Programmed to do good

The categorical imperative as a key to moral behavior of social robots

Fink, Matthias; Maresch, Daniela; Gartner, Johannes

DOI
10.1016/j.techfore.2023.122793

Publication date
2023

Document Version
Final published version

Published in
Technological Forecasting and Social Change

Technological Forecasting & Social Change 196 (2023) 122793

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

journal homepage: www.elsevier.com/locate/techfore

Programmed to do good: The categorical imperative as a key to moral behavior of social robots

Matthias Fink [a,b,*], Daniela Maresch [b], Johannes Gartner [c]

[a] Johannes Kepler University Linz, Austria
[b] Grenoble Ecole de Management, France
[c] Delft University of Technology, the Netherlands

ARTICLE INFO

Keywords:
Social robots
Ethical dilemmas
Categorical imperative
Algorithms
Societal impact

ABSTRACT

Social robots—such as autonomous vehicles, service robots, or healthcare robots—are designed to support tasks in a broad range of human activities. However, these robots face moral dilemmas because they must make decisions that may do good for one human but potentially inflict harm on another. We argue that Kant's categorical imperative provides a framework for algorithm-based moral decision-making. By systematically addressing ethical concerns from the outset in the development of the algorithms that steer social robots, their designers can help ensure that such robots promote the well-being of individuals, communities, and society. We conclude that those involved in the development of social robots need to embed ethics into their design and functioning. The solutions to the ethical dilemmas we advance in this paper can help improve the adoption and impact of social robots. The presented insights contribute to research, practice, and policy.

1. Introduction

Social robots include autonomous vehicles (De Moura et al., 2020; Tan et al., 2021), service robots (Van Wynsberghe, 2016; Pollmann et al., 2023), or healthcare robots (Na et al., 2023; Søraa et al., 2021). During their interactions with humans, social robots make decisions that may do good for one human but potentially inflict harm on another (Duffy, 2006; Frennert and Östlund, 2014; Sharkey, 2008; Schneiders et al., 2022; Pirni et al., 2021). This so-called "double-effect setup" (Bentzen, 2016) raises major moral dilemmas, because the social robot must decide who to help, who to harm, or how to distribute harm between humans (Hurtado et al., 2021). Oftentimes, the robot might simply stop and take no action at all (Tolmeijer et al., 2023), or it might take a sequence of actions over a longer period (Rosado et al., 2016) to avoid the dilemma. In other cases, however, the decision must be taken—potentially under immense time pressure and with severe consequences. The question of how the algorithms that steer social robots can account for these dilemmas has attracted vivid attention in public and scientific discourse (Boada et al., 2021; Deng, 2015).

In public discourse, the ethical quandary surrounding the actions of social robots has often been resolved by implementing Asimov's Laws, a set of three hierarchical principles that guide the behavior of social robots to prioritize human safety and human interests (Asimov, 1942). These principles state that (i) a robot must not injure a human or allow a human to come to harm through inaction, (ii) a robot must obey orders given by humans unless the order would injure a human or allow a human to come to harm, and (iii) a robot must protect its own existence as long as this is not in conflict with (i) or (ii). Asimov's Laws are "presented as a simple, comprehensive, and logically organized structure of inference for moral judgement of robots with their users" (Van Dang et al., 2018). However, when we attempt to apply these principles to today's social robots, their limitations soon become apparent. Asimov's simple laws reflect a context where a robot must choose between preset options; they cannot account for the complexity and dynamics of relationships between humans and social robots as well as how those relationships are expressed. Thus, in many cases these laws do not provide definite behavioral guidance (Murphy and Woods, 2009). Moreover, the question of how social robots should distribute harm between humans cannot be solved based on these principles (Awad et al., 2018). In a road accident, for example, autonomous vehicles will sometimes have to decide between running over pedestrians to prevent harm to their passengers or sacrificing themselves and their passengers to save the pedestrians (Bonnefon et al., 2016). Interestingly, when exploring how to address this ethical dilemma, earlier empirical findings

* Corresponding author at: Institute of Innovation Management, Johannes Kepler University Linz, Austria
E-mail addresses: matthias.fink@grenoble-em.com (M. Fink), daniela.maresch@grenoble-em.com (D. Maresch), johannes.gartner@jgg.at (J. Gartner).

https://doi.org/10.1016/j.techfore.2023.122793
Received 31 March 2023; Received in revised form 1 August 2023; Accepted 14 August 2023
Available online 31 August 2023

suggest a paradoxical situation. While participants would like others to buy utilitarian autonomous vehicles that would sacrifice their passengers for the greater good, they themselves would prefer to travel in autonomous vehicles that prioritize their own safety (Bonnefon et al., 2016).

Nor do Asimov's principles account for the human need for human attention. In the health industry, social robots may face the ethical dilemma of balancing a patient's independence and freedom against the need for medication adherence (Becker et al., 2013) and of systematic erroneous algorithmic decision-making that could lead to misdiagnoses or incorrect medication and treatment (Zarsky, 2016; Mittelstadt et al., 2016; Asaro, 2019). Exacerbating these challenges, moral preferences may vary globally and among different demographics (Awad et al., 2018).

