have to charge it. When it goes the wrong way or stops somewhere, put them on track manually. I have not thought about it beyond that, so I really focus on making the robot self-smart. And I actually don't think much about what it can do for us.*" (P5 - Engineer)

"*I am assuming that robots will be autonomous because if there were a manager to each robot then that would be way too inefficient unless some part of it is automated.*" (P3 - Designer)

In the context of the workplace collaboration, the robot was unanimously thought to be controlled by the managers of the organization, although participants had differing viewpoints on the type of interaction and the level of control given to the manager. Managers were understood to have control over the robot's autonomy in different tasks.

"*I am assuming that the kind of control given to the managers would be to designate roles, criteria to focus on, and maybe limit its functions with respect to people's privacy or limiting its function to respect the roles that the humans have.*" (P3 - Designer)

"*Because we (robot experts) cannot control remotely the robot. But this robot should have some kind of intelligent things like autonomously moving or AI chatbot. So, anything can happen. So, someone has to control and maintain the robot. I think that person should be the manager.*" (P4 - Designer)

"*Managers will operate the robot. Who else would?*" (P8 - Engineer)

A few participants also spoke about the safety of deploying robots. They defined this as dependent on the task carried out by the robot. The nature of the task would define appropriate "safety levels" that needs to be thought before implementation. Participants also reflected on privacy of any collected data by the robots. They were unclear on what is ethical in terms of data privacy, an aspect that they admitted to not have considered before. The robot's understanding of "sensitive" and "appropriate" information was also questioned through the FAB narratives.

"*Maybe it can harm people. It cannot counter its force. So, for example when it can give a high-five to people, someone's arm could be broken. It can also go and crash something*" (P7 - Engineer)

"*I am sure it's going to be more of being watched more precisely. Because in the past, the bosses, the big brothers were always watching people. But they were only humans. But robots can be everywhere. And they don't get tired. So, it's going to be more threatening.*" (P8 - Engineer)

"*They just write down all that is said during the meeting, but some part should be erased which is due to security reasons or some small talk. The robot doesn't need to write it or talking bad things behind someone. Robot can't share it all.*" (P7 - Engineer)

Values are context and people dependent. As in this case, privacy is a value held by a person (people with DDs) or it can be held by the organization. It can be intentionally embedded within a technology (monitoring) or materialized by the context of human interaction (writing in meetings).

Value in Terms of Business Needs

The business motivation and needs of this project were rarely forgotten by the participants. On the contrary, business costs, success and issues of profitability were brought up by all participants in one or more discussions. In cases where the futuristic decisions went against the business needs, participants expressed genuine concern. The justification of deploying robots even in less-than-ideal futuristic stories, was often based on a decision about profitability.

"*The employees and the organization have great expectations about collaborating with the robot because it can reduce the workload or somehow have good effects. Somehow the reasons for shutting down (the robot collaboration) is different from the needs of employees or the organization.*" (P2 - Manager)

"*And I don't know ultimately if not having managers will be profitable for the company and there will be more money to be shared among the employees. So, I guess reducing labor will have cost benefits.*" (P3 - Designer)

"*Yeah because using the robot must be cheaper or more efficient than hiring more managers. So, if problems happen, you will have to send for me anyway and that will cost a lot*" (P5 - Engineer)

While participants discussed robots replacing jobs, it was always the manager's job that was thought to be replaceable, not the employees with DDs. This is also aligned with the organizations mission statement to provide more employment to people with DDs.

The Eventuality of Robots

All the participants agreed that robots would be an everyday phenomenon in the future. The eventuality of robots coming into existence was compared with the likes of the industrial revolution and the "digital revolution." This is in line with the technologically deterministic framework (MacKenzie and Wajcman, 1999) of dynamics between technology and society, where society fills a passive role of accepting and adapting to the results of technological innovation. They seemed confident in the ability of users to "adapt" to the new technologies (with a few exceptions).

"*But as you already know, in the early 20th century, we already faced the industrial revolution. So, every businessman would like to reduce the costs of cleaning and other things. So if they just, buy one robot, then can replace 10–20 human beings. Then they can reduce the cost dramatically. This means that people cannot stop developing robots.*" (P7 - Engineer)

"*Maybe 10 or 20 years later, because nowadays people live with smartphones right, the kids. After 20 years, I think the customers can learn and adapt. It takes time, it just takes time.*" (P2 - Manager)

"*Some other generation like the elderly people like they can be afraid of this kind of new technologies. They can't imagine living with the robots. So I want them to use the technology easily. But somehow, someone is not able to. There can be a generation gap.*" (P4 - Designer)

In this eventuality, participants discussed the changing behavior of humans due to the collaboration with robots. They were asked to imagine any behavior that could possibly be "unnatural". They defined what is "unnatural" for them and how it would affect people with DDs through their narratives.

"*It is unnatural that the human follows the machine*" (P8 - Engineer)

"*If they started serving more human tasks that require more emotional attachment but is still in the form of a tool kind of, even if that's the case, people could feel like it's unnatural.*" (P3 - Designer)

One participant imagined people becoming extremely dependent on robots, causing them to lose their survival skills. This prediction of the future suggests the removal of the societal norm to work. This means that people with DDs won't need jobs as a measure to assert independence or build their skills.

"*I think many people will think that robots will make people lazy. For example: if robots cook for you all the time, all the human chefs will lose their jobs and after 100 years no human will be able to cook their own meals. If something happens to the robotic systems in the world, many people will die from hunger because they cannot cook. People might worry about this. Robots might remove the human ability to survive on their own because they will rely on robots for everything. More than society, everyone believes that you have to work to get paid and robots might change that so even if you don't*

work, robots can do all the jobs for you and you can play all the time." (P5 - Engineer)

Hence, participants in this context defined "unnaturalness" as a disruption of normal social relations and hierarchies which would have to be carefully handled when incorporating robots in the future. Unnaturalness as discussed with participants can have extreme consequences. People with DDs struggle with "normal" or "natural" routines. In case of robots doing all the works in the society, people with DDs would not need to struggle to learn "old survival" skills but will have to invest themselves in learning the new "survival skills" of operating a robot. This can also lead to extreme alienation of this vulnerable population.

DESIGN CONSIDERATIONS

The two-phase study of ethnographic observations and the design workshop with the FABs activity enabled a deep understanding of the organization, the users, and the ethical perspectives of stakeholders. In this section, we outline some of the design considerations we defined from this understanding.

Robot Roles in the Organization

During the FABs interviews, the executives and designers favored robots that were helpful in the tasks performed by or for users. Also, all participants justified the use of technology for being cost-effective and catering to business profitability. The executives, being true representatives of the people with DDs, stated that they would discontinue using robots if they become a physical or emotional threat to the employees. However, other participants believed that any loss of humanness by using robots is a small sacrifice for higher capabilities, productivity and power. Indeed, when looking at the company, most services adhere to a serialized organization of their processes that results in simple and repetitive tasks for each employee. This serialization reduces the occurrence of unexpected situations guaranteeing the emotional safety of the employees and also maintains the consistency of the outputs.

In this organizational structure, we identified a tension between the opportunities for personal development of the employees and the viability of the commercial operations. Robots, as assistants rather than supervisors, could effectively mediate between these two objectives, enhancing the work experience of employees and the possibilities for social and professional development while maintaining the efficiency of the service. At the same time, robots could enable new, more challenging or creative tasks, by providing a structured framework for activities that are not currently performed by the employees due to their higher level of complexity.

User Characteristics

The intended users, imagined by the participants during the activity, were both the managers and people with DDs. Managers were seen as primary users, while people with DDs were seen as secondary users of technology. This viewpoint of robots' design and use by the managers raises the issue of a socio-technical gap. In a case study (Grudin, 1988) of organizational interfaces or "group-ware" applications, the challenge of uneven distribution of benefits of these applications among members of the organization stands out. While these applications are designed to provide a collective benefit, some members of the organization may need to do more work, which may result in the rejection of the system. If we consider robots as a version of a new age "organizational interface," these uneven distributions might be more pronounced. While the managers are also new users to robotic technology, people with DDs might be affected more as they will go through higher adjustments. It raises the question, "Are we marginalizing the people with DDs?" Hence, we identified the need for simplistic and relatable interfaces and structures for both managers and employees to perform tasks.

Another aspect highlighted by the executives during the study was the passiveness of being a technology consumer. When discussing the problems a robot could have, P2 said: "But I don't answer any more because it is not a problem from our side." This reveals the dependency of users on designers and developers of technology and the vast responsibility they undertake without being aware of it. Technology development is often left in the hands of experts (Šabanovic, 2010) and hence understanding and mediating the values of these "experts" with those of the users using human-centered approaches become very important.

Adaptive Autonomy

Participants saw the robot as a collaborator that would enhance the work of the employees. They should perform tasks autonomously, but also side-by-side with the employees to overcome each other weaknesses. Participants mentioned that, in some cases, a human should take control of the robot leveraging her/his existing knowledge. *Sliding Autonomy* is a strategy that integrates the autonomous capabilities of robots with the reliability that human control can bring in the completion of complex tasks (Heger and Singh, 2006). By changing the level of autonomy of the robot, this strategy makes it possible to leverage and combine the capabilities of humans and robots and adapt the degree of control to address new or unexpected situations (Tang et al., 2016). In the case of our robotic platform, this is an aspect that could be considered to address technical limitations. For instance, navigation could be dynamically adapted, allowing managers or employees to temporarily guide the robots in specific situations, e.g. to overcome obstacles or find the way through a crowded space.

Most participants imagined the person controlling the robots to be the manager. They had, however not imagined negative consequences of this control, like "increased sense of surveillance," "abnormal workplace hierarchy," etc. While control should be given to the managers for a few complex tasks or in specific instances respecting the workplace hierarchy, an environment of distrust between the employees and the managers should be avoided. Hence, the control of the system should be distributed among all its users rather than a subset.

Beyond this, adaptive support could be provided by the robots to generate new learning opportunities for the employees with

DDs. By gradually decreasing the level of support the robot provides during the tasks, employees can increasingly gain independence in their work. However, the option to rely on a higher level of support remains available.

Managing Expectations

Through our research, we were able to identify the expectations stakeholders had of the robotic platform. While we found common points across groups (i.e., organization, managers, designers, engineers), it was interesting to identify the differences in their perspectives. As designers, we should be able to reach compromises that address these, sometimes, contradictory expectations.

In this case, all participants looked at future robots as a real work "partner." Roboticists displayed a high level of attachment to robots and were passionate advocates of technology. Their take on the futuristic stories had technology as a central theme, and their utmost concern was indeed the level of technological advancement, which would enhance or limit the robot's utility. Their imagined future problems were limited to "software," "sensors," "algorithms," or "control system." Executives were rather focused on the employees with DDs. They were conscious of creating and setting strict boundaries of user interaction with the robots. For example, the need to teach the people with DDs that hitting a robot is bad as these could translate to their real-world interaction with other humans. Indeed, collaboration with robots came with expectations based on the dynamics of human-human work collaboration.

All participants declared that if we wanted the robots to work with us, then we have to make them like us. This meant higher cognitive powers, ability to understand, process, respond to human actions and emotions humanely. However, the participants also appreciated and wanted robots to retain their "robotic" quality. "Robotic" quality was equated with consistency. Previous research has shown that when users perceive the robot's actions to be less predictable, they anthropomorphize the robots more to reduce the feeling of uncertainty (Waytz et al., 2010). This brings out the different yet coexisting perspectives of wanting the robots to be simple and objective and yet more human-like. Designers were aware of the probability of users over-estimating robot's technical capabilities as a major source of future discontent in the work environment. Indeed, managing user expectations is especially imperative in the current context of working with the vulnerable group of people with DDs. They are equivalent to the naive users discussed in Cheon and Su (2017), who have high expectations beyond what the roboticist intended to program. Often technology is introduced as a marketing gimmick that attracts and disillusions the users. As developers and designers of technology, it is important to manage user expectations through the physicality of robots' form, shape, and size.

CONCEPT PROPOSAL

Guided by the ethical considerations discussed in our findings, we defined the role of the robot to be the assistant of the employee with DDs. Indeed, it should not replace them in their activities, but rather improve the conditions of their work and augment human capabilities, aiming to increase their independence, agency, and learning opportunities. This was a central aspect for us and was in line with the concerns and expectations mentioned by the some of the participants during the FABs activity. Rather than replacing their dependency on managers with a dependency on robots (with a supervisory role), the robots should act as a support system ready to act in case of need. Also, as stated by the CEO of the company, the goal of the organization was that of employing the maximum possible number of people with DDs. This was considered a possibility by replacing or reducing the need of managers through introducing robotic automation. While, as designers of technology, we usually see the replacement of human labor as one of the potential risks of introducing automation technologies, in this case we found ourselves in a more complex setting with no simple answer, proving the importance of a deep understanding of every specific case and the involvement of every stakeholder in the process. Hence, a compromise was reached which was to reduce the ratio of managers per employee. It would reduce the costs and ultimately allow them to hire more people with DDs. These considerations were later considered when defining the features of the robot.

The plant management service was identified as the most suitable scenario for the introduction of the robotic platform during the participatory design workshop. The setting was ideal to incorporate automation to alleviate the workload of the managers while augmenting the skills of the employees by fostering independence and improving work conditions. In this service, teams of one manager and two employees with DDs manage the indoor plants of corporate offices around the city. They perform activities such as watering plants and cutting dead leaves with the manager having the additional responsibility of supervision. It is an activity that requires physical efforts, as they need to carry the water. Most employees work under the close supervision of their manager as even the most independent employees need regular advice for verification and unexpected situations (work emergency), which constitutes the following instances: spilled water, dead or fallen plants, and instances when questions were asked to them by the people working in the company they visit.

After identifying this scenario, we conducted additional interviews with the managers of the plant management service to complement the understanding we gained during the previous interviews, observations, and workshop activities. We performed a detailed analysis of the characteristics of the service environment, the tasks, the users, as well as the capabilities of our robotic platform, to identify the moments of intervention in which the robot could provide valuable support to the employees, and the best way to implement it.

The teams of one manager and two employees with DDs visit the plant management sites once a week and the time taken differs according to the number of plants it has (shortest: 40 plants >1.5 h and largest 350 plants >6–7 h). They need to perform their activities while employees of the other company are working in the office, so special consideration had to be taken to avoid

FIGURE 1A | Snippet of the questionnaire filled by the employees regarding their perception about the robots.

disrupting their activities. The employees are trained through a booklet which includes the basic step-by-step guide to do each task (watering, cutting leaves, etc.). It outlines the plant's need for water, sun, and shade with the plant's picture and name. Sometimes employees carry this booklet for consultation. The managers give repetitive generic instructions about the plants at the start of each shift. The interviews and written instructions helped us construct a detailed work journey. The major challenges of the service for the employees and moment of intervention were identified through this activity: carrying the heavy bucket of water, filling the water, and controlling the amount of water for each plant, since it changes based on size and season; need for assistance on where to cut the plant if it is brown or dying, missing a plant that needs to be given water.

Our earlier conducted observation studies helped us understand the characteristics of the people with DDs employed by the company. Indeed, 'people with cognitive disabilities can refer to a wide range of conditions that can go from very limited to high functioning individuals. We were able to understand the characteristics of the employees of this specific company by understanding their hiring process through interviews conducted with managers and HR. By requiring the ability to travel independently between home and work, and by asking to complete a number of manual tasks during the hiring process, they set clear boundaries for the capabilities that people with cognitive disabilities should have in order to work with them. The employees were transferred to different business units to acquire different skills. Additionally, the employees with DDs working in this service need to have certain specific qualities. There are often cases of outbursts (screaming, making loud noises, etc.) triggered by stressful situations given the conditions of the employees. A requirement for the employees of the plant management service is to a have very good control of their emotions, considering that they perform their tasks surrounded by the office workers of the client company.

We complemented our understanding with a questionnaire for the employees with DDs. The objective of this questionnaire was to provide a better understanding about the relationship employees have with technology and their perception about the robots. The questionnaire was filled out by ten employees (two females and eight males, aged 21–32 years old) who had worked in various business units. All respondents were comfortable using mobile phones, and most of them (8/10) used computers and TVs regularly. They used emojis (**Figure 1A**) to express their feeling and wrote down their impressions of how the robot looked. Most respondents were Surprised, Happy, or Neutral about the robots of our organization. They perceived it as a tool that could help them in general (move things around, guide people, give information, measure the environment), or in specific work activity (assistant in baking, automatic delivery, labeler for printing). Their positive responses indicated that they were curious about robots.