Drawing on moral philosophy, we argue that the categorical imperative (Kant, 2003; Kant, 1981; Paton, 1971) in combination with the life world concept (Schuetz and Luckmann, 1973) and the idea of a public discourse (Habermaas, 1973) can provide a key to solving moral dilemmas related to social robots. Similar ideas have been developed in practical philosophy in recent decades (Powers and Faden, 2006). We develop our argument along the following lines. The factual constraint (German: *Sachzwang*) argument that has underpinned much of the discourse so far (Apel, 1988; Awad et al., 2018) has been shown to be a naturalistic fallacy (Hume, 1739-40; Moore, 1903). If we accept that actors—be they human or machines—facing social dilemmas are not constrained by natural laws, we realize that they always have a choice. For actors, such choices come with the burden of justification, which makes human-social robot interactions a subject of practical ethics (Riek and Howard, 2014; Arif et al., 2017; Bolander, 2019). We therefore discuss the functions of rules, maxims, and laws—and thus the basis of moral behavior—using the life world concept (Schuetz and Luckmann, 1973) and the categorical imperative (Kant, 2003). We deduce that specific actions on an individual level cannot be morally legitimized through either factual constraints or considerations of utility. Public discourse (Habermaas, 1973) then provides the backdrop for developing a brief sketch of the discourse of determining aims, which discusses the communal utility of a maxim and establishes laws as the basis for the moral legitimation of autonomous actions. The paper concludes with the insight that individual utility considerations cannot provide the basis for morally legitimizing the specific actions of social robots. However, on the collective level of public discourse, considerations of utility can become part of formulating laws that provide a basis for legitimizing maxims, which, in turn, indirectly justify the decisions and resulting actions taken by social robots.

Our contribution has important implications for research, practice, and policy. Our paper shows that algorithms steering social robots cannot be developed based on optimization models that maximize utility by balancing the beneficial and adverse human impacts of social robots' actions. It becomes apparent that the outcome of such calculations can never be justified from a moral point of view. Rather, these algorithms need to assess each possible action of the social robot they steer using a simulation that tests whether one would want this action to become a generally valid law. We explain why the action may be performed only if the result of the simulation is in line with the consensus derived from the public discourse.

From a research perspective, the approach we suggest provides a theoretical basis for exploring ethical issues associated with social robots that interact with humans, and for developing empirical studies that investigate the impact of social robots on society. The multi-level framework that summarizes the results of our theoretical considerations can provide a starting point for attractive follow-up research.

From a practical perspective, our framework offers guidance for the developers and manufacturers of social robots, helping them to ensure that their products are designed and deployed in an ethically responsible manner. The solution we suggest also implies that programmers need to be aware of their moral responsibility. Finally, from a policy perspective, our framework provides a starting point for developing regulations and guidelines for the use of social robots, with a focus on ensuring that they are aligned with the moral values and principles of society as a whole.

2. Why social robots “can” and “ought to” act morally

The key to social robots' ethical dilemmas is maxims, because they can be translated into algorithms that steer the robots' actions in human-machine interactions. In the remainder of this chapter, we explain why.

2.1. Unmasking a naturalistic fallacy

The starting point for our argument is the observation that utility considerations on the individual level cannot provide a basis to justify moral decisions. This paradox can easily be explained following Lutz-Bachmann (2018). The aim of designing (practical) laws for social robots' interactions with humans is to establish certain actions as desirable and others as forbidden. As the maxims derived from these laws prescribe specific actions in specific situations, they restrict a robot's scope for action (McNair, 2000; Lukow, 2003). In this regard, the laws are prescriptive statements; they only make sense if there are incentives for one to adhere to them when developing the algorithms (maxims) for the actions of a social robot (Herman, 1990). If we justify the legitimacy of a law based on its long-term utility, it is difficult to comprehend why we need such a law in the first place. If it is in the long-term interest of the completely rational social robot to act according to maxims, neither maxims that restrict the robot's behavior nor the respective governing laws are needed. Social robots that do not adhere to a maxim that is in line with the long-term utility for the user violate their own interest and thus act irrationally. Because social robots act rationally, any behavior that breaks the law cannot exist.

As long as the theoretical framework assumes actors such as social robots to be fully rational, we cannot use utility considerations on the individual level to legitimize maxims that restrict behavior. Nor can value systems such as religion serve as a basis for legitimizing maxims. So, how can we legitimize maxims for social robots?

Applied sciences such as informatics and economics traditionally claim to be value-free: involving no normative judgments at all (Robbins, 1932; Nordgren, 2012; Ulrich, 2008). The normative postulates of engineers and economists seem to require no justification as long as we claim that factual constraints bind the actors to these postulates. If we accept that social robots make decisions in social dilemmas based on factual constraints and comparative utility considerations, we do not need a moral benchmark to evaluate their actions. This would result in social robots having ethics without morals (Ulrich, 2008). The right actions result directly from the conditions and, thus, individual utility considerations. In this way, a “must” that almost resembles a natural law replaces an “ought” that is open to discussion and has to be justified. Economists and engineers hence escape their responsibility to justify the social robots' actions by pointing to the factual constraints that can purportedly be found in the social context in which the interaction is embedded (Shleifer, 2004).