Our robotic platform (**Figure 1B**) is able to carry objects, autonomously navigate through spaces and interact with the users through a GUI. We proposed a number of additional functionalities that we deemed required for this context. Indeed, the plant management service consists of several tasks that require different levels of automation and collaboration with the employee. The robot should be able to adapt to provide the right level of support to both managers and the employees. At the same time, we had to consider the special needs of the employees with DDs when defining the HRI elements.

Based both on ethical considerations as well as on the specific characteristics of the plant management service, we proposed a robotic platform that would provide physical, cognitive and enable emotional support to the employees with DDs during their activities in the plant management service.

In terms of physical support, the robots would assist the employees by carrying the water, one of the most physically demanding tasks mentioned during the interviews by the managers. In order to do this, the robot would be motorized and able to move independently, as well as to follow the employees during their routes in the offices. Indeed, we proposed this navigation mode, i.e., the robot following rather than guiding the employee, to give agency to the employee over the robot and avoid putting them in a passive role.

The platform would also provide cognitive support by giving information about the amount of water needed by each type of plant, according to the humidity conditions and the season. Also, the robot would guide them on how to maintain the plants, especially which leaves to cut and how to do it properly. Aiming to facilitate learning, the information displayed would reduce progressively over time to allow employees to strengthen their own knowledge, with the robot ready to provide additional information or feedback when needed.

Finally, we planned the platform to enable emotional support for the employees, even in the absence of managers nearby. Employees would be able to easily request manager assistance in stressful situations through the platform. The managers would then be able to monitor and use the robot as a proxy to communicate with the employee and help them by assessing the situation and decide if additional assistance would be required. This type of remote monitoring and interaction would allow one manager to supervise several employees at the same time.

FIGURE 1B | The robotic platform developed by our organization.

One important aspect to be considered in the following design stages, which was discussed during our findings, is what level of control would managers get over the robot and how they should handle the remote monitoring to avoid a sense of surveillance. It is possible to solve this problem by giving varying control of the robots to the manager. It means that employees and the managers are responsible for their individual robots and they can communicate with each other through the robots when required. To mitigate the sense of surveillance, the managers would oversee the location of each employee robot in real time when the robot performs only the autonomous tasks such as filling water and reporting to the office floor. Both the employees and the managers would have access to the shift report, which includes information such as the area covered (plants watered) in the office.

The physical aspect of the robots as well as the interactions are planned to present the robot as a tool and minimize any anthropomorphism to prevent any emotional attachments and false expectations regarding possible interactions. A touch and visual-based interface (buttons and screen) is proposed with screen and light feedback to maintain familiarity with the mode of operation by users.

CONCLUSION

Throughout this paper we have tried to articulate some of the complexities involved in designing a robotic platform for employees with cognitive and developmental disabilities. We have taken a practical approach to resolving the ethical stakes in our project, and treated them, from a methodological point of view, as emergent issues rather than as *a priori* considerations.

We do not intend to suggest that high-level debates about ethical, responsible AI is not important. The themes that characterize current research on the ethics of AI, such as privacy concerns, responsibility and the delegation of decision making, transparency, and bias (Coeckelberg, 2020) are all very much pertinent to our scenario. But we also take on the analytic perspective that rules and norms of behavior are situational and negotiable, and therefore emerge and are made relevant in and through practice (Phillips, 1992). This is consistent with our tradition of ethnography-based design, and in our project, we therefore treated socio-technical and ethical issues as practical, emergent matters to be understood from the perspective of the actors we are designing for. And as Dewsbury et al. (2004) point out, an applied technology project moves forward not through political rhetoric but through recommendations for design.

As we pointed out in the introduction, the context-specific dimension of ethical questions requires attention that cannot simply be satisfied by high-level ethical principles alone. Everyone in our project can claim to have the well-being of the disabled employees at heart. And yet this does not mean that everyone will agree on what role technology and automation can or should play in ensuring that. Different parties will also be subject to different organizational imperatives or incentives. All this has to be unpacked and from a methodological point of view attention has to paid to the uniqueness of each project and scenario. This implies an analytic stance which is not necessarily oriented to the generalizability of the findings. And the people we involved in our study should not be though of as a "sample"—we do not examine our own motivation as researchers through a process of reductionism, and consequently we do not advocated doing so for other stakeholders in a collaborative project.

In terms of understanding the emergent ethics of our scenario, this project faced two broad challenges. The first challenge was that we had very limited access to the more vulnerable of our end-users, the disabled employees. While this is obviously not ideal, access to this population of users was owned and mediated by our partner organization, and this is also a scenario that is not uncommon in applied research conducted on behalf of and partnership with private and commercial entities. Given the nature of our partner organization, we knew they would most likely have their own mission statement with respect to their employees, and that our first responsibility was to understand how their own ethics influenced and were influenced by their organization of work, processes and managerial actions, which is to say how their ethics were practiced (Clegg et al., 2007).

The access we did have in the course of our brief observational study, along with the interviews with the service managers and CEO of the organization, allowed us to identify what, through the lens of our own agenda (to provide a positive role for our technology platform), appeared to be an interesting practical compromise between providing opportunities for personal development of the employees and the viability of a commercial operation. This for us represented an opportunity for proposing a technology design scenario that would enhance the intrinsic interest of the work itself for the employees without compromising (if not enhancing) the efficiency of the workflow.

The second major challenge, which is one faced by many technology design projects, is that the other stakeholders involved came from various disciplines (interaction and user experience designers, machine learning experts, and mechanical engineers) and were likely to have different agendas and ways of framing the ethical stakes involved. In fact, many of the people involved in our project (including ourselves) belong to professional categories

that do not, for the most part, have a clearly defined professional code of ethics, and consequently may not have been in the habit of managing ethical considerations as part of their work in the first place.

The design exercises based on futuristic autobiographies, which we conducted with our stakeholders in the second phase of the project, forced everyone involved to confront each other's priorities and concerns on the use of robots for people with DDs. It also allowed us to identify a broader set of ethical issues, not necessarily to provide a generalizable theoretical contribution to the ethics of AI, but to better manage the expectations of everyone involved while appreciating the risks of unintended consequences that were not obvious to us at the start of the project, and to ultimately agree on a shared scenario and set of features.

All of this is just a first step and careful and iterative empirical testing of the design concept will be required, ideally with more direct involvement of the disabled employees themselves, as we would not claim that preliminary studies and participatory design will ever allow us to anticipate all the ways a technology might be appropriated and how things might go wrong. But we do hope that we have made a compelling argument for the value of a practice-based understanding of ethics and that our discussion of how we approached the challenges in our project from a methodological point of view has been informative.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

TC: writing of abstract, introduction, conclusion, understanding the setting, and conducted ethnographic observational study. DG: writing of design considerations and concept proposal, co-organizer and animator of participatory design workshop, co-responsible for original service concept design, assisted in the ethnographic observational study. SS: writing of related work, design approach and findings, coorganizer, and animator of participatory design workshop, analysis of FABs outputs, and co-responsible for original service concept design. YI: coorganizer and animator of participatory design workshop, organizer of ethnography (observational study), general editing tasks. SC: co-organizer and animator of participatory design workshop, organizer of ethnography (observational study), general editing tasks. The first three authors (TC, DG, and SS) share a co-first authorship for this publication.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frobt.2021.643160/full#supplementary-material

REFERENCES

Baxter, P., Lightbody, P., and Hanheide, M. (2018). "Robots Providing Cognitive Assistance in Shared Workspaces," in Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction -HRI '18, March 5-8, 2018 (Chicago, IL, USA: ACM Press), 57–58. doi:10.1145/3173386.3177070

Bell, W., and Olick, J. K. (1989). An Epistemology for the Futures Field: Problems and Possibilities of Prediction. *Futures* 21 (2), 115–135. doi:10.1016/0016-3287(89)90001-3

Belman, J., Nissenbaum, H., and Flanagan, M. (2011). "Grow-A-Game: A Tool for Values Conscious Design and Analysis of Digital Games," in 5th International Conference on Digital Research Association: Think Design Play, DiGRA 2011, September 14-17, 2011 (Utrecht, Netherlands)

Bleecker, J. (2009). Design Fiction: A Short Essay on Design, Science, Fact and Fiction. *Near Future Lab.*, 49.

Blythe, M., Andersen, K., Clarke, R., and Wright, P. (2016). "Anti-Solutionist Strategies," in Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16, May 7–12, 2016 (Santa Clara, California, USA: ACM Press), 4968–4978. doi:10.1145/2858036.2858482

Blythe, M. (2017). "Research Fiction," in Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17, May 6–11, 2017 (Denver, Colorado, USA: ACM Press), 5400–5411. doi:10.1145/3025453.3026023

Blythe, M. (2014). "Research through Design Fiction," in Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14, April-May 26-01, 2014 (Toronto, Ontario, Canada: ACM Press), 703–712. doi:10.1145/2556288.2557098

Bosch, T. (2012). Sci-Fi Writer Bruce Sterling Explains the Intriguing New Concept of Design Fiction. Available at: https://slate.com/technology/2012/03/ bruce-sterling-on-design-fictions.html. Accessed August 09, 2021.

Brandt, E., and Messeter, J. (2004). "Facilitating Collaboration through Design Games," in Proceedings of the eighth conference on Participatory design Artful integration: interweaving media, materials and practices - PDC 04, July 27-31, 2004 (Toronto, Ontario, Canada: ACM Press), 121. doi:10.1145/1011870.1011885

Burgard, W., Cremers, A. B., Fox, D., Hähnel, D., Gerhard, G., Schulz, D., et al. (1999). Experiences with an Interactive Museum Tour-Guide Robot. *Artif. Intelligence* 114, 3–53. doi:10.1016/S0004-3702(99)00070-3

Caleb-Solly, P., Dogramadzi, S., Ellender, D., Fear, T., and Heuvel, H. v. d. (2014). "A Mixed-Method Approach to Evoke Creative and Holistic Thinking about Robots in a home Environment," in Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction - HRI '14, March 3-6, 2014 (Bielefeld, Germany: ACMPress), 374–381. doi:10.1145/2559636.2559681

Chen, T. L., and Kemp, C. C. (2010). "Lead Me by the Hand," in Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction - HRI '10, March 2-5, 2010 (Osaka, Japan: ACM Press), 367. doi:10.1145/1734454.1734579

Cheon, E., and Su, N. M. (2017). "Configuring the User," in Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17, February-March 25-1, 2017 (Portland, Oregon, USA: ACM Press), 191–206. doi:10.1145/2998181.2998329

Cheon, E., and Su, N. M. (2018). "Futuristic Autobiographies," in Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction - HRI '18, March 5-8, 2018 (Chicago, IL, USA: ACM Press), 388–397. doi:10.1145/3171221.3171244

Cheon, E., and Su, N. M. (2016). "Integrating Roboticist Values into a Value Sensitive Design Framework for Humanoid Robots," in 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), March 7-10, 2016 (Christchurch, New Zealand: IEEE), 375–382. doi:10.1109/HRI.2016.7451775

Clegg, S., Kornberger, M., and Rhodes, C. (2007). Business Ethics as Practice. *Br. J. Manag.* 18, 107–122. doi:10.1111/j.1467-8551.2006.00493.x

Coeckelberg, M. (2020). *AI Ethics*. Cambridge, United Kingdom: The MIT Press.

Colombino, T., Willamowski, J., Grasso, A., and Hanrahan, B. (2019). "Deeper into the Wild: Technology Co-creation Across Corporate Boundaries," in *Into the Wild: Beyond the Design Research Lab*. Editors A. Chamberlain and A. Crabtree. Studies in Applied Philosophy, Epistemology and Rational Ethics. Springer.

Crawford, S. Y., Grussing, P. G., Clark, T. G., and Rice, J. A. (1998). Staff Attitudes about the Use of Robots in Pharmacy before Implementation of a Robotic Dispensing System. *Am. J. Health-System Pharm.* 55, 1907. doi:10.1093/ajhp/55.18.1907

Dewsbury, G., Clarke, K., Randall, D., Rouncefield, M., and Sommerville, I. (2004). The Anti-social Model of Disability. *Disabil. Soc.* 19 (2), 145–158. Available at: http://www.tandfonline.com/doi/abs/10.1080/0968759042000181776. doi:10.1080/0267303042000249224

Dibia, V., Trewin, S., Ashoori, M., and Erickson, T. (2015). "Exploring the Potential of Wearables to Support Employment for People with Mild Cognitive Impairment," in Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS '15), Lisbon, Portugal, October 26-28, 2015 (New York, NY, USA: ACM), 401–402. doi:10.1145/2700648.2811390

Dietsch, J. (2010). People Meeting Robots in the Workplace [Industrial Activities]. *IEEE Robot. Automat. Mag.* 17, 15–16. doi:10.1109/MRA.2010.936950

Dourish, P., and Bell, G. (2014). "Resistance Is Futile": reading Science Fiction Alongside Ubiquitous Computing. *Pers Ubiquit Comput.* 18, 769–778. doi:10.1007/s00779-013-0678-7

Evans, J., Krishnamurthy, B., Barrows, B., Skewis, T., and Lumelsky, V. (1992). Handling Real World Motion Planning A Hospital Transport Robot. *IEEE Control. Syst.* 12 (1), 15–19. doi:10.1109/37.120445

Fitzpatrick, G. (2015). "Inscribing Futures through Responsible Design by Responsible Designers," in *Responsible Innovation*. Editors A. Bogner, M. Decker, and M. Sotoudeh (Baden-Baden: Nomos Verlagsgesellschaft mbH and Co. KG), 71–78. doi:10.5771/9783845272825-71

Fleischmann, K. R., and Wallace, W. A. (2009). Ensuring Transparency in Computational Modeling. *Commun. ACM* 52, 3. doi:10.1145/1467247.1467278

Fleischmann, K. R., and Wallace, W. A. (2010). Value Conflicts in Computational Modeling. *Computer* 43 (7), 57–63. doi:10.1109/MC.2010.120

Fong, T., Nourbakhsh, I., and Dautenhahn, K. (2003). A Survey of Socially Interactive Robots. *Robotics Autonomous Syst.* 42, 143–166. doi:10.1016/S0921-8890(02)00372-X

Forlizzi, J., and DiSalvo, C. (2006). "Service Robots in the Domestic Environment: A Study of the Roomba Vacuum in the Home," in HoRI '06: Prceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction, Salt Lake City Utah, USA, March 2-3, 2006, 258–265. doi:10.1145/1121241.1121286

Friedman, B., and Hendry, D. (2012). "The Envisioning Cards," in Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12, March 2-3, 2006 (May 5-10, 2012: ACM Press), 1145. doi:10.1145/2207676.2208562

Garfinkel, H. (2002). *Ethnomethodology's Program: Working Out Durkheim's Aphorism*. Lanham, MD: Rowman & Littlefield.

Garfinkel, H. (1967). *Studies in Ethnomethodology*. Cambridge: Polity Press.

Gaudion, K. (2016). *Building Empathy: Autism and the Workplace*. London: The Helen Hamlyn Centre for Design. Royal College of Art.