However, the “is-ought” problem, or Hume's Law (Hume, 1739-40), is the thesis that we cannot infer the right action (that which ought to be) from past, current, and future conditions (that which is). Hence, we cannot infer what social robots ought to do from factual constraints or past data. Rather, the actions of social robots depend on their programmers' will (Kant: *Willkür*). In fact, the ostensibly factual constraints are cognitive constraints in the thinking of the programmers (Ulrich, 2002), as factual constraints only prevail where natural law determines a cause-and-effect relation. Kant (2003) distinguishes between cognitive constraints (“causality of freedom”; Kant: *Gründe*) and factual constraints (“causality of nature”; Kant: *Ursachen*). Factual constraints can never determine the interactions between humans and social robots: even if every human in the interaction behaved in a strictly rational manner and only ever aimed at maximizing personal utility, the actions

would still be taken by humans and therefore depend on their will. In such a scenario, the ultimate aim of humans' will would be maximum utility. Even if the process of determining aims is based on reason, the result is an act of will (decision) and therefore has to be justified. Thus, Ulrich (2008) concludes that economic determinism makes no sense as an empirical hypothesis. The constraints experienced in the utility maximization model used for algorithms cannot serve as a justification for the social robots' inability to act morally. Otherwise, economic theories would be mistaken for reality (Ulrich, 2008).

The programmers of social robots cannot derive the right actions from the conditions to be experienced because the ostensibly factual constraints of the interaction between humans and social robots are not natural laws but rather cognitive constraints. A social robot's actions are the result of a reason-based algorithm designed by a programmer with a free will (McNair, 2000). Thus, we need principles for these actions that can guide the programmers who design the algorithms. Laws are such principles for action that are generally valid within a community. This brings us to the question of how laws are rooted in maxims and how they develop normative power for the individual social robot via the programmer.

2.2. Heteronomy, autonomy, and duty

Social robots' moral behavior is determined by the programmers who create the algorithms that steer them (Pirni et al., 2021). Thus, it is important to focus on the programmers who design these algorithms (Allen et al., 2005). According to Kant, individuals are "citizens of two worlds": the world of appearances and the world of reason (Kant, 2003). In the world of appearances, individuals are embedded in a web of causal connections. Within the framework of social relationships, actions, and experience, knowledge elements are categorized and stored in their individual knowledge pool before being partly transferred, in a socialized form, into the community's collective knowledge pool (Schuetz and Luckmann, 1973). At this level, individuals use their experience as the starting point for establishing their will and the actions arising from it. Individuals' behavior is thus steered by appearances and the mutual dependencies they underlie. It does not originate from the individual but is externally determined; heteronomous.

However, the world of appearances cannot provide a universal benchmark for evaluating actions as "good" or "bad," as individual opinions and experiences can differ (Schwarz, 2006). Laws as a moral benchmark must therefore have a different foundation. Kant finds this foundation in human reason—specifically, practical reason, which is directed towards actions that are the expression of will. It is important to see that in Kant's thinking the will is free from external influence and only originates from pure reason. Kant calls this form of will-formation autonomy. This autonomy establishes an individual's freedom, which is not the freedom to do whatever one wants but rather the freedom to act in accordance with duty. Duty is not imposed from outside, but arises from reason itself (O'Neill, 1989a, 1989b). If the individual is reasonable, i.e., if they follow reason, which Kant assumes, they have no other choice but to do their duty. But how does this duty come about? And how can the social robot's programmer discover which maxims they are bound by?

2.3. Hypothetical versus categorical imperative and the normative force of maxims

The first question—how this duty comes about—leads us to the construct of the categorical imperative. Imperatives express an *ought* (Kant, 2003). What distinguishes the categorical imperative from its hypothetical counterpart is the way in which it develops normative force (Schwarz, 2006). In the case of hypothetical imperatives, the means can be logically derived from the end. For example, if we wish to get from Rome to London quickly, we should take a plane. Hence, the imperative is tied to the desired end, i.e., getting to London quickly. The normative force depends on this end, and thus on the subject following this end (Korsgaard, 1996). Hypothetical imperatives are therefore not generally valid and cannot become the foundation of a moral system (Mackie, 1977). The categorical imperative, however, only refers to the form of the action. It is anchored in the principle of morality. What counts is the good intention and not the desired end (Korsgaard, 2002). Thus, the internal alignment of the will with moral laws is crucial (Schneewind, 1992).

Only a principle for action that is based on reason can claim to be generally valid. This principle must be abstracted from both subject and object (Galvin, 1999; Fink et al., 2023). Therefore, only a formal principle for action can be generally valid (Arntzen, 1991). Kant suggests a thought experiment for individuals to check the moral quality of their subjective principles for action, which he calls *maxims* (Fig. 1). If the maxim governing one's actions is of such a nature that one would want it to become a universal law, the action derived from this maxim is moral (Schneewind, 1992). This thought experiment has to be conducted for every maxim. The normative force that arises from this thought experiment does not contradict free will. Individuals are free from the constraints of the world of appearances (level 1) if they act autonomously, but, because they are reasonable, they must act in line with their duty, as imposed by the thought experiment of the categorical imperative (level 2). Implementing the categorical imperative leads to autonomous action that is in accordance with maxims. This is only possible because the free will is committed to the material reason that it has autonomously given itself (Korsgaard, 2002). Autonomous action is then oriented towards duty, formulated as maxims that are derived from moral reasoning based on the categorical imperative (level 3). As a result, what an actor wants is identical to what they ought to do. The individual programmer needs to design the algorithms that steer the social robots in a way that it performs the thought experiment for any decision that arises in its interaction with humans. But how do maxims link to the aims of communities who are affected by social robots?