Grudin, J. (1988). "Why CSCW Applications Fail: Problems in the Design and Evaluationof Organizational Interfaces," in Proceedings of the 1988 ACM conference on Computer-supported cooperative work - CSCW '88, September 26-28, 1988 (Portland, Oregon, United States: ACM Press), 85–93. doi:10.1145/62266.62273

Heger, F., and Singh, S. (2006), Sliding Autonomy for Complex Coordinated Multi-Robot Tasks: Analysis & Experiments. *Robotics*. doi:10.15607/RSS.2006.II.003

Holmquist, L. E., and Forlizzi, J. (2014). Introduction to Journal of Human-Robot Interaction Special Issue on Design. *J. Hum.Robot Interaction* 3, 1. doi:10.5898/jhri.3.1.holmquist

Javed, H., Burns, R., Jeon, M., Howard, M. A., and Park, C. (2018). A Robotic Framework to Facilitate Sensory Experiences for Children with Autism Spectrum Disorder: A Preliminary Study. *ACM Trans. Hum.-Robot Interact.* 1, 18. doi:10.1145/3359613

Kidal, J., Martín, M., Ipiña, I., and Maurtua, I. (2019). Empowering Assembly Workers with Cognitive Disabilities by Working with Collaborative Robots: a Study to Capture Design Requirements. *Proced. CIRP* 81, 797–802. doi:10.1016/j.procir.2019.03.202

Kidal, J., Maurtua, I., Martin, M., and Ipiña, I. (2018). "*Towards Including Workers with Cognitive Disabilities in the Factory of the Future*," in ASPSETS '18: roceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility, Galway, Ireland, October 22-24, 2018, 426–428. doi:10.1145/3234695.3241018

Knobel, C., and Bowker, G. C. (2011). Values in Design. *Commun. ACM* 54, 26–28. doi:10.1145/1965724.1965735

Kuno, Y., Sadazuka, K., Kawashima, M., Yamazaki, K., Yamazaki, A., and Kuzuoka, H. (2007). "Museum Guide Robot Based on Sociological Interaction Analysis," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '07, San Jose, California, USA, April-May 28-3, 2007 (San Jose, California, USA: ACM Press), 1191–1194. doi:10.1145/1240624.1240804

Kuzuoka, H., Pitsch, K., Suzuki, Y., Kawaguchi, I., Yamazaki, K., Yamazaki, A., Kuno, Y., Luff, P., and Heath, C. (2008). "Effect of Restarts and Pauses on Achieving a State of Mutual Orientation between a Human and a Robot," in Proceedings of the ACM 2008 conference on Computer supported cooperative work - CSCW '08, November 8-2, 2008 (San Diego, CA, USA: ACM Press). doi:10.1145/1460563.1460594

Lee, H. R., Šabanovic, S., and Stolterman, E. (2014). "Stay on the Boundary," in Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14, April-May 26-01, 2014 (Toronto, Ontario, Canada: ACM Press), 1471–1474. doi:10.1145/2556288.2557395

Lichtenthäler, C., Peters, A., Griffiths, S., and Kirsch, A. (2013). "Be a Robot! Robot Navigation Patterns in a Path Crossing Scenario," in Proceedings of the 8th ACM/IEEE International Conference on Human Robot Interaction, Tokyo Japan, March 3-6, 2013 (Tokyo, Japan). Available at: https://hal.archives-ouvertes.fr/hal-01684317doi:10.1109/hri.2013.6483561

Ljungblad, S., Kotrbova, J., Jacobsson, M., Cramer, H., and Niechwiadowicz, K. (2012). "Hospital Robot at Work," in Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12, Seattle, Washington, USA, February 11-15, 2012 (Seattle, Washington, USA: ACM Press), 177. doi:10.1145/2145204.2145233

Hautop Lund, H. (2009). "Modular Playware as a Playful Diagnosis Tool for Autistic Children," in IEEE International Conference on Rehabilitation Robotics, Kyoto, Japan, June 23-26, 2009 (IEEE). doi:10.1109/ICORR.2009.5209606

MacKenzie, D., and Wajcman, J. (1999). "Introductory Essay: the Social Shaping of Technology," in *The Social Shaping of Technology*. Editors D. MacKenzie and J. Wajcman. 2nd ed. (Buckingham, UK: Open University Press) ISBN 9780335199136.

Mankoff, J., Hayes, G. R., and Kasnitz, D. (2010). "Disability Studies as a Source of Critical Inquiry for the Field of Assistive Technology," in Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility - ASSETS '10, Orlando, FL, USA, October 25-27, 2010 (Orlando, Florida, USA: ACM Press), 3. doi:10.1145/1878803.1878807

Martin, D., and Sommerville, I. (2004). Patterns of Cooperative Interaction. *ACM Trans. Comput. Hum. Interact.* 11 (1), 59–89. doi:10.1145/972648.972651

Millen, D. R. (2000). "Rapid Ethnography," in Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques (DIS '00), New York City, NY USA, August 17-19, 2000 (New York, NY USA: Association for Computing Machinery), 280–286. doi:10.1145/347642.347763

Morkes, J., Kernal, H. K., and Nass, C. (1999). Effects of Humor in Task-Oriented Human-Computer Interaction and Computer-Mediated Communication: A Direct Test of SRCT Theory. *Human-Computer Interaction* 14 (Dec. 1999), 4395–4435. doi:10.1207/S15327051HCI1404_2

Muller, M., and Liao, V. Q. (2017). "Exploring AI Ethics and Values through Participatory Design Fictions," in HCIC 2017: Design Futures, Pajaro Dunes, Watsonville, CA, June 25-29 2017.

Mutlu, B., and Forlizzi, J. (2008). "Robots in Organizations," in Proceedings of the 3rd international conference on Human robot interaction - HRI '08, Amsterdam, Netherlands, March 12-15, 2008 (Amsterdam, Netherlands: ACM Press), 287. doi:10.1145/1349822.1349860

Oudshoorn, N., Rommes, E., and Stienstra, M. (2004). Configuring the User as Everybody: Gender and Design Cultures in Information and Communication Technologies. *Sci. Technol. Hum. Values* 29, 30–63. doi:10.1177/0162243903259190

Phillips, N. (1992). "Understanding Ethics in Practice: An Ethnomethodological Approach to the Study of Business Ethics," in *Business Ethics Quarterly* (Cambridge University Press), 2.

Randall, D., Harper, R., and Rouncefield, M. (2007). *Fieldwork for Design: Theory and Practice*. New York: Springer-Verlag. doi:10.1007/978-1-84628-768-8

Richardson, K. (2015). *An Anthropology of Robots and AI: Annihilation Anxiety and Machines*. New York, NY: Routledge. doi:10.4324/9781315736426

Robinette, P., Li, W., Allen, R., Howard, A. M., and Wagner, A. R. (2016). "Overtrust of Robots in Emergency Evacuation Scenarios," in 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), New Zealand, March 7-10, 2016, 101–108. doi:10.1109/HRI.2016.7451740

Royakkers, L., and van Est, R. (2015). *Just Ordinary Robots: Automation from Love to War*. 1 ed. Boca Raton: CRC Press. doi:10.1201/b18899

Šabanovic, S. (2010). Robots in Society, Society in Robots: Mutual Shaping of Society and Technology as a Framework for Social Robot Design. *Int. J. Soc. Robotics* 2, 4439–4450. doi:10.1007/s12369-010-0066-7

Sauppé, A., and Mutlu, B. (2014). "Design Patterns for Exploring and Prototyping Human-Robot Interactions," in Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14, Toronto Ontario Canada, April-May 26-01, 2014 (Toronto, Ontario, Canada: ACM Press), 1439–1448. doi:10.1145/2556288.2557057

Scherer, D. C. (2014). Movie Magic Makes Better Social Robots: the Overlap of Special Effects and Character Robot Engineering. *J. Human-Robot Interaction* 3, 123–141. doi:10.5898/JHRI.3.1

Schulte, B. F. (2016). "Using Design Fiction to Reflect on Autonomy in Smart Technology for People Living with Dementia," in Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct - UbiComp '16, Heidelberg Germany, September 12-16, 2016 (Heidelberg, Germany: ACMPress), 1110–1113. doi:10.1145/2968219.2972717

Shiomi, M., Iio, T., Kamei, K., Sharma, C., and Hagita, N. (2014). "User-Friendly Autonomous Wheelchair for Elderly Care Using Ubiquitous Network Robot Platform," in In Proceedings of the Second International Conference on Human-Agent Interaction - HAI '14, Tsukuba, Japan, October 29-31, 2014 (Tsukuba, Japan: ACM Press), 17–22. doi:10.1145/2658861.2658873

Shukla, J., Cristiano, J., Oliver, J., and Puig, D. (2019). Robot Assisted Interventions for Individuals with Intellectual Disabilities: Impact on Users and Caregivers. *Int. J. Soc. Robotics* 11, 631–649. doi:10.1007/s12369-019-00527-w

Siino, R. M., and Hinds, P. J. (2005). "Robots, Gender & Sensemaking: Sex Segregation's Impact on Workers Making Sense of a Mobile Autonomous Robot," in Proceedings of the 2005 IEEE International Conference on Robotics and Automation, Barcelona, Spain, April 18-22, 2005 (Barcelona, Spain: IEEE), 2773–2778. doi:10.1109/ROBOT.2005.1570533

Sirkin, D., Mok, B., Yang, S., and Ju, W. (2016). "Oh, I Love Trash: Personality of a Robotic Trash Barrel," in Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion - CSCW '16 Companion, San Francisco, California, USA, February-March 26-2, 2016 (San Francisco, California, USA: ACM Press), 102–105. doi:10.1145/2818052.2874336

Smids, J., Nyholm, S., and Berkers, H. (2019). Robots in the Workplace: a Threat To-Or Opportunity For-Meaningful Work? *Philos. Technol.* 33, 503–522. doi:10.1007/s13347-019-00377-4

Stöhr, M., Schneider, M., and Henkel, C. (2018). "Adaptive Work Instructions for People with Disabilities in the Context of Human Robot Collaboration," in 2018 IEEE 16th International Conference on Industrial Informatics (INDIN), PORTO, Portugal, July 18-20, 2018 (IEEE, Porto), 301–308. doi:10.1109/INDIN.2018.8472070

Suchman, L. (2007). *Human-Machine Reconfigurations: Plans and Situated Actions*. Cambridge, United Kingdom: Cambridge University Press.

Suchman, L. (2011). Subject Objects. *Feminist Theor.* 12 (2), 119–145. doi:10.1177/1464700111404205

Sung, J., Christensen, H. I., and Grinter, R. E. (2009). "Sketching the Future: Assessing User Needs for Domestic Robots," in RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication, Toyama, Japan, September-October 27-2, 2009 (Toyama, Japan: IEEE), 153–158. doi:10.1109/ROMAN.2009.5326289

Tanenbaum, T. J., Tanenbaum, K., and Wakkary, R. (2012). "Design Fictions," in TEI '12: Proceedings of the Sixth International Conference on Tangible, Embedded and Embodied Interaction, Kingston Ontario Canada, February 19-22, 2012, 347–350. doi:10.1145/2148131.2148214

Tang, F., Mohammed, M., and Longazo, J. (2016). "Experiments of Human-Robot Teaming under Sliding Autonomy," in 2016 IEEE International Conference on Advanced Intelligent Mechatronics (AIM), Banff, Canada, July 12-15, 2016, 113–118. doi:10.1109/AIM.2016.7576752

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019). The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems*. First Edition. IEEE. Available at: https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html.

The National Autistic Society (2016). The Autism Employment Gap: Too Much Information in the Workplace. Available at: https://www.autism.org.uk/get-involved/media-centre/news/2016-10-27-employment-gap.aspx (Accessed December 10, 2019).

Wallach, W., and Allen, C. (2009). *Moral Machines*. New York, NY: Oxford University Press.

Waytz, A., Cacioppo, J., and Epley, N. (2010). Who Sees Human? *Perspect. Psychol. Sci.* 5 (3), 219–232. doi:10.1177/1745691610369336

Wong, R. Y., Mulligan, D. K., Van Wyk, E., Pierce, J., and Chuang, J. (2017). Eliciting Values Reflections by Engaging Privacy Futures Using Design Workbooks. *Proc. ACM Hum.-Comput. Interact.* 1, 1–26. doi:10.1145/3134746

Yamazaki, K., Yamazaki, A., Ikeda, K., Liu, C., Fukushima, M., Kobayashi, Y., et al. (2016). "I'll Be There Next". *ACM Trans. Interact. Intell. Syst.* 5 (4), 1–20. doi:10.1145/2844542

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Robotics and AI

ORIGINAL RESEARCH
published: 29 April 2022
doi: 10.3389/frobt.2022.676767

Trust and Cooperation

*Benjamin Kuipers**

Computer Science and Engineering, University of Michigan, Ann Arbor, MI, United States

We AI researchers are concerned about the potential impact of artificially intelligent systems on humanity. In the first half of this essay, I argue that ethics is an evolved body of cultural knowledge that (among other things) encourages individual behavior that promotes the welfare of the society (which in turn promotes the welfare of its individual members). The causal paths involved suggest that *trust* and *cooperation* play key roles in this process. In the second half of the essay, I consider whether the key role of trust exposes our society to existential threats. This possibility arises because decision-making agents (humans, AIs, and others) necessarily rely on simplified models to cope with the unbounded complexity of our physical and social world. By selecting actions to maximize a utility measure, a well-formulated game theory model can be a powerful and valuable tool. However, a poorly-formulated game theory model may be uniquely harmful, in cases where the action it recommends deliberately exploits the vulnerability and violates the trust of cooperative partners. Widespread use of such models can erode the overall levels of trust in the society. Cooperation is reduced, resources are constrained, and there is less ability to meet challenges or take advantage of opportunities. Loss of trust will affect humanity's ability to respond to existential threats such as climate change.

Keywords: ethics, cooperation, trust, society, evolution, unknown unknowns, existential threat

OPEN ACCESS

Edited by:
Martim Brandão,
King's College London,
United Kingdom

Reviewed by:
David Gunkel,
Northern Illinois University,
United States
Saeed Hamood Alsamhi,
Ibb University, Yemen

***Correspondence:**
Benjamin Kuipers
kuipers@umich.edu

Specialty section:
This article was submitted to Ethics in Robotics and Artificial Intelligence, a section of the journal Frontiers in Robotics and AI

Received: *06 March 2021*
Accepted: *21 February 2022*
Published: *29 April 2022*

Citation:
Kuipers B (2022) Trust and Cooperation. Front. Robot. AI 9:676767. doi: 10.3389/frobt.2022.676767

1 INTRODUCTION AND OVERVIEW

Like many researchers in Artificial Intelligence (AI), I am concerned about the impact of the increasing success of our field on the welfare of humanity. This has led many of us to look for ideas in the fields of Ethics, both philosophical and applied. And of course, to the work of anthropologists, psychologists, sociologists, historians, and others who have contributed important ideas about the roles of ethics in human society. Even in the last few years, these efforts have led to numerous books and journal articles, at least two major international conferences, and many workshops.

Although originally trained in pure mathematics, I have spent my career as an AI researcher focused on commonsense knowledge, especially cognitive maps of the spatial environment, and more generally knowledge of foundational domains (e.g., space, dynamical change, objects, actions, etc.) that help an intelligent agent make sense of its world in a computationally tractable way. This has involved reviewing literature across multiple disciplines for insights and constraints on useful representations for states of incomplete knowledge that arise during development, learning, planning, and acting.

Ethics can be viewed as another domain of foundational knowledge–a critical one at this point in time. In this essay, I describe a view from AI and robotics of certain roles that ethics plays in the welfare of humanity, and the implications of that view for how AI systems should function.

1.1 Terminology

This paper uses a set of terms that are familiar to many people, but which are used quite differently by different people and in different disciplines and contexts. Here are some key definitions, describing how I use these terms, followed by commentary.

An *agent* is an entity (natural or artificial) that perceives its environment, builds an internal representation, and takes actions to pursue its goals within its model of that environment [(Russell and Norvig, 2010), p.4].

A *society* is a collection of agents that share an environment and interact with each other [(Rawls, 1999), p.4]. Therefore, the environment for each agent includes the actions of other agents and their effects.

Cooperation is the process of two or more agents acting together for a common purpose or benefit (Tomasello et al., 2012). Coordinated individual efforts can result in greater benefits than the sum of what the individuals can accomplish (Wright, 2000; Nowak and Roger, 2011).

Trust is defined here as "a psychological state comprising the intention to accept vulnerability based on positive expectations of the intentions or behavior of another" [(Rousseau et al., 1998), p.1998]. This builds on a seminal model of trust (Mayer et al., 1995) that includes ability, benevolence, and integrity as three factors contributing to perceived trustworthiness.

Ethics[1] is a body of knowledge describing how a person should act in particular situations, and what sort of person one should try to be [(Shafer-Landau, 2013), p.xi]. Ethical knowledge is generally shared by members of a given society [(Tomasello, 2019), p.249].

1.2 Commentary

The term "agent" is used here as in the fields of artificial intelligence and multi-agent systems, encompassing both human and artificial goal-oriented actors [(Russell and Norvig, 2010), p.4]. This is not the sense of "agent" meaning someone who acts for another, the principal.

All agents, human and non-human, act to pursue goals. However, virtually all observed actions are motivated by subgoals within plans to achieve higher-level subgoals, perhaps quite distant from any ultimate goal.

In setting the foundation for his theory of justice, John Rawls writes [(Rawls, 1999), p.4] that "a society is a more or less self-sufficient association of persons who in their relations to one another recognize certain rules of conduct as binding and who for the most part act in accordance with them."

Human agents belong to many overlapping societies, each of which may have its own ethics. The individual agent has the task of deciding what ethical knowledge applies to the current situation. The relationship between artificial agents and human societies is an important research topic.