2.4. Establishing maxims in the discourse of determining aims for communities

By adopting a maxim, the actor commits themselves to the underlying law, but no maxim is valid beyond the individual level. In the thought experiment of the categorical imperative, however, the individual takes the whole community into account through the claim of a fictitious general validity, by including other people's ends in their deliberations (Paton, 1971). Here, the interface between the individual and collective levels can be identified: individual-level maxims are linked to community-level laws.

On the level of the community, the only possible criterion for morally evaluating actions is whether the claims made for a law's validity can be argued to be universal. In order to formulate, through public discourse, a law that is generally valid for the community, perspectives must be intersubjectively exchangeable between all the actors and people

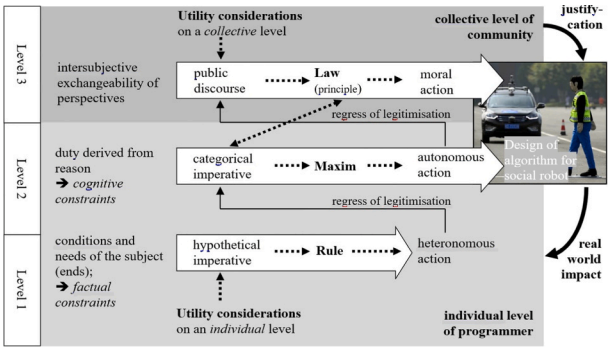

Fig. 1. Multi-level model of moral reasoning of social robots.

involved (Ulrich, 2008). Individuals debate the results of their individual thought experiments through public discourse; if they can reach consensus, appropriate laws emerge on the collective level. In turn, each individual then uses these laws as the basis for legitimizing their maxims (Fink et al., 2023). Hence, the maxim first formulated in the thought experiment (which is therefore subjective) is made explicit through community-level public discourse. In this way, laws are developed and established in the community as the benchmark of morality.

3. Multi-level framework of ethics for social robots

We summarize the results of the above discussion in a multi-level framework (see Fig. 1). At the center of our argument is the maxim. Maxims emerge from the thought experiment (categorial imperative) if it is desirable for an individual rule to be generally valid. The thought experiment takes the appraisal out of the context of the actor's own lifeworld and preferences. Actions rooted in maxims are autonomous because they are not motivated by individuals' utility considerations. At the same time, the discourse on the collective level of communities frames the maxims of their members because the principles accepted throughout the community impact on the rational considerations of the individual during the thought experiment. As a result, maxims derive duty from reason, but still incorporate the intersubjective principles of the community. Thus, maxims are ideal to guide the design of algorithms steering social robots.

Interestingly, the categorial imperative is a perfect fit for the formulation of a set of maxims for social robots interacting with a specific community. The logical form of a thought experiment matches the technique of AI-based simulation, and the public discourse within a community can be integrated via big data covering scientific publications and media releases. The digital era, thus, provides the perfect conditions for employing the categorical imperative for developing algorithms that enable social robots to act in ways that are collectively perceived as good in the community they interact with. Clearly, the set of maxims will differ between communities, and the maxims will evolve through machine learning processes built into the social robots. It is important to recognize not only that ethical considerations develop over time, but also that many different discourses are unfolding in communities in parallel (Niemi et al., 2022). While AI accelerates the decision process, it can only mirror the ethical stance that frames the discourses within communities. “AI itself is, per default, not irrational or biased; it just extrapolates patterns that exist in the real-world data that we give it to learn and to exploit these patterns in order to distinguish between the potential decision alternatives” (Antretter et al., 2020; p. 1).

4. Discussion and conclusion

Social robots are designed to support a broad range of human activities including mobility, customer service, healthcare, daily domestic life, and education (Boada et al., 2021; Borghi and Mariani, 2022). However, the use of social robots raises ethical questions such as privacy and data protection (Lutz et al., 2019; Lutz and Tamó-Larrieux, 2020; Chatterjee et al., 2021), safety, and responsibility, as well as safeguarding and transparency. Consequently, the ethical principles of nonmaleficence, beneficence, autonomy, and fairness should be transposed to robotics (Körtner, 2016).

Thus, the introduction of social robots into society will require that they follow ethical principles that go beyond consequentialism (Bentzen, 2016). We argue that Kant's categorical imperative is a general principle that provides a framework for making moral decisions that can be applied to social robots (Zoshak and Dew, 2021). This principle can be used in the development of algorithms for social robots to evaluate the ethical implications of different actions or decisions arising from interactions with humans (Powers and Faden, 2006). If a design choice would result in harm or injustice if it were universally applied, it would not be in accordance with the categorical imperative and, thus, would not be an acceptable option.