Cooperation is a relationship among agents, which each have goals of their own, requiring the agents to resolve conflicts among individual and collective goals, as illustrated by the Prisoner's Dilemma and other laboratory games. The collective behavior of a system of components, where the components are not "agents" capable of choosing actions to pursue their goals within the environment as they perceive it, is not considered "cooperation" by the definition used here. For example, robust distributed communication protocols such as the Internet's TCP/IP (Cerf and Kahn, 1974) and Drone/IoT communication (Alsamhi et al., 2019) are sometimes described in terms such as "collaboration" or "cooperation" because each node in a network maintains and updates a table of accessible nodes, and the protocol selects paths for transmitting packets based on the connectivity represented by these distributed tables. Although the similarities are undeniable, we consider this case to be outside of our definition of "agent" because of the limited state and decision freedom of the nodes.

Cooperation often involves vulnerability, due to the risk of exploitation by one's cooperative partners, who might contribute less than their share, or might take more than their share of the rewards. Therefore, voluntary cooperation requires trust of one's partners, accepting vulnerability in the confident belief that it will not be exploited. Some cases described as "cooperation without trust" (Mayer et al., 1995) involve coerced cooperation, where credible threat of punishment eliminates risk of exploitation. Other cases (Cook et al., 2005) rely on a much stronger definition of "trust", closer in meaning to "devoted love", so some examples of cooperation do not involve "trust" in this strong sense.

The definition of trust above (from (Rousseau et al., 1998), inspired by (Mayer et al., 1995)) is clearly motivated by interpersonal trust between individuals who know each other, such as the trust between partners in crime in the Prisoner's Dilemma. Of course, the word "trust" is used in many other contexts, typically with overlapping but not identical meanings. For example: trust in an attribute of an inanimate object, such as the strength of a rope, or the accuracy of a sensor; trust in the individual or corporation that manufactured or supplied that inanimate object; trust in corporate or government entities, such as the security of a savings account in a bank, or the safety and efficacy of medications allowed on the market by the FDA; trust in generic (not individually known) members of my community, such as believing that other drivers will virtually always stop at red lights, allowing me to drive confidently through a green light; and even, interpersonal trust "because we think you take our interests to heart and encapsulate our interests in your own" (Cook et al., 2005). Some of these cases are enforced by law, but it is widely recognized by legal scholars that voluntary compliance with social norms, rather than the threat of legal penalties, is primarily responsible for widespread trustworthy behavior (Posner, 2000; Posner, 2007). While these are different contexts and senses of the word "trust", they share the social benefits described in **Section 3**.

In this paper, I extend the terms "cooperation" and "trust" to situations described by "social norms", where the cooperative partners are not identified individuals, but are generic other members of the same society. For example, we trust that other drivers will stay on the correct side of the road as they drive, and

[1]Some scholars distinguish between "morality" (meaning personal beliefs) and "ethics" (meaning societal teachings), though occasional writers swap the two meanings. I follow the example of philosopher Peter Railton, who writes "I will be using 'morality' and 'ethics' (and 'moral' and 'ethical') interchangeably" [(Railton, 2003), Note 1, p.xx].

will behave appropriately at stop signs and traffic lights. Near-universal obedience to these norms (and many others) makes vehicle transportation safer and more efficient for everyone involved.

Ethical knowledge is generally (though not perfectly and universally) shared by members of a given society, but it varies significantly over historical time and geographical space. Traditionally, ethical knowledge is only possessed by humans, but scholars have begun to consider how ethics applies to non-human agents such as AIs and institutions.

1.3 Overview: The Importance of Trust

The first half of this essay proposes a relationship among these key concepts (**Figure 1**), drawing on related work in philosophy (**Section 2**), cooperation and trust (**Section 3**), and evolution (**Section 4**).

Humanity is made up of individual humans, the agents who make decisions about how to act. Humans organize themselves into societies. Early in human evolution, societies were small isolated bands of hunter-gatherers (Tomasello, 2019). Since then, societies have grown larger, more complex, nested and overlapping in various ways. A society gets resources from the efforts of its individual members, and the individual members are supported and protected by the physical and cultural strength of the society (Wright, 2000; Christakis, 2019).

Among the assets of a society are bodies of accumulated cultural knowledge that are distributed among its individual members. This includes a great deal of "how-to" knowledge such as how to prepare specific foods and how to build specific artifacts (Henrich, 2016). The shared body of cultural knowledge also includes the *ethics* of the society, which helps to direct individuals away from possibly-tempting action choices, and toward actions that are better for the society in the long run, and therefore also better for the individual (Beauchamp and Childress, 2009; Fedyk, 2017).

We observe important similarities and striking variation in the content of the ethical knowledge in different societies, both across historical and pre-historical time, and across the different societies and cultures that exist around the world. Within a given society, knowledge is transmitted from one generation to the next through a variety of mechanisms including imitation and explicit teaching. These imperfect learning methods introduce variations, some of which fade away while others grow, persist, and displace other beliefs. The structural similarities with Darwinian evolution suggest that cultural evolution is a real and important process complementing the properties of genetic evolution (Dawkins, 1976; Richerson and Boyd, 2005; Pinker, 2011; Buchanan and Powell, 2018).

A society gets resources from the efforts of its individual members, but those efforts can be multiplied through cooperation. Mechanisms for cooperation include teamwork, specialized expertise, division of labor, economies of scale, military organization and discipline, markets, capital investments, common infrastructure, and many others. Cooperation benefits the society as a whole, as well as the individuals directly involved (Curry et al., 2016; Curry et al., 2019).

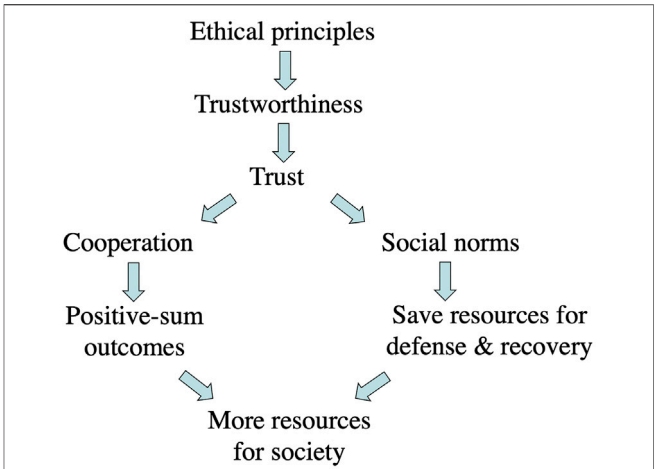

FIGURE 1 | From ethics to resources.

Trust and trustworthiness are widely recognized as important to the successful functioning of society (Fukuyama, 1995). A particularly important role for trust is the support of cooperation, which involves vulnerability to one's cooperative partners. Another important role of trust is to reduce complexity and uncertainty, making it feasible to make plans by focusing on only a few possible alternatives (Luhmann, 1979; Nissenbaum, 2001).

One role of the ethical principles of a society is to help individual members of the society know how to be trustworthy, and how to recognize when others are trustworthy. **Figure 1** summarizes some of the relationships among ethics, trust, cooperation, and resources for society. (This is not to argue that support for trust and cooperation are the *only* functions of ethics.)

The ethical principles of a society determine what it is to be trustworthy, and thus who or what is trusted. Trust enables cooperation which produces more resources. Trusted social norms can be counted on, saving resources. The nature and degree of trust in the society determines whether the society will have a shortage or plenty of resources, and hence whether it thrives or not in future generations.

Given the centrality of trust to the processes that provide resources for society (as shown in **Figure 1**), if trust is eroded, society is threatened. Lack of trust decreases both willingness to cooperate and confidence in social norms, making it harder to meet threats or exploit opportunities, resulting in scarcity of resources. As societies get larger and more complex, they increasingly rely on trust–of individuals, of institutions, and of social norms. Erosion of trust and loss of resources can bring a successful, complex society to the point of collapse (Tainter, 1988; Diamond, 2005).

1.4 Overview: The Vulnerability of Trust

The second half of this essay addresses the question of how trust can erode in a successful complex society.

The physical and social world we inhabit is unboundedly complex. To reason effectively, we necessarily create simplifying *models* to capture a few relevant elements of that world for current purposes, leaving all of the rest of the

complexity out. Technical fields in science and engineering explicitly study the creation and evaluation of models, but simplifying models are unavoidable in everyday life and common sense as well.

A model is created for a particular purpose, and it explicitly describes a limited set of elements of the world and the relations among them. We might call these the "*known unknowns*." We need to provide values for some of these elements in order to reason with the model, and the relationships within the model help us determine values for the others. Everything else about the infinitely complex world is treated as *negligible*–aspects of the world that we assume may be neglected for the purposes of this model. We might call these aspects the "*unknown unknowns*".

Reasoning with incomplete knowledge–models–carries risk, but is also necessary to make it possible to draw useful conclusions. For example, reasoning about how gravity determines orbits is impossible without the simplifying "point mass assumption" that treats each body–Sun, planets, spacecraft–as if its entire mass were concentrated at a single point at its center of mass. This, of course, abstracts away geography, so that within this model, the distinction between, say, Western/European and Eastern/Asian cannot even be expressed. All this means is that one must use one model to reason about orbits, and a different one to reason about geography.

This essay presents and uses simplified, incomplete, descriptions of ethics, cooperation, trust, and evolution. Are these therefore "bad models" in the sense discussed later (in **Section 7**), purely by virtue of being incomplete and omitting major aspects of those topics? Not necessarily, any more than the point mass model of orbiting bodies is a bad model. After developing appropriate preliminaries, I will distinguish between harmful and useful models, drawing attention to certain types of models that may be harmful to trust in our society, leading to potentially catastrophic consequences.

To complete the astronomy analogy, suppose our goal is to predict eclipses. In the first step, a simplified model embodying the point mass assumption is used to identify precise orbits for the Earth and Moon about the Sun. The second step uses a different model, treating the Sun, Earth, and Moon as extended bodies of certain sizes and shapes (whose relative motions are now known), so we can reason about the shadows they cast and where those shadows will fall. Neither model is adequate by itself, and combining the two models is too complex, but the problem can be solved by applying one model to the first sub-problem and the other to the second.

In many cases, the simplification embodied by a model is reasonable and makes inference more efficient. But in cases where the elements omitted from the model are important, then conclusions drawn from that model may be badly wrong. The proper and improper creation and use of models is discussed in more detail in **Sections 6**, **7**.

One dramatic example is the "Prisoner's Dilemma", where a straight-forward application of the powerful modeling method of game theory (von Neumann and Morgenstern, 1953; Leyton-Brown and Shoham, 2008) leads to a bad outcome due to over-simplified modeling assumptions. Another dramatic example relevant to autonomous vehicles (AVs) is the "Moral Machine" (Awad et al., 2018) where a narrowly-framed model forces a choice between two terrible evils, while a wider framing would provide a more plausible, realistic, and favorable solution. (Both are discussed in **Section 7**).

One possible impact of an improperly simplified model is to erode trust between potential partners and make cooperation less likely in the future. If the utility measure in a game theory model is not sensitive to trust, cooperation, or the welfare of society, then the algorithm will deliberately choose actions that exploit the vulnerabilities of other players. The overly-simple formulation of the decision model not only leads to a bad outcome, but it "poisons the well" for further decisions by discouraging trust. A generalized lack of trust can lead to inability to respond effectively to existential threats such as climate change (**Section 8**).

2 RELATED WORK IN PHILOSOPHY

2.1 Traditional Schools of Thought in Philosophical Ethics

Morality and ethics have been important to human society for thousands of years.

What is ethics? One philosopher responds, "*At the heart of ethics are two questions: 1) What should I do?, and 2) What sort of person should I be?*" [(Shafer-Landau, 2013), p.xi]. Another philosopher says, "*At its most basic, ethics is about . . . the kind of life that is most worthy of a human being, the kind of life worth choosing from among all the different ways we might live*" [(Vallor, 2016), p.2].

For centuries, moral philosophers have searched for principles to describe the moral judgments that people should make. Strong candidates include virtues (Hursthouse and Zalta, 2013), duties (Alexander et al., 2015), contractual agreements (Ashford and Mulgan, 2018; Cudd and Eftekhari, 2018), and utility maximization (Driver, 2014; Sinnott-Armstrong, 2015). No consensus has been reached.[2] However, a repeated theme is that ethics helps balance the selfish interests of the individual decision-maker against the interests of other individuals or of the society as a whole.

Virtue ethics describes ethics in terms of the characteristic virtues of exemplary individuals, and how they confront particular problems. Aristotle (Aristotle, 1999) compares virtues to skills like carpentry, gained through experience and practice until they become automatic. A current philosopher like Shannon Vallor (Vallor, 2016) proposes "technomoral virtues" extending the traditional virtues to meet the demands of modern

[2]Consider the lesson of the children's poem, "*The Blind Men and the Elephant*" (Saxe, 1949). Six men, highly educated but blind, conclude that the elephant must be very much like a wall, or a snake, or a leaf, or a spear, or a tree, or a rope, corresponding to the part that each has experienced, while none grasps the complex whole. Fragmentary truths may be useful and important, but must be recognized as incomplete.

technological developments. The computational methods for AI knowledge representation best suited for virtue ethics are *case-based reasoning* (López, 2013) and *analogical reasoning* (Forbus et al., 2018). These methods describe specific situations in the world, actions taken, and their results and evaluations. Actions applied in past situations can be retrieved and adapted to new situations, leading to increasing experience and expertise.

Deontology is the study of duty (*deon* in Greek), which describes ethics in terms of obligations and prohibitions, offering simplicity, clarity, and ease of explanation, but raising the question of how the duties are determined. Immanuel Kant responded in 1785 with his categorical imperative: "*Act only according to that maxim whereby you can, at the same time, will that it should become a universal law*." To apply this concept to the complexity and diversity of modern society, John Rawls (Rawls, 1999) proposed that "*The principles of justice are chosen behind a veil of ignorance*," meaning without knowledge of the situation that one would personally occupy under those principles. The obligations and prohibitions of deontology are well suited to the expressive power of computational rules and constraints, which are standard tools for knowledge representation and inference in AI (Russell and Norvig, 2010). Isaac Asimov's Three Laws of Robotics (Asimov, 1952) have a deontological character, but they also illustrate (through fiction) how an apparently straight-forward duty, for example "*A robot may not injure a human being or, through inaction, allow a human being to come to harm*", can be complex and ambiguous in practical application.

Utilitarianism is the position that "*the morally right action is the action that produces the most good*" (Driver, 2014). It is a form of consequentialism, that "*the right action is understood entirely in terms of consequences produced*" (Driver, 2014). In philosophical utilitarianism, one maximizes *everyone's* good, not just the good of the decision maker. This is in contrast with the computational methods of *game theory* (von Neumann and Morgenstern, 1953; Leyton-Brown and Shoham, 2008). On the one hand, game theory provides a powerful mathematical formalism for utilitarian calculations, including concepts of probability, discounting, and expected utility. On the other hand, the focus of game theory is on each decision-maker's efforts to maximize their *own* utility measure (called "egoism" in (Driver, 2014)). Nonetheless, thanks to its computational power and conceptual clarity, game theory has become a near-standard for action selection in artificial intelligence and is often treated as the definition of "rationality" [(Russell and Norvig, 2010), p.611]. Recently, advocates for this "standard" view of rationality in AI have begun to reconsider their position (Russell, 2019).

2.2 Ethics and Artificial Intelligence

In recent decades, AI researchers have begun to create artificial entities capable of learning from data, representing knowledge, solving problems, making decisions, and taking action in our physical and social environment. Whether these entities are embodied as robots such as autonomous vehicles or are disembodied decision support systems deciding whether people get jobs, credit, or parole, they are effectively participating as members of human society.

Interest in the field of AI Ethics has grown rapidly, driven by important concerns about the impact of AI technology on human society: safety, privacy, surveillance, facial recognition, bias and fairness, polarization, etc (Christian, 2020; Kearns and Roth, 2020). Early contributions (Anderson and Anderson, 2006; Wallach and Allen, 2009; Lin et al., 2012) drew heavily on the major schools of thought in philosophical ethics.

Work in the AI Ethics research community is directed at several questions: 1) What sorts of ethical impacts are implemented AI systems likely to have on humans and human society? 2) How can AI systems be designed to make their ethical impacts on humans more positive, or at least, less negative? 3) How can we analyze and measure the impact of a particular implemented AI system on humans?