The solution to the ethical dilemmas faced by social robots advanced here can help improve such robots' impact by ensuring that they are designed in ways that align with the universal moral principles of the community where they are used. Systematically addressing ethical concerns from the outset in the development of the algorithms that steer social robots can help ensure that these robots are used in ways that promote the well-being of individuals, communities, and society. Engineers and designers involved in the development of social robots need to embed ethics into their functioning. The approach suggested in this paper calls for responsible innovation, which is “a transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view on the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products” (Von Schomberg, 2011, p. 9). To realize this responsible innovation approach in the realm of social robots, those involved in the product innovation process, particularly programmers, must be aware of their responsibility (Shea and Hawn, 2019). The suggested multi-level framework can offer a good starting point because it speaks to the social situatedness of responsible product innovation (Scherer and Voegtlin, 2020). While maxims enable actors to assess behavioral options autonomously, their embeddedness in the community involves them in the public discourse (Blok and Lemmens, 2015; Jordan, 2008) and thus infuses their rational deliberations with the elements of responsiveness, inclusion, anticipation, and reflexivity (Stilgoe et al., 2013; Smith and Semin, 2004; Wood and Williams, 2014).

Specifically, we show that the categorical imperative is a powerful principle for embedding ethics in social robots, because it provides a formal framework that dovetails with the functioning of algorithms and can consider factors on both the individual and the community level. This cross-level legitimization of ethical decisions, however, also imposes certain limitations on the application of the categorical imperative to social robots. Firstly, the application of the formal framework in specific situations via simulation requires extensive data and computational capacity. Secondly, this database needs to be context-sensitive, and data is created differently across contexts (Dąbrowska et al., 2022). Every society has a specific pool of experiences and knowledge underpinning its set of laws, which in turn provide the moral evaluation framework for the maxims steering the individual social robot. Such contextual differences and their change over time must be considered. Third, embedding ethics in social robots is costly. These costs occur in the development of the algorithm, during testing, and over robots' entire lifespan. Given the obvious benefits for individuals, communities, and society at large, a feasible way to cover these costs is the introduction of a global “robo-tax” that is charged on every unit sold. So far, such taxes have only been discussed as a way to compensate for robots' negative effects on labor markets (Costinot and Werning, 2018; Guerreiro et al., 2020). Finally, we do not suggest a specific ethics for social robots, and nor do we provide the rules, maxims, and laws to support it. Rather, we develop a framework that can help to develop a path towards such an ethics for social robots.

The ideas developed in this paper contribute to research, practice, and policy. We develop an argument for why algorithms for social robots cannot be based on optimization models that balance the beneficial and adverse impacts of their actions on humans. We show that the categorical imperative is a possible key to embed ethics in algorithms steering social robots. This approach provides a theoretical basis for exploring ethical issues associated with social robots. It also offers guidance for practitioners such as developers and manufacturers to ensure their products are ethically responsible. When users of social robots are making their purchase decisions, they can weigh up the algorithms implemented in the products available. For policymakers, our insights provide a starting point for developing regulations and guidelines for the development and use of social robots that align with society's moral values and principles.

CRediT authorship contribution statement

Matthias Fink: Conceptualization, Data collection and curation; Methodology, Formal analysis, Visualization, Supervision, Validation, Writing original draft; Writing - review & editing

Daniela Maresch: Conceptualization, Data curation; Writing - original draft; Writing - review & editing

Johannes Gartner: Conceptualization, Validation, Writing - original draft; Writing - review & editing

Data availability

No data was used for the research described in the article.

References

Allen, C., Smit, I., Wallach, W., 2005. Artificial morality: top-down, bottom-up, and hybrid approaches. Ethics Inf. Technol. 7, 149–155.

Antretter, T., Blohm, I., Siren, C., Grichnik, D., Malmström, M., Wincent, J., 2020. Do algorithms make better-and fairer-investments than angel investors? Harv. Bus. Rev. https://hbr.org/2020/11/do-algorithms-make-better-and-fairer-investments-than-angel-investors.

Apel, K.O.D., 1988. Verantwortung. Das Problem des Übergangs zur postkonventionellen Moral. Suhrkamp, Frankfurt am Main.

Arif, D., Ahmad, A., Bakar, M.A., Ihtisham, M.H., Winberg, S., 2017. Cost effective solution for minimization of medical errors and acquisition of vitals by using autonomous nursing robot. In: Proceedings of the 2017 International Conference on Information System and Data Mining, pp. 134–138.

Arntzen, S., 1991. Kant on Willing a Maxim as a Universal Law. In: Funke, G. (Hrsg.) (Ed.), Akten des Siebenten Internationalen Kant-Kongresses. Bd. II/1. Bonn, pp. 265–275.

Asaro, P.M., 2019. AI ethics in predictive policing: from models of threat to an ethics of care. IEEE Technol. Soc. Mag. 38 (2), 40–53.

Asimov, I., 1942. Runaround – a short story. In: Astounding Science Fiction, May, pp. 94–103.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., Rahwanet, I., 2018. The moral machine experiment. Nature 563, 59–64.

Becker, H., Scheermesser, M., Früh, M., Treusch, Y., Auerbach, H., Hüppi, R.A., Meier, F., 2013. Robotik in Betreuung und Gesundheitsversorgung. In: TA-SWISS 58/2013. vdf Hochschulverlag. Zürich.

Bentzen, M.M., 2016. The Principle of Double Effect Applied to Ethical Dilemmas of Social Robots. In Robophilosophy/TRANSOR, pp. 268–279.

Blok, V., Lemmens, P., 2015. The emerging concept of responsible innovation. Three reasons why it is questionable and calls for a radical transformation of the concept of innovation. In: Koops, B.J., Oosterlaken, I., Romijn, H., Swierstra, T., van den Hoven, J. (Eds.), Responsible Innovation 2. Springer, Cham.

Boada, J.P., Maestre, B.R., Genís, C.T., 2021. The ethical issues of social assistive robotics: a critical literature review. Technol. Soc. 67, 101726.

Bolander, T., 2019. What do we loose when machines take the decisions? J. Manag. Gov. 23, 849–867.

Bonnefon, J.F., Shariff, A., Rahwan, I., 2016. The social dilemma of autonomous vehicles. Science 352 (6293), 1573–1576.

Borghi, M., Mariani, M.M., 2022. The role of emotions in the consumer meaning-making of interactions with social robots. Technol. Forecast. Soc. Chang. 182, 121844.

Chatterjee, S., Chaudhuri, R., Vrontis, D., 2021. Usage intention of social robots for domestic purpose: from security, privacy, and legal perspectives. Inf. Syst. Front. 1–16.

Costinot, A., Werning, I., 2018. Robots, Trade, and Luddism: A Sufficient Statistic Approach to Optimal Technology Regulation (No. w25103). National Bureau of Economic Research.

Dąbrowska, J., Almpanopoulou, A., Brem, A., Chesbrough, H., Cucino, V., Di Minin, A., Giones, F., Hakala, H., Marullo, C., Mention, A.-L., Mortara, L., Nørskov, S., Nylund, P.A., Oddo, C.M., Radziwon, A., Ritala, P., 2022. Digital transformation, for better or worse: a critical multi-level research agenda. R&D Manag. 52, 930–954.

De Moura, N., Chatila, R., Evans, K., Chauvier, S., Dogan, E., 2020, October. Ethical decision making for autonomous vehicles. In: In 2020 IEEE Intelligent Vehicles Symposium (Iv). IEEE, pp. 2006–2013.

Deng, B., 2015. Machine ethics: the robot's dilemma. Nature 523, 24–26.

Duffy, B.R., 2006. Fundamental issues in social robotics. Int. Rev. Inf. Ethics 6, 31–36.

Fink, M., Gartner, J., Harms, R., Hatak, I., 2023. Ethical orientation and research misconduct among business researchers under the condition of autonomy and competition. J. Bus. Ethics 183 (2), 619–636.

Frennert, S., Östlund, B., 2014. Seven matters of concern of social robots and older people. Int. J. Soc. Robot. 6, 299–310.

Galvin, R.F., 1999. Slavery and universalizability. In: Kant-Studien, 90 (2), pp. 191–203.

Guerreiro, J., Rebelo, S., Teles, P., 2020. Should Robots Be Taxed?, NBER Working Paper Series. Working Paper 23806,. National Bureau of Economic Research. http://www.nber.org/papers/w23806.

Habermaas, J., 1973. Erkenntnis und Interesse, Frankfurt am Main.

Herman, B., 1990. Morality as Rationality. A Study of Kant's Ethics, New York – London.

Hume, D.A., 1739-40. Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects. London.

Hurtado, J.V., Londoño, L., Valada, A., 2021. From learning to relearning: a framework for diminishing bias in social robot navigation. Front. Robot. AI 8, 650325.

Jordan, A., 2008. The governance of sustainable development: taking stock and looking forward. Environ. Plan. C Pol. Space 26, 17–33.

Kant, I., 1981. Grounding for the Metaphysics of Morals, translated by J. Ellington, Hacket. M. (2006). Prospects for a Kantian Machine. Intelligent Systems, 21(4), pp. 46–51.

Kant, I., 2003. Critique of Pure Reason (Kritik der praktischen Vernunft, 1788), Translated by Kemp Smith, N., Basingstoke.

Korsgaard, Ch., 1996. Creating the Kingdom of Ends, Cambridge.

Korsgaard, Ch., 2002. Self-Constitution: Action. Identity and Integrity, The Locke Lectures (I-IV).

Körtner, T., 2016. Ethical challenges in the use of social service robots for elderly people. Z. Gerontol. Geriatr. 49 (4), 303–307.

Lukow, P., 2003. Maxims, Moral Responsiveness, and Judgment. In: Kant-Studien. 94 (3), pp. 405–425.

Lutz, C., Tamó-Larrieux, A., 2020. The robot privacy paradox: understanding how privacy concerns shape intentions to use social robots. Human-Mach. Commun. 1, 87–111.

Lutz, C., Schöttler, M., Hoffmann, C.P., 2019. The privacy implications of social robots: scoping review and expert interviews. Mob. Media Commun. 7 (3), 412–434.