The "technomoral virtues" proposed by philosopher Shannon Vallor (Vallor, 2016) recognize that new technologies may present new and demanding ethically fraught situations requiring new (or newly framed) virtues extending the more traditional virtue ethics framework. Philosopher John Sullins (Sullins, 2020) further explores Vallor's categories of technomoral trust and honesty, observing with concern that humans appear to have an innate tendency to trust others that can be exploited by designers of robots (Robinette et al., 2016). While humans do often exhibit initial trust, it is well known that trust can be lost and may or may not be regained. Indeed, the TIT-FOR-TAT algorithmic strategy that won two successive tournaments of the Repeated Prisoner's Dilemma game starts with initial trust, and then responds according to the partner's action on the previous cycle (Axelrod, 1984).

Value Sensitive Design (VSD) (Friedman et al., 2013; Friedman et al., 2021) is a general methodology for designing information systems to be compatible with human values. AI and robotic systems are embodied information systems, embedded along with humans in the physical world, so they are an important particular case for VSD methods. The concept of trust, especially for online activities, has also been analyzed by VSD researchers (Friedman et al., 2000; Nissenbaum, 2001).

Most people feel that ethical human decision-makers should be able to provide comprehensible explanations for their conclusions, and that AI decision-makers should be held to the same standard. Unfortunately, current state-of-the-art decision performance comes from deep neural network systems trained on extremely large training sets, and both their training and their operation are too complex for comprehensible explanation. This is often seen as a choice between high-performance but incomprehensible systems, vs. explainable but lower-performing systems. Wachter, et al. (Wachter et al., 2018) take a different approach, explaining the decision outcome for a given case by synthesizing artificial cases, similar to the given case, but with small differences sufficient to change the decision outcome. These "counterfactual" cases provide an explanation, not of the actual mechanism of the decision, but of the features of the case most responsible for its outcome. In a more recent paper, Mittelstadt and Wachter (Mittelstadt et al., 2019) contrast typical human styles of

explanation with the model-based approaches typical in explainable-AI research. Focusing on model-based explanation of complex AI models such as deep neural networks, they discuss the limitations of simple human-comprehensible models as approximations to DNN models.

Philosophers, computer scientists, AI researchers, and experts in other areas have focused on specific aspects of AI and ethics. Computer scientist Noel Sharkey is a leader in the movement to ban killer robots (Sharkey, 2012). Philosopher Patrick Lin was among the first to propose a "Trolley Problem" analogy for autonomous vehicles (Lin, 2013), which has gone on to inspire the "Moral Machine" online survey experiment (**Figure 4**) (Awad et al., 2018). Some philosophers express skepticism about the relevance of ethics for robots because of supposed fundamental differences between humans and robots (van Wynsberghe and Robbins, 2019; Nyholm and Smids, 2020). Some of my own previous papers (Kuipers, 2018; Kuipers et al., 2020) explore the importance of trust to society, the appropriateness of different AI representations to ethical knowledge, and examples from several domains of what humans would want to count on from non-human agents.

Many scientific, professional, governmental, and public interest organizations in the United States, United Kingdom, and EU have formulated principles and recommendations for ethical constraints on artificial intelligence and its deployment (Cath et al., 2017). Drawing on these, the 2018 AI4People report (Floridi et al., 2018) categorizes the risks and opportunities from AI research and deployment, proposes five general principles (beneficence, non-maleficence, autonomy, justice, and explicability), the first four based on well-understood principles from applied biomedical ethics (Beauchamp and Childress, 2009). The report assumes without a definition that the reader understands the terms "trust" and "trustworthiness." The report concludes with a list of 20 action recommendations intended to help create a "Good AI Society" based on AI technologies.

In 2019, the European Commission's High Level Expert Group on Artificial Intelligence published its "Ethics Guidelines for Trustworthy AI" (High Level Expert Group on AI, 2019a), and in 2020 published an expanded "Assessment List for Trustworthy AI (ALTAI)" (High Level Expert Group on AI, 2020a). Two additional reports provided policy and investment recommendations (High Level Expert Group on AI, 2019b; High Level Expert Group on AI, 2020b).

These Guidelines begin with three abstract ethical principles–respect for human autonomy, prevention of harm, and fairness and explicability–plus the need to assess both benefits and risks of AI deployment, with particular attention to vulnerable groups. The Guidelines provide seven key requirements that implemented AI systems should meet: 1) human agency and oversight, 2) technical robustness and safety, 3) privacy and data governance, 4) transparency, 5) diversity, non-discrimination, and fairness, 6) environmental and societal well-being, and 7) accountability. Finally, it provides an assessment list (updated in 2020) for evaluating an implemented system.

The Guidelines provide a definition for trust in its glossary: "Trust is viewed as: 1) a set of specific beliefs dealing with benevolence, competence, integrity, and predictability (trust in beliefs); 2) the willingness of one party to depend on another in a risky situation (trusting intention); or 3) the combination of these elements" (Siau and Wang, 2018). The definition I use (**Section 1**) subsumes clauses (1) and (3) under a statement similar to (2), but the meanings are quite similar.

3 RELATED WORK ON COOPERATION AND TRUST

Evolutionary theorists characterize *homo sapiens* as a "hyper-cooperative species," and attribute our success as a species to the positive-sum results of cooperative action (Wright, 2000; Richerson and Boyd, 2005; Tomasello, 2019). Cooperation among individuals often yields rewards much greater than the total those individuals could obtain separately. Cooperation provides substantial advantages when faced with threats from human enemies or other predators, or when taking advantage of opportunities for obtaining more resources.

However, in a cooperative enterprise, each partner is vulnerable to exploitation by the other partners. Successful cooperation requires *trust*:

> *"Trust is a psychological state comprising the intention to accept vulnerability based on positive expectations of the intentions or behavior of another." [(Rousseau et al., 1998), p.1998]*

Where that trust exists, cooperation is possible, the society benefits from more positive-sum ("win-win") interactions, and it tends to grow in resources. Where that trust does not exist, cooperation is much less viable, interactions are more often zero-sum or negative-sum, and the society tends to lose resources. Fewer resources, and decreased ability to mount a cooperative response to a crisis (external attack, ecological failure, epidemic disease, climate change, etc.), means that a society that once could surmount a crisis through cooperative action, no longer can, and may collapse (Tainter, 1988; Diamond, 2005).

A society has its own set of norms that show its individual members how to act in order to be considered trustworthy (Posner, 2000). They also show what sorts of behavior by others provides evidence that they are (or are not) trustworthy. Some norms, such as prohibitions against killing, stealing, breaking promises, or driving on the wrong side of the road, provide direct benefits in terms of safety. Other norms, like customs in clothing, speech, and table manners, signal that one belongs to a particular society. The presumption that in-group members are more likely to be trustworthy, while out-group members are less likely to be, encourages trust and cooperation among members of the society. However, this mechanism also encourages discrimination and racism against non-members (Posner, 2000; Van Bavel and Packer, 2021).

A contrary argument by Cook, Hardin and Levi in "*Cooperation Without Trust*?" (Cook et al., 2005) depends on

a restrictive definition of trust: "According to this conception of trust, we trust you because we think you take our interests to heart and encapsulate our interests in your own. . . .. By 'encapsulate' we mean that to some extent our interests become yours in the trust relation between us" [(Cook et al., 2005), p.5]. Further: "Note that the conception of trust as encapsulated interest implies that *many interactions in which there is successful coordination or cooperation do not actually involve trust*." [(Cook et al., 2005), p.8, emphasis theirs]. Under the broader definition cited above, the acceptance of vulnerability necessary for cooperation does require trust.

3.1 Is Ethics Only for Cooperation?

Anthropologist Oliver Scott Curry and his colleagues present a theory, "Morality as Cooperation" (MAC) (Curry et al., 2016; Curry et al., 2019), arguing that "morality consists of a collection of biological and cultural solutions to the problems of cooperation recurrent in human social life" [(Curry et al., 2019), p.48].

Curry and others quote an array of philosophers back to Plato and Aristotle in support of the strong connection between morality and cooperation and the common good. Based on evolutionary biology and game theory, they describe seven different problems of cooperation: 1) the allocation of resources to kin; 2) coordination to mutual advantage; 3) social exchange; 4) hawkish and 5) dovish displays of traits for resolving conflicts; 6) division; and 7) possession. Cooperative solutions to these problems explain corresponding types of morality: 1) family values; 2) group loyalty; 3) reciprocity; 4) bravery; 5) respect; 6) fairness; and 7) property rights.

Curry, Mullins, and Whitehouse (Curry et al., 2019) describe the predictions of the MAC theory for what should be considered good or bad in particular cultures, and present the results of testing those predictions against 60 societies studied by anthropologists and described in the Human Relations Area File (HRAF). They found that the predicted cooperative behaviors were almost always noted in the HRAF description, and that the descriptions were uniformly positive.

Although they make a strong case for a link from ethics to the welfare of society via cooperation, Curry et al. (Curry et al., 2019) deliberately and explicitly fall into a trap that philosopher Allen Buchanan calls the Cooperation Dogma: the claim that morality is *nothing but* a mechanism for encouraging cooperation [(Buchanan, 2020), pp.12–14]. Such a strong claim invites falsification by examples of issues that are clearly moral, but that are not about cooperation. Critics of the Cooperation Dogma present a variety of phenomena, including disgust reactions, sexual practices, the treatment of dead bodies, and the treatment of cattle in India, to argue against the "*nothing but*" claim [(Curry et al., 2019), Comments].

Buchanan makes a more limited point:

> *"I cheerfully acknowledge that moralities originally were all about cooperation, and that moralities remain essential for successful cooperation today and always will be. I also heartily endorse the hypothesis that the basic features of human moral psychology, the moral mind, came about through natural selection because they contributed to cooperation and thereby to reproductive fitness. Nevertheless, I will argue that some moralities are more than a collection of solutions to cooperation problems." [(Buchanan, 2020), p.13, his emphasis]*

Buchanan's cheerful acknowledgment and hearty endorsement suggest that the role of trust might be part of a more nuanced understanding of the purpose of ethics.

3.2 Roles for Trust

My claim in this essay is that trustworthiness, and therefore properly earned trust, are key steps on the path from ethics to a thriving society via cooperation (**Figure 1**). With adequate trust, individuals can cooperate, producing (on average) outcomes with net positive gains for the society as a whole. When people can be trusted (most of the time) to follow social norms, then individuals can count on those social norms when they make their plans and act to achieve their goals.

Some norms (e.g., "*Keep your promises*") are obviously important for cooperation. Other norms (e.g., "*Drive on the correct side of the road*") are conventional, but if everyone can count on them, everyone's travel becomes safer and more efficient. Yet others (e.g., "*Wear business attire when doing this job*") are also conventional and seem to have little to do with cooperation, but signal membership in some group, providing evidence for trustworthiness.

Some moral principles (e.g., "*Care for elderly and disabled members of your community*" or "*Care for the dead bodies of your fallen comrades*") explicitly direct resources toward individuals who cannot contribute productively to the society. However, they are clearly grounded in trust, by the members of a community, that their community will continue to support them even when they are unable to contribute. That trust supports risky types of cooperation, for example, participation in dangerous hunts or warfare. Similarly, trust enables commitments that accept lifelong opportunity costs in order to benefit society, for example raising children, devotion to a religious vocation, or academic pursuit and conveyance of knowledge.

Trust also provides practical benefits for the computational complexity of reasoning about the effects of actions on the world. Philosopher Helen Nissenbaum (Nissenbaum, 2001) describes important insights about the function of trust from the social theorist Niklas Luhmann (Luhmann, 1979).

> *"Luhman characterizes trust as a mechanism that reduces complexity and enables people to cope with the high levels of uncertainty and complexity of contemporary life. Trust makes uncertainty and complexity tolerable because it enables us to focus on only a few possible alternatives. Humans, if faced with a full range of alternatives, if forced to acknowledge and calculate all possible outcomes of all possible decision nodes, would freeze in uncertainty and indecision. In this state, we might never be able to act in situations that call for action and decisiveness. In trusting, Luhmann says,*

'one engages in an action as though there were only certain possibilities in the future.' Trust also enables 'co-operative action and individual but coordinated action: trust, by the reduction of complexity, discloses possibilities for action which would have remained improbable and unattractive without trust—which would not, in other words, have been pursued.' According to this account, trust expands people's capacity to relate successfully to a world whose complexity, in reality, is far greater than we are capable of taking in." [(Nissenbaum, 2001), p.106 (footnotes omitted)]

The observations in this section support the structure described in **Figure 1** connecting ethics to trust to cooperation–both explicit cooperation with selected partners and implicit cooperation through social norms–leading to regularities that one can count on, and thus to a safer, more prosperous, and more secure society.

4 RELATED WORK ON EVOLUTION

The ethical principles of societies around the world, and across historical and pre-historical time, have much in common, but there are also dramatic differences. This pattern of diversity, changing over time, suggests the results of an evolutionary process. Since the ethics of a society consists of shared knowledge, that evolutionary process must operate at a cultural level, as well as (perhaps) at a biological level.

Figure 2 illustrates how biological evolution incrementally changes the distribution of genotypes in a population from one generation to the next. Over extended time, these incremental shifts can result in major qualitative changes. This pattern can be generalized to describe the accumulation and change of cultural knowledge, including ethics (Richerson and Boyd, 2005; Henrich, 2016).

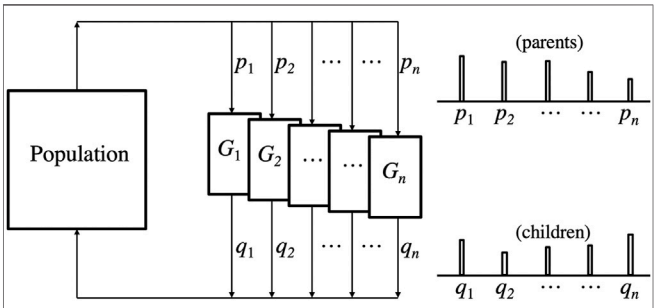

FIGURE 2 | A simple sketch of biological evolution. Consider a population of individuals who can be described as having certain *genotypes*, $G_1, \ldots, G_n$ with proportions $p_1, \ldots, p_n$. The genotype of each individual determines its *phenotype* (not shown) which determines the proportions $q_1, \ldots, q_n$, of the genotypes surviving into the population of the next generation. The two histograms (parents and children) illustrate distributional change from one generation to the next. Over many generations, new genotypes may become dominant, while others become rare or disappear entirely.

One selective pressure on the ethical beliefs of a society is the ability of that society to engage in cooperative activities that increase its resources and security. To accomplish this, a society must encourage its individual members to trust each other and the institutions of the society. This evolutionary process has a number of related aspects.

4.1 The Evolution of Shared Intentionality

An important cognitive skill is the ability to understand the behavior of oneself or others as *agents*; that is, in terms of *actions* taken in particular *situations* to pursue one's *goals*. Knowledge of any two of these provides some degree of information about the third.

Observing an agent's actions, predict its goals.

Knowing an agent's goals, predict its actions in a given situation.

Knowing an agent's goals and observing its actions, predict its beliefs about the current situation.

From an evolutionary perspective, it is obviously of great value for an agent to have the ability to predict the goals, beliefs, and actions of other agents, whether they are potential cooperative partners, enemies, or prey. Michael Tomasello calls this capability, shared by humans, great apes, and other animals, *individual intentionality* (Tomasello, 2019).

Based on data from similar tests administered to chimpanzees, orangutans, and human two-and-a-half-year-old children, Tomasello's group found strong similarities between great apes and human children in physical cognition (e.g., space, objects, and causality), and dramatic differences in social cognition and cooperation (Herrmann et al., 2007). Tomasello explains the extraordinary levels of cooperation seen in *Homo sapiens* in terms of two distinct levels of *shared intentionality*, rarely observed in non-human animals.

"In this view, humans' abilities to cooperate with one another take unique forms because individuals are able to create with one another a shared agent "we", operating with shared intentions, shared knowledge, and shared sociomoral values. The claim is that these abilities emerged first in human evolution between collaborative partners operating dyadically in acts of joint intentionality*, and then later among individuals as members of a cultural group in acts of* collective intentionality*." [(Tomasello, 2019), p.7, emphasis added]*

Tomasello argues that joint and collective intentionality are genetically encoded biological capabilities, acquired by the species through natural selection. *Joint intentionality* appeared about 400,000 years ago (in *Homo heidelbergensis*), driven by climate change, which made food harder to come by. Humans able to cooperate with partners, for example to capture larger animals, had a competitive advantage over those who could only seek food

as individuals. Likewise, the ability to cooperate pairwise in the raising of young children would be a selective advantage.