Lutz-Bachmann, M., 2018, March. Metaphysik. Überlegungen zu einem Konzept von Philosophie im Anschluss an Kant. In: Wozu Metaphysik? Verlag Karl Alber, pp. 79–94.

Mackie, J.M., 1977. Ethics. Inventing Right and Wrong, London.

McNair, T., 2000. Universal Necessity and Contradiction in Conception. In: Kanz-Studien. 91 (1), pp. 25–43.

Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., Floridi, L., 2016. The ethics of algorithms: mapping the debate. Big Data Soc. 3 (2) (2053951716679679).

Moore, G.E., 1903. Principia Ethica, Cambridge.

Na, E., Jung, Y., Kim, S., 2023. How do care service managers and workers perceive care robot adoption in elderly care facilities? Technol. Forecast. Soc. Chang. 187, 122250.

Murphy, R.R., Woods, D.D., 2020. Beyond Asimov: The three laws of responsible robotics. In: Machine Ethics and Robot Ethics. Routledge, pp. 405–411.

Niemi, L., Stenholm, P., Hakala, H., Kantola, J., 2022. Immanent sensemaking by entrepreneurs and the interpretation of consumer context. Int. Small Bus. J. 40 (8), 966–990.

Nordgren, A., 2012. The web-rhetoric of companies offering home-based personal health monitoring. Health Care Ann. 20, 103–118.

O'Neill, O., 1989a. Consistency in action. In: O'Neill (Ed.), Constructions of Reason. Cambridge, pp. 81–104.

O'Neill, O., 1989b. Universal laws and ends-in-themselves. In: O'Neill (Ed.), Constructions of Reason. Cambridge, pp. 126–144.

Paton, H.J., 1971. The Categorical Imperative: A Study in Kant's Moral Philosophy, vol. 1023. University of Pennsylvania Press.

Pirni, A., Balistreri, M., Capasso, M., Umbrello, S., Merenda, F., 2021. Robot care ethics between autonomy and vulnerability: coupling principles and practices in autonomous systems for care. Front. Robot. AI 184.

Pollmann, K., Loh, W., Fronemann, N., Ziegler, D., 2023. Entertainment vs. manipulation: personalized human-robot interaction between user experience and ethical design. Technol. Forecast. Soc. Chang. 189, 122376.

Powers, M., Faden, R.R., 2006. Social Justice: The Moral Foundations of Public Health and Health Policy. Oxford University Press, USA.

Riek, L., Howard, D., 2014. A code of ethics for the human-robot interaction profession. In: Proceedings of We Robot. Available at SSRN. https://ssrn.com/abstract=2757805.

Robbins, L., 1932. An Essay on the Nature and Significance of Economic Science. London.

Rosado, A.L., Chien, S., Li, L., Yi, Q., Chen, Y., Sherony, R., 2016. Certainty and critical speed for decision making in tests of pedestrian automatic emergency braking systems. IEEE Trans. Intell. Transp. Syst. 18 (6), 1358–1370.

Scherer, A.G., Voegtlin, C., 2020. Corporate governance for responsible innovation: approaches to corporate governance and their implications for sustainable development. Acad. Manag. Perspect. 34, 182–208.

Schneewind, J.B., 1992. Autonomy, obligation and virtue: an overview of Kant's moral philosophy. In: Guyer, P. (Ed.), The Cambridge Companion to Kant. Cambridge, pp. 309–341.

Schneiders, E., Cheon, E., Kjeldskov, J., Rehm, M., Skov, M.B., 2022. Non-dyadic interaction: a literature review of 15 years of human-robot interaction conference publications. ACM Trans. Human-Robot Interact. 11 (2), 1–32.

Schuetz, A., Luckmann, T., 1973. The Structures of the Life-World. (Strukturen der Lebenswelt.), Translated by Zaner, R. M. and Engelhardt, H. T., Evanston.

Schwarz, M., 2006. Der Begriff der maxim bei Kant. Eine Untersuchung des maximsbegriffs in Kants praktischer Philosophie. In: Vossenkuhl, W. (Ed.), Philosophie im Kontext, Berlin.

Sharkey, N., 2008. The ethical frontiers of robotics. Science 322 (5909), 1800–1801.

Shea, C.T., Hawn, O.V., 2019. Microfoundations of corporate social responsibility and irresponsibility. Acad. Manag. J. 62, 1609–1642.

Shleifer, A., 2004. Does competition destroy ethical behavior? Am. Econ. Rev. 94 (2), 414–418.

Smith, E.R., Semin, G.R., 2004. Socially situated cognition: cognition in its social context. Adv. Exp. Soc. Psychol. 36, 53–117.

Søraa, R.A., Nyvoll, P., Tøndel, G., Fosch-Villaronga, E., Serrano, J.A., 2021. The social dimension of domesticating technology: interactions between older adults, caregivers, and robots in the home. Technol. Forecast. Soc. Chang. 167, 120678.

Stilgoe, J., Owen, R., Macnaghten, P., 2013. Developing a framework for responsible innovation. Res. Policy 42, 1568–1580.
Tan, S.Y., Taeihagh, A., Tripathi, A., 2021. Tensions and antagonistic interactions of risks and ethics of using robotics and autonomous systems in long-term care. Technol. Forecast. Soc. Chang. 167, 120686.
Tolmeijer, S., Arpatzoglou, V., Rossetto, L., Bernstein, A., 2023. Trolleys, crashes, and perception—a survey on how current autonomous vehicles debates invoke problematic expectations. AI Ethics 1–12.
Ulrich, P., 2002. Ethics and economics. In: Zsolnai, L. (Ed.), Ethics in the Economy. Handbook of Business Ethics. Oxford/Bern, pp. 9–37.
Ulrich, P., 2008. Integrative Economic Ethics: Foundations of a Civilized Market Economy. Cambridge.
Van Dang, C., Jun, M., Shin, Y.B., Choi, J.W., Kim, J.W., 2018. Application of modified Asimov's laws to the agent of home service robot using state, operator, and result (Soar). Int. J. Adv. Robot. Syst. 15 (3) (1729881418780822).
Van Wynsberghe, A., 2016. Service robots, care ethics, and design. Ethics Inf. Technol. 18 (4), 311–321.
Von Schomberg, R., 2011. Towards Responsible Research and Innovation in the Information and Communication Technologies and Security Technologies Fields. Publications Office of the European Union, Luxembourg.
Wood, M.S., Williams, D.W., 2014. Opportunity evaluation as rule-based decision making. J. Manag. Stud. 51, 573–602.
Zarsky, T., 2016. The trouble with algorithmic decisions an analytic road map to examine efficiency and fairness in automated and opaque decision making. Sci. Technol. Hum. Values 41 (1), 118–132.
Zoshak, J., Dew, K., 2021. Beyond kant and bentham: how ethical theories are being used in artificial moral agents. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–15.

Matthias Fink is a professor of innovation management at JKU Linz, Austria and a professor of strategy at Grenoble Ecole de Management. Matthias previously had professorial roles at ARU Cambridge (UK), Leuphana University Lueneburg (Germany) and was researcher at the Institute for Small Business Management and Entrepreneurship as well as head of the Research Institute for Liberal Professions at WU Vienna University of Economics and Business (Austria). Matthias holds a Ph.D. and a postdoctoral qualification (habilitation) from WU Vienna University of Economics and Business and was a visiting professor at several universities e.g., University of Sydney (Australia), Universitat Autònoma de Barcelona (Spain), and University of Twente (The Netherlands). His research concerns Strategy as Practice (SAP) in the context of innovation (generation of novel processes and valuable market offerings), socio-technical transformation (interplay between social change, technological progress and business models) and business venturing (in new and established organizations). His emphasis is on three underlying challenges (1) overcoming behavioral uncertainty with interpersonal trust, (2) bridging the gap between intention and action, and (3) responsibility of socially, regionally and temporally embedded actors. Matthias addresses these challenges with a combination of qualitative and quantitative research methods with the aim to generate impact on theory, practice and policy. His research has been published in journals such as Journal of Management Studies, Technological Forecasting and Social Change, Journal of Banking and Finance, British Journal of Management, Journal of Business Ethics, Regional Studies, Journal of Business Venturing, as well as Entrepreneurship Theory & Practice. Additionally, he has published four monographs, two edited volumes and has guest edited several special issues to international journals. He serves as associate editor for Technology Forecasting and Social Change.

Daniela Maresch is an associate professor at GEM, where she joined the Management, Technologies and Strategy Department in December 2020. Daniela also hold a fractional position as associate professor at the University of Southern Denmark (DK). Daniela has a PhD in Business Administration and an LL.M.(WU) in Business Law from WU Vienna University of Business and Economics (Austria) as well as a postdoctoral qualification (Habilitation) from Johannes Kepler University Linz (Austria). Before joining academia in 2014, she gained practical experience in financial reporting and corporate law working for a major Austrian utility and a renowned Viennese law firm. In her research, Daniela employs her interdisciplinary expertise. The focus of her research is on topics at the intersection of entrepreneurship, regional development and innovation, such as entrepreneurial finance, technology entrepreneurship, social and migrant entrepreneurship, and entrepreneurship education. The results of her work have been published both in scholarly journals such as Technological Forecasting and Social Change, Journal of Corporate Finance, and Entrepreneurship and Regional Development, as well as in transfer publications such as the European Central Bank Working Paper Series. Daniela's research has attracted funding from institutions such as the EU and the Swedish Kamprad Family Foundation for Entrepreneurship, Research & Charity as well as private enterprises.

Johannes Gartner is an assistant professor at Delft University of Technology, the Netherlands. Johannes holds a PhD in Social and Economics Science from the Johannes Kepler University Linz, a Diploma in Computer Science from UAS Mittweida, Germany, a Master of Business Administration from WU Vienna University and a Master in Arts and Business in Information Security Management from University of Applied Sciences Upper Austria (Campus Hagenberg). His research was published in Journals such as Technological Forecasting and Social Change, Journal of Business Research and Creativity and Innovation Management. During his extra-occupational studies, Johannes gained professional management experience in international IT companies such as Hewlett Packard. He is co-founder and CEO of a company, specialized in online community building and is the issuer of 3Druck.com, the leading online magazines of additive manufacturing in the German language area. Furthermore, he is the initiator of GameSurvey, a research project aiming at the application of gamification in research.