He argues that *collective intentionality* appeared around 100,000 years ago (in *Homo sapiens*), driven by increasing human population and increasing competition among human groups. Those capable of organizing into bands or tribes for collective support and defense would have an important advantage. Individuals in such groups who were incapable of learning and following the group's social norms would face exclusion and death.

Darwin, in *The Descent of Man* (Darwin, 1874), recognized this selective pressure.

> *"When two tribes of primeval man, living in the same country, came into competition, if (other circumstances being equal) the one tribe included a great number of courageous, sympathetic and faithful members, who were always ready to warn each other of danger, to aid and defend each other, this tribe would succeed better and conquer the other. . . .. A tribe rich in the above qualities would spread and be victorious over other tribes; but in turn overcome by some other tribe still more highly endowed."*

The "we" underlying joint intentionality is an abstract "agent" representing the shared intention, the shared understanding of the situation, and the roles in the shared activity. For example, a joint intention to hunt antelope might have roles for the chaser and the spearer. Each individual in the dyad has an obligation to the joint "we" to fill one of the roles, and the right to share in the rewards of the joint action.

In collective intentionality, the abstract agent "we" represents the entire community, and defines rights and obligations for members of the community. Some of these obligations are norms specifying the expected behavior of members in good standing of the society, including behaving in trustworthy ways when cooperating with others within the society. Other norms may define seemingly arbitrary behavioral regularities (e.g., of dress, food, and language, etc.) that signal membership in a specific society, allowing other members to distinguish "insiders" from "outsiders" even when the society is too large to recognize everyone individually. Social psychologists Jay Van Bavel and Dominic Packer (Van Bavel and Packer, 2021) describe the positive and negative impacts of these group-based identities on individuals and societies.

The abilities to reason about individual, joint, and collective intentionality are closely related to "Theory of Mind" in child development (Gopnik and Wellman, 1992; Wellman, 2014).

> *"Mirroring the phylogenetic sequence, this maturational process unfolds in two basic steps: first is the emergence of joint intentionality at around nine months of age, and second is the emergence of collective intentionality at around three years of age." [(Tomasello, 2019), p.8]*

Parents invest substantial effort in teaching these skills and social norms to their children, since survival may depend on them.

4.2 The Evolution of Cultural Knowledge

Biological evolution through natural selection of genes that enhance successful reproduction is a slow process. This is plausible for the genetic evolution of the biological (neural) capacity for joint and collective intentionality over several hundred thousand years. However, the last 10,000 years or so has seen dramatic changes in the structure of our civilization, in part due to changes in the nature and scope of cooperation (Wright, 2000). These rapid changes suggest a process of cultural evolution operating at a faster time-scale.

Richerson and Boyd (Richerson and Boyd, 2005) argue that cultural evolution is a distinct process within the framework of Darwinian evolution.

> *"Culture is information capable of affecting individuals' behavior that they acquire from other members of their species through teaching, imitation, and other forms of social transmission." [(Richerson and Boyd, 2005), p.5]*
>
> *"Some beliefs make people more likely to be imitated, because the people who hold those beliefs are more likely to survive or more likely to achieve social prominence. Such beliefs will tend to spread, while beliefs that lead to early death or social stigma will disappear." [(Richerson and Boyd, 2005), p.6]*
>
> *" . . .the human cultural system arose as an adaptation, because it can evolve fancy adaptations to changing environments rather more swiftly than is possible by genes alone. Culture would never have evolved unless it could do things that genes can't." [(Richerson and Boyd, 2005), p.7]*

It is important to recognize that cultural evolution is a kind of evolution by natural selection, but the analogy with biological evolution is not comprehensive. New variations are not generated through random mutations, but through inspiration or errors by individual humans. They are not selected purely through differential survival and reproduction, but by ease and accuracy of transmission of ideas from some human minds to others (Dawkins, 1976).

Joseph Henrich, in *The Secret of Our Success* (Henrich, 2016), sets out to explain the unique dominance of *homo sapiens* over the other species on our planet. Even before the beginning of recorded history, early humans had settled over a larger and more diverse geographical range than any other species. Henrich argues that this success is not due to our brain-power, but rather due to our cumulative culture.

> *"Probably over a million years ago, members of our evolutionary lineage began learning from each other in such a way that culture became cumulative. . . . After several generations, this process produced a sufficiently large and complex toolkit of practices and techniques that individuals, relying only on their own ingenuity and personal experience, could not get anywhere close to figuring out over their lifetime. . . . Once these useful skills and practices began to accumulate and improve over generations, natural selection had to favor individuals who were better cultural learners, who*

> *could more effectively tap into and use the ever-expanding body of adaptive information available." [(Henrich, 2016), p.3]*

Cumulative cultural knowledge includes technological knowledge like the "know-how" to create arrows or kayaks or compasses, and institutional knowledge like the structure of corporations, churches, and governments. Cultural evolution allows the incremental accumulation of sophisticated designs that could not have been created by any individual during a single lifetime.

In spite of the differences in typical time-scales of biological and cultural evolution, Henrich (Henrich, 2016) provides persuasive examples of gene-culture co-evolution. For example, the cultural acquisition of how-to knowledge about cooking has influenced the biological evolution of the digestive tract in *homo sapiens*. Another example describes how cultural adaptations in tracking and water storage set the context for biological adaptations that have made humans into pre-eminent long-distance runners, able to capture much faster prey by pursuing them to exhaustion.

Some accumulated cultural information is highly adaptive, like the technologies that have allowed humans to inhabit a wider range of environments than any other species on Earth. Others eventually die out, like human sacrifice among the Aztec and Inca, or universal celibacy among the Shakers. The social and individual costs of some cultural beliefs eventually lead to their extinction.

Culture, then, is an evolved adaptation that fills a critical gap in scope and time-scale between biological evolution and individual learning and problem-solving (Richerson and Boyd, 2005). The biological evolution of *Homo sapiens* included the cognitive capacity for shared intentionality (Tomasello, 2019), and social emotions such as shame, guilt, and loyalty (Haidt, 2012).

4.3 The Evolution of Social Structures

In *Non-zero: The Logic of Human Destiny* (Wright, 2000), Robert Wright argues that there is a clear direction of progress in human history, visible in the increasing scale of social structures and technologies for supporting cooperation. An organizing theme is the creation of non-zero-sum (i.e., win-win) interactions that result in increasing resources for the society as a whole.

Early humans lived in small egalitarian bands of individuals who cooperated with each other to obtain food through hunting and gathering, and cooperated to protect the band from threats. As the size of human groups increased, egalitarian bands grew into tribes. The successful leader of a tribe, sometimes called a Big Man, was able to accumulate capital and organize the division of labor necessary for building larger-scale technologies such as whale boats and large rabbit nets. Organized hunts using these technologies could bring in much greater resources for the tribe than would be possible even for a very cooperative egalitarian band.

The capture of a whale or many rabbits gives the group a larger supply of perishable meat than it can consume, and therefore an opportunity for trade with other groups–the paradigm win-win interaction. Surplus meat is much more valuable to hungry neighbors who have not had a successful hunt and, in the presence of sufficient trust, can be traded for a commitment to share when circumstances are reversed. Sharing a surplus increases the tribe's status at relatively low cost, while helping to protect it from future uncertainties. The ability to establish trustworthiness and to recognize and use these forms of cooperation is a selective advantage for a group, which enhances the survival and reproductive opportunities of its individual members.

With new technologies such as agriculture, and increasing scale spanning multiple settlements, tribes grew into chiefdoms. Continued growth, supporting and supported by information technologies such as writing, money, law, and markets, leads to state-level organization: "civilization." The common link between these information technologies and societal growth is trust. Writing increases trust in promises. Money provides portable, trustworthy value. Published law allows people to trust in the reliability of rules for acceptable behavior. Markets allow trade between people who are willing to trust each other without knowing each other personally. Access to the market motivates people to follow its norms and to punish those who refuse to do so. Increasing scope and benefits of cooperation are supported by political and organizational developments such as democracy, and technological developments such as the industrial revolution(s), the computing revolution, and the Internet.

4.4 Taking Stock

Sections 2–4 are intended to support the claim that the human species, consisting of individuals and their societies, is the result of biological and cultural evolution. Biological (genetic) evolution takes place through individual reproductive success. However, individual reproductive success, especially as societies become more complex, depends on the success of the society in accumulating resources including various forms of cultural knowledge.

Cooperation is a large family of mechanisms whereby a society can accumulate more resources. Trust is a relation that is generally necessary for cooperation, both among groups of prospective cooperative partners, and across the entire society in the case of respect for social norms.

Among other roles in human life, the ethics of a society instructs individuals in what it means to be trustworthy, both in one's own decisions and in recognizing whether others are worthy of trust. Thus ethics (among other things) encourages trust, which encourages cooperation, which helps the society thrive, which helps its individual members thrive, including in terms of individual reproductive success.

My argument in the first half of this essay is that this causal chain contributes to humanity's success, even up to our very complex modern society. However, the second half of this essay argues that certain links in the chain are vulnerable, and could lead to existential threats.

5 TRUST AND VULNERABILITY

As we have seen, the definition of trust involves vulnerability among individuals: "*Trust is a psychological state comprising the intention to accept vulnerability based on positive expectations of the intentions or behavior of another.*" [(Rousseau et al., 1998), p.1998].

The vulnerability that individuals accept is vulnerability to cooperative partners (trusting that partners will respect and protect each others' vulnerabilities, resulting in greater benefits for everyone), and the vulnerability of following social norms (incurring opportunity costs, in confidence that others will do the same, resulting in regularities that make planning easier and reduce the need for defense and repair, for everyone).

In both of these cases, accepting vulnerability by trusting others can result (if the others are trustworthy) in a significantly better outcome than actively defending the vulnerability against exploitation. We can therefore consider trust and cooperation to represent "non-obvious self-interest", obtaining payoffs from cooperation though prudent acceptance of vulnerability to trustworthy partners.

As described in **Figure 1**, trust plays a central role in many cooperative processes, ranging from pairs, to larger groups of partners, to the entire society (for social norms). These processes generate the resources that help a society thrive by defending against threats, taking advantage of opportunities, and generally providing benefits for its individual members.

Loss of trust decreases willingness to cooperate and confidence in social norms, resulting in scarcity of resources, making it difficult for the society to plan, and to meet threats or exploit opportunities. Given the centrality of trust in **Figure 1**, if trust is eroded, society is threatened.

The larger and more complex societies get, the more they rely on trust–of individuals, of institutions, and of social norms (Luhmann, 1979; Wright, 2000). Erosion of trust and loss of resources can bring a successful, complex society to the point of collapse (Tainter, 1988; Diamond, 2005). For our own society, climate change poses an existential threat. Meeting that threat will require serious amounts of trust and cooperation, at a time when trust is being eroded.

6 REASONING WITH MODELS

Before returning to the problem of existential threats, we need to consider how we make predictions and action decisions in a world that is essentially infinitely complex. Neither we humans, nor any conceivable computing device, can reason with the full complexity of the physical and social world we inhabit.

Instead, we (ordinary people using common sense as well as scientists and engineers) reason and make decisions using *models* that identify a limited set of relevant factors. We treat all other factors as *negligible*. When the relevant factors are well chosen, a simplified model can efficiently draw conclusions, making predictions, plans, and action decisions that are adequate for the purpose of the model.

The big question of model-building is which few aspects of the unbounded complexity of the world should be explicitly included in the model, omitting everything else. For inference to be feasible, a model must have a small number of elements (variables and constraints in the case of a numerical, algebraic, or differential equation model; constants, variables, relations, and sentences in case of a logical theory; other elements for other types of models). Everything else is left out.

A model makes explicit a relatively small set of "*known unknowns*"—the elements that are relevant to its predictions. The values of some of these known unknowns must be found and provided as inputs; others are derived by inference within the model. The many other aspects of the world not explicitly described in the model are the "*unknown unknowns.*"[3] For a well-constructed model, omitting the unknown unknowns simply makes the model more efficient.

6.1 Deciding What to Do: Game Theory

How do we decide what to do in complex situations with multiple motivated decision-makers and uncertain outcomes? Inspired by recreational games, *game theory* is a powerful framework for creating simple models of these complex situations and interactions (Leyton-Brown and Shoham, 2008; von Neumann and Morgenstern, 1953).[4] The core idea behind game theory is that each player selects the action that maximizes his own *expected utility*, recognizing that the other players are doing the same. In their seminal book defining game theory [(von Neumann and Morgenstern, 1953), sect.3], von Neumann and Morgenstern show that for any consistent set of preferences that an agent might have over states in the state space, there is a real-valued utility function such that the ordering of its values expresses the agent's preferences. Unfortunately, we do not have a guarantee that this function is the same as the one provided in the problem statement.

In game theory, action selection by utility maximization is defined as "*rational.*" As in economics and other disciplines, the leading textbook in Artificial Intelligence states that "a rational agent should choose the action that maximizes the agent's expected utility" [(Russell and Norvig, 2010), p.611].

With a good model, including an appropriate utility measure, game theory can find optimal strategies responding to complex situations, including the optimal choices of other players. Game theory can be effective in real-world circumstances where the stakes and the relationships among the participants are clear–for example in economic interactions such as auctions.

For many decision problems, the game theory models–state and action spaces, transition probabilities, and utility measures–seem to be clear and straight-forward translations of the problem statement. Applying the power of game theory seems to be a

[3]A model *builder* may knowingly and deliberately omit an aspect of the world from a model, but that aspect is then unknown to the model itself, and very likely to the model *user*, especially if that model user is an artificially intelligent creature. Inferences with that model treat the missing aspects as invisible–they are "unknown unknowns."

[4]The *state space* describes the situations the game can be in, such as the board position in chess and whose move it is; or in baseball, the scores, team at bat, balls and strikes, and runners on bases. Each state has a set of possible *actions*, and a *probability distribution* over the possible results of taking each action. Each player has a *utility value* for each state in the state space, which represents that player's preference for that state of the game. A *discount factor* decreases the present value of future rewards exponentially with time. The *expected utility* of an action taken in a state is the probability-weighted average of the discounted utilities of all possible future states following from this action.

matter of plugging in the relevant values, computing expected values, and identifying the maximum. Is this correct?

Unfortunately, bad models lead to bad results. The Prisoner's Dilemma (**Figure 3**) is famous because, using the straightforward model, utility maximization gives a poor outcome. A number of other laboratory-scale games provide closely related results, including the Public Goods Game (Rand and Nowak, 2013), the Ultimatum Game (Thaler, 1988), and the Tragedy of the Commons (Hardin, 1968).

Generalizing the Trolley Problem, a Deadly Dilemma (**Figure 4**) occurs when an agent is faced with two deadly alternatives. The Moral Machine online survey experiment (Awad et al., 2018) probes the nature of the utility function by which the agent selects the lesser of the two evils. Human participants are shown simulated scenarios where several passengers in an autonomous vehicle are speeding toward several pedestrians on a narrow street. Its only options are to hit the pedestrians, killing all of them, or to crash into a barrier, killing all the passengers. Participants are given demographic features of the potential victims and are asked which choice the AV should make. Assuming that participants are maximizing expected utility, the researchers infer the utilities they assign to those demographic features.

7 THE DANGERS OF BAD MODELS

A good model provides a simplified description of the complex world that can be used efficiently to accomplish the purpose of the model. On the other hand, a bad model can make seriously wrong predictions with unwarranted confidence, failing to predict genuine threats or overlooking genuine opportunities. Particularly dangerous cases occur when the model's predictions are mostly correct, earning the user's confidence, but the model is blind to unusual situations where its predictions diverge strongly from reality.

7.1 The Problem of Unknown Unknowns

An important failure mode for a model is to omit a factor that proves to be important. This is the infamous "unknown unknown"—a factor missing from the model whose absence is not even suspected, but that leads to an importantly incorrect prediction.

You and your partner in crime have been captured, separated, and each is offered this deal: "*If you testify against your partner, you will go free, and your partner goes to jail for 4 years. If neither of you testifies, you each go to jail for 1 year, but if you both testify, you both get 3 years.*"

Cooperating with your partner (action *C*) means refusing to testify. Defecting (action *D*) means to testify against your partner. The entries in this array are the utility values for (you, partner), and they reflect individual rewards (years in jail).

	C	D
C	$-1,-1$	$-4,0$
D	$0,-4$	$-3,-3$

(1)

No matter which choice your partner makes, you are better off choosing action *D*. The same applies to your partner, so the Nash equilibrium (the "rational" choice of action) is (*D*, *D*), which is collectively the worst of the four options. To attain the much better cooperative outcome (*C*, *C*) by choosing *C*, you must trust that your partner will also choose *C*, accepting your vulnerability to your partner choosing *D*.

This can be due to modeling error: without thinking about it, the model-builder omits a factor that turns out to be important. For example, in a model of health-care services, an insurance company used cost of treatment as a proxy for severity of disease, failing to recognize that the training data reflected historical racial biases, where minority patients received less (and less costly) treatment for a given severity of disease. For the same clinical evidence, the resulting model categorized diseases as less severe in minority patients than in majority patients (Obermeyer et al., 2019).

You and your partner in crime have been captured, separated, and each is offered this deal: "*If you testify against your partner, you will go free, and your partner goes to jail for four years. If neither of you testifies, you each go to jail for one year, but if you both testify, you both get three years.*" Cooperating with your partner (action C) means refusing to testify. Defecting (action D) means means to testify against your partner. The entries in this array are the utility values for (you, partner), and they reflect individual rewards (years in jail).

	C	D
C	$-1,-1$	$-4,0$
D	$0,-4$	$-3,-3$

(1)

No matter which choice your partner makes, you are better off choosing action D. The same applies to your partner, so the Nash equilibrium (the "rational" choice of action) is (D, D), which is collectively the worst of the four options. To attain the much better cooperative outcome (C, C) by choosing C, you must trust that your partner will also choose C, accepting your vulnerability to your partner choosing D.

FIGURE 3 | The Prisoner's dilemma (Axelrod, 1984).

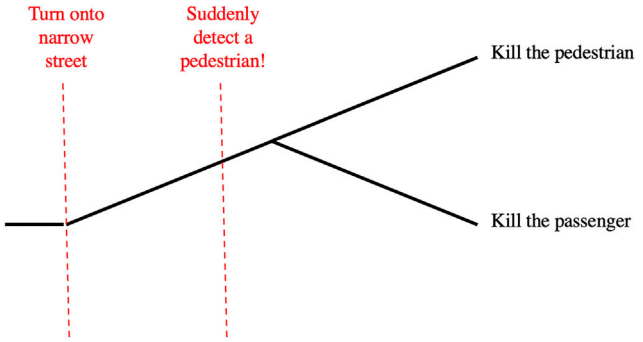

FIGURE 4 | The Deadly Dilemma (abstracted from Awad et al., 2018).

This failure mode can also arise when a model that works well in one regime is applied outside that regime, where a simplifying assumption is no longer valid. For example, the effect of air resistance is negligible in a model to predict the result of jumping from my garage roof. But if I consider jumping from a flying airplane, the model *must* include air resistance, or it will be unable to predict the benefit of a parachute.

7.2 A Bad Model of the Prisoner's Dilemma

As the Prisoner's Dilemma (**Figure 3**) is presented, the translation from story problem to game theory model seems straight-forward. The choice of action is obvious: Cooperate or Defect. The utility measure is obvious: number of years in prison. Utility maximization clearly shows that Defect is the best choice for each player, no matter what the other player chooses. Shockingly, the outcome (*D*, *D*) from this choice is the *worst* collective result.

More sophisticated games show that this problem generalizes to larger numbers of players (the Public Goods Game (Rand and Nowak, 2013)) and management of limited resources (the Tragedy of the Commons (Hardin, 1968)).

The far better cooperative result (*C*, *C*) is available if each player trusts the other, accepting vulnerability to the other's defection. However, game theory assumes that each player chooses actions to maximize its own utility measure (as in recreational games). And trust and trustworthiness are unknown, with no role in the utility measure for this model.

If we change the model, adding a component to the utility measure that reflects the player's demonstrated trustworthiness (say, +1 for *C*, −1 for *D*), then the payoff matrix (1) changes

	C	D
C	$-1,-1$	$-4,0$
D	$0,-4$	$-3,-3$

$$\Rightarrow$$

	C	D
C	$0,0$	$-3,-1$
D	$-1,-3$	$-4,-4$

(2)

and the best choice for each player is *C*, regardless of the other player's choice, so utility maximization within this improved model gives the optimal outcome (*C*, *C*).

A reader might argue that the updated payoff matrix (2) no longer represents the Prisoner's Dilemma, but this is exactly the point. When the payoff from the utility-maximizing choice is obviously worse than available alternatives, then the model of the decision is likely to be wrong. An unknown unknown (in this case trustworthiness) has been omitted from the model. Changing the model improves the outcome.

Much effort has gone into designing models like the Repeated Prisoner's Dilemma (Axelrod, 1984), aiming to explain how maximizing the expected utility of the future stream of rewards can make the cooperative choice (*C*, *C*) into the optimum. Unfortunately, these efforts have proved to be fragile, for example giving different results for finite and infinite sequences of games, and depending for tractability on repeating the same game. Getting robust decisions for cooperation seems to require trustworthiness to be an explicit element of the model, included in the utility function.

Some critics argue that a game theory model should include only "objective" utilities such as money, mortality, or jail time, rather than "character" or reputation attributes such as trustworthiness. However, note that in a game like poker, expertise clearly includes the ability to estimate an opponent's character, such as willingness and ability to bluff. A game theory model whose utility measure is based only on the expected values of given hands of cards will play poor poker. The need to estimate trustworthiness in potential cooperative partners is analogous.

In general, an overly simple utility measure will treat important concerns as negligible, leading to a bad outcome.

7.3 A Bad Model of the Deadly Dilemma

Likewise, as the Deadly Dilemma (**Figure 4**) is described, the choice of action is obvious: kill the pedestrians or kill the passengers. The utility measure also seems obvious, especially in the light of the demographic information provided: quality and quantity of lost life (not just number of deaths). As presented, the scenario begins when the autonomous vehicle first senses the pedestrians in its path, and recognizes that its speed and the constrained environment requires it to choose to kill the pedestrians or to kill the passengers. Both alternatives are terrible, so the decision-maker must select the lesser of two evils.[5]

In our society, however, beginning drivers are taught situational awareness: continually monitoring the environment and evaluating whether their speed allows them to respond appropriately to sudden developments.[6] A better model for this decision would include the "upstream decision point" where the environment changes (e.g., the road narrows and loses shoulders), making the vehicle's speed excessive. At that point, the utility-maximizing decision is to slow down to preserve the ability to make a safe emergency stop in the future, in case a hazard is detected (See **Figure 5**).

Suddenly facing a deadly dilemma, an individual driver (human or AV) cannot go back in time to the upstream decision point. But

[5]The Moral Machine experiment (Awad et al., 2018) required participants to make this choice in several different scenarios varying on demographic features for the passengers and pedestrians including number, age, gender, social class, criminality, and species. The researchers used the participants' choices to determine the utilities assigned to those demographic features as a function of the participants' demographic features including geographical region. The methodological validity and larger significance of this experiment are quite controversial.

[6]A well-known question on written driving tests asks students to choose how to react to a ball bouncing into the street in the path of their car. The correct answer is to anticipate that a child could be following the ball, so the driver should slow or stop.

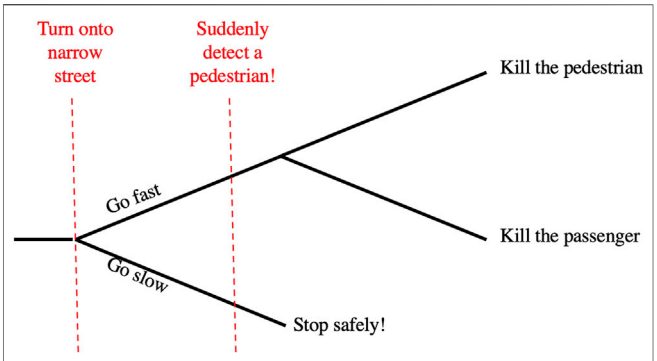

FIGURE 5 | Identify an upstream decision point to avoid the Deadly Dilemma.

an educator or an algorithm designer has the responsibility to anticipate such problems, and ensure that the driver has the situational awareness to detect the upstream decision point and make the choice that avoids the deadly dilemma.

Some readers may argue that a Deadly Dilemma is possible, no matter how unlikely, so an autonomous vehicle should be programmed to make the "right choice" if that should happen. At the point when such a tragic dilemma appears, there is no good option; there is only the lesser of two serious evils. Only in the larger model including the upstream decision point is there an opportunity to make a choice resulting in a good utility value. Therefore, the "moral" choice for the design of an autonomous vehicle is to be prepared for a Deadly Dilemma, use the larger model, recognize the upstream decision point, and choose the option that avoids both evils.

7.4 Bad Models Can Target the Vulnerability of Trust

In the Prisoner's Dilemma, the desirable payoff of the cooperative solution depends on each player trusting the other: accepting vulnerability to defection, confident in the other's choice to cooperate. In the original model (1), with no utility for trustworthiness, each player is tempted by the even higher payoff from defecting on a trusting partner. However, the symmetry of the game means that the tempting outcome is lost, and both "rational" players do poorly.

When a player's trust is violated, the victim's trust for the exploiter is lost, and can only be restored slowly, if at all. Even worse, a reputation for being untrustworthy means that the exploiter will be offered fewer opportunities for cooperation in the future. These are among the reasons why ethical and trustworthy behavior can be considered "non-obvious self-interest."

Moving beyond the individual to the society, a widespread belief or custom that encourages exploitation results in widespread loss of trust, discouraging cooperation, leading to a tragedy of the commons (Hardin, 1968). This concern has become mainstream, illustrated by a recent discussion in *CACM*, the flagship journal of the Computer Science professional association, of the need for regulating false and polarizing posts on social media platforms.

> *"Yet moral hazard may not be a strong enough term to describe what could happen. . . . another motivation for platform businesses to self-regulate more aggressively is the potential for a "tragedy of the commons." This phrase refers to a situation where individuals or organizations narrowly pursue their own self-interest, as with moral hazard, but in the process deplete an essential common resource that enabled their prosperity to begin with. Think of the native on Easter Island who cut down the last tree from a once-bountiful forest to make a fire—and then left everyone with an island that had no more trees. With online platforms, we can view the essential common resource as user trust in a relatively open Internet that has become a global foundation for digital commerce and information exchange." [(Cusumano, 2021), p.17]*

The erosion of trust can quite possibly lead, not just to economic loss for the exploiters, but to an existential threat to the society as a whole.

8 EXISTENTIAL THREATS TO HUMAN SOCIETY?

Our society has grown enormously in size, wealth, complexity, and quality of life over centuries (Pinker, 2018) and millenia (Wright, 2000), due in part to our ability as humans to trust and cooperate with each other, producing net gains in resources for the society as a whole. However, growth and prosperity are not inevitable. Indeed, a number of complex, thriving societies have gone on to collapse due to factors such as overpopulation and ecological disaster (Tainter, 1988; Diamond, 2005).[7]

Decreasing resources can make it more difficult for the society to respond to threats or to take advantage of opportunities. Challenges that were manageable in the past might become insurmountable.

The high-level description in **Figure 1** of the roles of ethics, trust, and cooperation in generating society's resources suggests that trust could be a critical point of vulnerability for a society. A general societal failure of trust could decrease effective cooperation, decreasing available resources.

Are there potential existential threats to our society? Yes, several.

8.1 Superintelligent AI

There are concerns about the possibility of an "intelligence explosion" leading to the emergence of an uncontrollable super-intelligent AI that could be an existential threat to humanity. The intelligence explosion was initially proposed by mathematician I. J. Good in 1965 (Good et al., 1965), and explored by computer scientist Vernor Vinge in 1993 (Vinge, 1993), philosopher Nick Bostrom in 2014 (Bostrom, 2014), and computer scientist Stuart Russell in 2019 (Russell, 2019), among many others. Since artificial

[7]There is controversy about sudden societal "collapse" (McAnany and Yoffee, 2010), but general recognition that societies rise and fall, thrive and deteriorate at various points in their history.

intelligence today is a product of human intelligence, attaining human-level AI could enable an exponentially self-improving process, possibly resulting in an artificial entity with super-human powers, incomprehensible and uncontrollable by mere humans. Humanity could be eliminated deliberately or by accident. Compelling analogies are presented to the slow rise and sudden take-off of exponential growth curves. Less attention is paid to competing analogies with equally fundamental mathematical phenomena such as the damping effects of resource constraints, and limits to prediction due to sensitive dependence on initial conditions.

Both fictional and non-fictional explorations of this scenario suggest that the existential threat is not actually "super-intelligence" but rather "super-power." That is, the existential threat follows from putting an AI system with decidedly sub-human levels of intelligence in control of a source of power that poses an existential threat, such as nuclear weapons.

8.2 Oversimplified Capitalism

In its abstract ideal form, capitalism is a powerful form of societal cooperation, harnessing feedback cycles among many production and consumption decisions to allocate investment, produce wealth, and distribute that wealth among stakeholding members of society.

The original insight was that, *under appropriate conditions*, a successful economic system need not depend on central coordination to maximize everyone's utility. As Adam Smith wrote in 1776:

> *"It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest. We address ourselves, not to their humanity but to their self-love, and never talk to them of our necessities but of their advantages." (Smith, 1776)*

Those appropriate conditions include market-based competition among many buyers and sellers, both concerned with both quality and price, and all participants being small relative to the size of the market. When this simplified model is appropriate, negative feedback from producer and consumer choices drives the system as a whole toward equilibrium states that satisfy certain optimality critera.

One failure mode for this model occurs when a seller (e.g., of a necessary product) or a buyer (e.g., an employer buying work) dominates their part of the market, to the point where negative feedback can no longer compel them to change their ways. This can easily result in high monopoly prices and low captive-worker wages. Marketplace rules are intended to prevent these possibilities, but a sufficiently powerful player may find it more profitable to manipulate the rules than to improve their offering in the marketplace.

In an influential 1970 article (Friedman, 1970) titled "*The social responsibility of business is to increase its profits*", the economist Milton Friedman argued that firms in a marketplace should focus purely on profits, without concern for other societal factors. Even when concern for the local community is in the firm's long-term best interest, Friedman criticized action on that concern as "hypocrisy." The fictional character Gordon Gecko in the 1987 movie *Wall Street* expressed Friedman's position with his famous line, "*Greed . . . is good!*" The reader should be reminded of the oversimplified Prisoner's Dilemma model (1) with a utility measure sensitive to years in jail but not to trustworthiness, leading to poor decisions with bad outcomes.

These ideas, treating non-financial aspects of the economy (e.g., trust) as negligible, spread from economics and business to the culture generally. Former President Ronald Reagan's 1986 quote, "*The nine most terrifying words in the English language are: 'I'm from the government, and I'm here to help*'," is an explicit attack on the trustworthiness of government (Andersen, 2020).

Well-regulated capitalism is a valuable tool for cooperative enterprise in society. But explicitly discouraging trust also discourages cooperation, reducing resources and threatening the long-term viability of the society.

8.3 Climate Change

Climate change is an existential threat to human society, and possibly even to the human species. We've passed the "upstream decision point" where a genuine solution might have been possible, but mitigating the destructive impact of climate change will require substantial cooperation among individuals and nations. That cooperation will require trust, which involves vulnerability. Given the global set of actors involved, it is safe to assume that vulnerability will be exploited in some cases. To avoid catastrophe, we will need resources, including trust and cooperation. Can we do it? Nobody knows (Robinson, 2020; Gates, 2021; Kolbert, 2021).[8]

9 CONCLUSION

As an AI researcher, I am concerned about the potential impact of artificially intelligent systems on humanity. The focus of my research has been on understanding the structure of knowledge in commonsense foundational domains (space, dynamical change, objects, actions, and now, ethics), including how this knowledge is created, how it is learned, and how it might be applied to solve tangible problems facing intelligent agents in a complex world.

In the first half of this essay, I present an argument, based on work by Tomasello (Tomasello, 2019), Richerson and Boyd (Richerson and Boyd, 2005), Henrich (Henrich, 2016), Curry (Curry et al., 2016; Curry et al., 2019), Buchanan (Buchanan, 2020), and others, that ethics is an evolved body of cultural knowledge that serves to encourage

[8]Humanity has already faced the existential threat of nuclear weapons, capable of destroying our civilization and possibly our species. Somehow, so far, we have found ways to trust and cooperate well enough to keep this existential threat at bay.

individual behavior that promotes the welfare of the society (which in turn promotes the welfare of its individual members). A high-level (and partial) representation of the causal paths involved (**Figure 1**) suggests that *trust* plays a key role in this process.

In the second half of the essay, I consider whether that key role could be a bottleneck, even a vulnerability, exposing the society to existential threats. This possibility depends on the fact that we (humans, AIs, corporations, and governments) necessarily rely on simplifying models to cope with the unbounded complexity of our physical and social world. Well-formulated models are essential tools. But when important unknown unknowns are omitted, poorly-formulated models can draw dangerously wrong conclusions.

By selecting actions to maximize a utility measure, a well-formuilated game theory model can be a powerful and valuable tool. However, a poorly-formulated game theory model may be uniquely harmful, in cases where the action it recommends deliberately exploits the vulnerability and violates the trust of cooperative partners. Widespread use of such models can erode the overall levels of trust in the society. Cooperation is reduced, resources are constrained, and there is less ability to meet challenges or take advantage of opportunities.

We are experiencing a variety of social, economic, and political forces that promote models that erode trust in our society and its institutions and could result in resource limitations. At the same time, humanity is facing the existential threat of climate change, which will require material resources, as well as trust and cooperation.

This argument about the critical importance of trust is not only relevant to robots and other AI systems, important though they may be. Like robots and AIs, corporate and governmental systems make action decisions based on formal representations of simplified models. Human commonsense inference is also subject to errors due to incorrectly simplified models, but most humans have the capability of detecting and correcting model failures, a capability seldom implemented in AI systems.

DATA AVAILABILITY STATEMENT

The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

BK conceived the topic, did the scholarly research, structured the argument, and wrote the paper.

REFERENCES

Alexander, L, and Moore, M. (2015). "Deontological Ethics," in *The Stanford Encyclopedia of Philosophy. Spring 2015 Edition*. Editor E. N. Zalta. http://plato.stanford.edu/archives/spr2015/entries/ethics-deontological/.

Alsamhi, S., Ma, O., Ansari, M., and Gupta, S. (2019). Collaboration of Drone and Internet of Public Safety Things in Smart Cities: An Overview of QoS and Network Performance Optimization. *Drones* 3 (1), 13. doi:10.3390/drones3010013

Andersen, K. (2020). *Evil Geniuses: The Unmaking of America, A Recent History*. New York, NY, USA: Random House.

Anderson, M., and Anderson, S. L. (2006). Guest Editors' Introduction: Machine Ethics. *IEEE Intell. Syst.* 21 (4), 10–11. doi:10.1109/mis.2006.70

Aristotle (1999). *Nicomachean Ethics*. second edition. Indianapolis, IN: Hackett. 1999/349 BCE. Terence Irwin, translator.

Ashford, E., and Mulgan, T. (2018). "Contractualism," in *The Stanford Encyclopedia of Philosophy. Summer 2018 Edition*. Editor E. N. Zalta.

Asimov, I. (1952). *I, Robot*. New York, NY, USA: Grosset & Dunlap.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., et al. (2018). The Moral Machine experiment. *Nature* 563, 59–64. doi:10.1038/s41586-018-0637-6

Axelrod, R. (1984). *The Evolution of Cooperation*. New York, NY, USA: Basic Books.

Beauchamp, T. L., and Childress, J. F. (2009). *Principles of Biomedical Ethics*. sixth edition. Oxford, UK: Oxford University Press.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford, UK: Oxford University Press.

Buchanan, A. (2020). *Our Moral Fate: Evolution and the Escape from Tribalism*. Cambridge, MA USA: MIT Press.

Buchanan, A., and Powell, R. (2018). *The Evolution of Moral Progress: A Biocultural Theory*. Oxford, UK: Oxford University Press.

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., and Floridi, L. (2017). Artificial Intelligence and the 'Good Society': the US, EU, and UK Approach. *Sci. Eng. Ethics* 24 (2), 505–528. doi:10.1007/s11948-017-9901-7

Cerf, V. G., and Kahn, R. E. (1974). A Protocol for Packet Network Interconnection. *IEEE Trans. Comm. Tech.* 22 (5), 627–641. doi:10.1109/tcom.1974.1092259

Christakis, N. A. (2019). *Blueprint: The Evolutionary Origins of a Good Society*. New York City: Little, Brown Spark.

Christian, B. (2020). *The Alignment Problem*. New York, NY, USA: W. W. Norton.

Cook, K. S., Hardin, R., and Levi, M. (2005). *Cooperation without Trust?* New York, NY, USA: Russell Sage Foundation.

Cudd, A., and Eftekhari, S. (2018). "Contractarianism," in *The Stanford Encyclopedia of Philosophy. Summer 2018 Edition*. Editor E. N. Zalta.

Curry, O. S. (2016). "Morality as Cooperation: A Problem-Centred Approach," in *The Evolution of Morality*. Editors T. K. Shackelford and R. D. Hansen (Berlin, Germany: Springer), 27–51. doi:10.1007/978-3-319-19671-8_2

Curry, O. S., Mullins, D. A., and Whitehouse, H. (2019). Is it Good to Cooperate? Testing the Theory of Morality-As-Cooperation in 60 Societies. *Curr. Anthropol.* 60 (1), 47–69. doi:10.1086/701478

Cusumano, M. A. (2021). Section 230 and a Tragedy of the Commons. *CACM* 64 (10), 16–18.

Darwin, C. (1874). *The Descent of Man and Selection in Relation to Sex*. 2nd edition. New York: American Home Library.

Dawkins, R. (1976). *The Selfish Gene*. Oxford, UK: Oxford University Press.

Diamond, J. (2005). *Collapse: How Societies Choose to Fail or Succeed*. New York City: Viking.

Driver, J. (2014). "The History of Utilitarianism," in *The Stanford Encyclopedia of Philosophy. Winter 2014 Edition*. Editor E. N. Zalta.

Posner, E. A. (Editor) (2007). *Social Norms, Nonlegal Sanctions, and the Law* (Cheltenham, UK: Edward Elgar Publishing).

Fedyk, M. (2017). *The Social Turn in Moral Psychology*. Cambridge, MA USA: MIT Press.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., et al. (2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds & Machines* 28, 689–707. doi:10.1007/s11023-018-9482-5

Forbus, K. D., Hinrichs, T., and Rabkina, I. (2018). "Building Analogy Systems: Some Lessons Learned," in *Advances in Cognitive Systems*.

Friedman, B., Harbers, M., Hendry, D. G., van den Hoven, J., Jonker, C., and Logler, N. (2021). Eight Grand Challenges for Value Sensitive Design from the 2016 Lorenz Workshop. *Ethics Inf. Technol.* 23, 5–16. doi:10.1007/s10676-021-09586-y

Friedman, B., Kahn, P. H., Jr., Borning, A., and Huldtgren, A. (2013). "Value Sensitive Design and Information Systems," in *Early Engagement and New Technologies: Opening up the Laboratory* (Berlin, Germany: Springer), 55–95. doi:10.1007/978-94-007-7844-3_4

Friedman, B., Khan, P. H., Jr., and Howe, D. C. (2000). Trust Online. *Commun. ACM* 43 (12), 34–40. doi:10.1145/355112.355120

Friedman, M. (1970). *The Social Responsibility of Business Is to Increase its Profits*. New York City: The New York Times Magazine.

Fukuyama, F. (1995). *Trust: The Social Virtues and the Creation of Prosperity*. New York, NY, USA: Simon & Schuster.

Gates, B. (2021). *How to Avoid a Climate Disaster*. New York, NY, USA: Alfred A. Knopf.

Good, I. J. (1965). "Speculations Concerning the First Ultraintelligent Machine," in *Advances in Computers*. Editors F. L. Alt and M. Rubinoff (Cambridge, MA USA: Academic Press), 6, 31–88.

Gopnik, A., and Wellman, H. M. (1992). Why the Child's Theory of Mind Really *Is* a Theory. *Mind Lang.* 7 (1), 145–171. doi:10.1111/j.1468-0017.1992.tb00202.x

Haidt, J. (2012). *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. New York, NY, USA: Vintage Books.

Hardin, G. (1968). The Tragedy of the Commons. *Science* 162, 1243–1248. doi:10.1126/science.162.3859.1243

Henrich, J. (2016). *The Secret of Our Success*. Princeton, NJ, USA: Princeton University Press.

Herrmann, E., Call, J., Hernàndez-Lloreda, M. V., Hare, B., and Tomasello, M. (2007). Humans Have Evolved Specialized Skills of Social Cognition: The Cultural Intelligence Hypothesis. *Science* 317, 1360–1366. doi:10.1126/science.1146282

High Level Expert Group on AI (2019). Ethics Guidelines for Trustworthy AI. Technical Report, European Commission. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

High Level Expert Group on AI (2019). Policy and Investment Recommendations for Trustworthy Ai. Technical Report, European Commission. Available at: https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence.

High Level Expert Group on AI (2020). Sectoral Considerations on the Policy and Investment Recommendations for Trustworthy Artificial Intelligence. Technical Report, European Commission. Available at: https://futurium.ec.europa.eu/en/european-ai-alliance/document/ai-hleg-sectoral-considerations-policy-and-investment-recommendations-trustworthy-ai.

High Level Expert Group on AI (2020). The Assessment List for Trustworthy Artificial Intelligence (ALTAI). Technical Report, European Commission. Available at: https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai.

Hursthouse, R. (2013). "Virtue Ethics," in *The Stanford Encyclopedia of Philosophy. Fall 2013 Edition*. Editor E. N. Zalta.

Kearns, M., and Roth, A. (2020). *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford, UK: Oxford University Press.

Kolbert, E. (2021). *Under a White Sky*. New York City: Crown.

Kuipers, B. (2018). How Can We Trust a Robot? *Commun. ACM* 61 (3), 86–95. doi:10.1145/3173087

Kuipers, B. (2020). "Perspectives on Ethics of AI: Computer Science," in *Oxford Handbook of Ethics of AI*. Editors M. Dubber, F. Pasquale, and S. Das (Oxford, UK: Oxford University Press), 421–441.

Leyton-Brown, K., and Shoham, Y. (2008). *Essentials of Game Theory*. San Rafael, CA, USA: Morgan & Claypool.

Lin, P. (2013). *The Ethics of Autonomous Cars*. New York, NY, USA: The Atlantic Monthly.

Lin, P., Abney, K., and Bekey, G. A. (Editors) (2012). *Robot Ethics: The Ethical and Social Implications of Robotics* (Cambridge, MA, USA: MIT Press).

López, B. (2013). *Case-Based Reasoning: A Concise Introduction*. San Rafael, CA, USA: Morgan & Claypool.

Luhmann, N. (1979). "Trust: A Mechanism for the Reduction of Social Complexity," in *Trust and Power: Two Works by Niklas Luhmann 8* (Hoboken, NJ, USA: Wiley).

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *Amr* 20 (3), 709–734. doi:10.5465/amr.1995.9508080335

McAnany, P. A., and Yoffee, N. (Editors) (2010). *Questioning Collapse: Human Resilience, Ecological Vulnerability, and the Aftermath of Empire* (Cambridge: Cambridge University Press).

Mittelstadt, B., Russell, C., and Wachter, S. (2019). "Explaining Explanations in AI," in *Conf. On Fairness, Accountability, and Transparency (FAT*'19)*. arXiv: 1811.01439. doi:10.1145/3287560.3287574

Nissenbaum, H. (2001). Securing Trust Online: Wisdom or Oxymoron? *Boston Univ. L. Rev.* 81 (3), 635–664.

Nowak, M. A., and Roger, H. (2011). *Super Cooperators: Altruism, Evolution, and Why We Need Each Other to Succeed*. New York, NJ, USA: Free Press.

Nyholm, S., and Smids, J. (2020). Automated Cars Meet Human Drivers: Responsible Human-Robot Coordination and the Ethics of Mixed Traffic. *Ethics Inf. Technol.* 22 (4), 335–344. doi:10.1007/s10676-018-9445-9

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science* 366 (6464), 447–453. doi:10.1126/science.aax2342

Pinker, S. (2018). *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress*. New York City: Viking.

Pinker, S. (2011). *The Better Angels of Our Nature: Why Violence Has Declined*. New York City: Viking Adult.

Posner, E. A. (2000). *Law and Social Norms*. Cambridge, MA, USA: Harvard University Press.

Railton, P. (2003). *Facts, Values and Norms: Essays toward a Morality of Consequence*. Cambridge: Cambridge University Press.

Rand, D. G., and Nowak, M. A. (2013). Human Cooperation. *Trends Cogn. Sci.* 17, 413–425. doi:10.1016/j.tics.2013.06.003

Rawls, J. (1999). *A Theory of Justice*. revised edition. Cambridge, MA, USA: Harvard University Press.

Richerson, P. J., and Boyd, R. (2005). *Not by Genes Alone: How Culture Transformed Human Evolution*. Chicago, Illinois: University of Chicago Press.

Robinette, P., Allen, R., Li, W., Howard, A. M., and Wagner, A. R. (2016). "Overtrust of Robots in Emergency Evacuation Scenarios," in ACM/IEEE Int. Conf. Human Robot Interaction (HRI), 101–108.

Robinson, K. S. (2020). *The Ministry for the Future*. London, UK: Orbit Books.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., and Camerer, C. (1998). Not so Different after All: a Cross-Discipline View of Trust. *Amr* 23 (3), 393–404. doi:10.5465/amr.1998.926617

Russell, S., and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Third edition. Hoboken, NJ, USA: Prentice-Hall.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. New York City: Viking.

Saxe, J. G. (1949). "The Blind Men and the Elephant [poem]," in *Childcraft* (Chicago, Illinois, U.S.: Field Enterprises, Inc.), Vol. 2, 122–123.

Shafer-Landau, R. (Editor) (2013). *Ethical Theory: An Anthology*. second edition (Hoboken, NJ, USA: Wiley-Blackwell).

Sharkey, N. E. (2012). The Evitability of Autonomous Robot Warfare. *Int. Rev. Red Cross* 94 (886), 787–799. doi:10.1017/s1816383112000732

Siau, K., and Wang, W. (2018). Building Trust in Artificial Intelligence, Machine Learning, and Robotics. *Cutter Business Technol. J.* 31, 47–53.

Sinnott-Armstrong, W. (2015). "Consequentialism," in *The Stanford Encyclopedia of Philosophy. Winter 2015 Edition*. Editor E. N. Zalta.

Smith, A. (1776). *An Inquiry into the Nature and Causes of the Wealth of Nations*. London: W. Strahan and T. Cadell.

Sullins, J. P. (2020). "Trust in Robots," in *Routledge Handbook of Trust and Philosophy*. Editor J. Simon (Abingdon-on-Thames, Oxfordshire, UK: Taylor & Francis), 313–325.

Tainter, J. A. (1988). *The Collapse of Complex Societies*. Cambridge: Cambridge University Press.

Thaler, R. H. (1988). Anomalies: The Ultimatum Game. *J. Econ. Perspect.* 2 (4), 195–206.

Tomasello, M. (2019). *Becoming Human: A Theory of Ontogeny*. Cambridge, MA, USA: Harvard University Press.

Tomasello, M., Melis, A. P., Tennie, C., Wyman, E., and Herrmann, E. (2012). Two Key Steps in the Evolution of Human Cooperation. *Curr. Anthropol.* 53 (6), 673–692. doi:10.1086/668207

Vallor, S. (2016). *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford, UK: Oxford University Press.

Van Bavel, J. J., and Packer, D. J. (2021). *The Power of Us*. New York City: Little, Brown Spark.

van Wynsberghe, A., and Robbins, S. (2019). Critiquing the Reasons for Making Artificial Moral Agents. *Sci. Eng. Ethics* 25 (3), 719–735. doi:10.1007/s11948-018-0030-8

Vinge, V. (1993). The Coming Technological Singularity. *Whole Earth Rev.*.

von Neumann, J., and Morgenstern, O. (1953). *Theory of Games and Economic Behavior*. third edition. Princeton, NJ, USA: Princeton University Press.

Wachter, S., Mittelstadt, B., and Russell, C. (2018). Counterfactual Explanation without Opening the Black Box: Automated Decisions and the GDPR. *Harv. J. L. Technol.* 31 (2).

Wallach, W., and Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. Oxford, UK: Oxford University Press.

Wellman, H. M. (2014). *Making Minds: How Theory of Mind Develops*. Oxford, UK: Oxford University Press.

Wright, R. (2000). *Nonzero: The Logic of Human Destiny*. New York City: Pantheon.

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership