On formal ethics versus inclusive moral deliberation

Christian Herzog né Hoffmann¹

Received: 12 January 2021 / Accepted: 22 February 2021 / Published online: 5 March 2021
© The Author(s) 2021

Abstract

In this article, I will advocate caution against a formalization of ethics by showing that it may produce and perpetuate unjustified power imbalances, disadvantaging those without a proper command of the formalisms, and those not in a position to decide on the formalisms' use. My focus rests mostly on ethics formalized for the purpose of implementing ethical evaluations in computer science—artificial intelligence, in particular—but partly also extends to the project of applying mathematical rigor to moral argumentation with no direct intention to automate moral deliberation. Formal ethics of the latter kind can, however, also be seen as a facilitator of automated ethical evaluation. I will argue that either form of formal ethics presents an obstacle to inclusive and fair processes for arriving at a society-wide moral consensus. This impediment to inclusive moral deliberation may prevent a significant portion of society from acquiring a deeper understanding of moral issues. However, I will defend the view that such understanding supports genuine and sustained moral progress. From this, it follows that formal ethics is not per se supportive of moral progress. I will illustrate these arguments by practical examples of manifest asymmetric relationships of power primarily from the domain of autonomous vehicles as well as on more visionary concepts, such as artificial moral advisors. As a result, I will show that in these particular proposed use-cases of formal ethics, machine ethics risks to run contrary to their proponents' proclaimed promises of increasing the rigor of moral deliberation and even improving human morality on the whole. Instead, I will propose that inclusive discourse about automating ethical evaluations, e.g., in autonomous vehicles, should be conducted with unrelenting transparency about the limitations of implementations of ethics. As an outlook, I will briefly discuss uses formal ethics that are more likely to avoid discrepancies between the ideal of inclusion and the challenge from power asymmetries.

Keywords Artificial intelligence · Formal ethics · Inclusive moral deliberation · Power · Moral progress · Moral understanding

1 Introduction

Works dealing with computational approaches to ethics often correctly claim that ethics is “done at a low level of formality” [1]. This renders the programming of automated ethical evaluations highly non-trivial. There is already ample work attempting to tackling this challenge, see, e.g., [2–5]. Many of these pursue the laudable goal “to ‘make the world better’” and they commonly try to do so “with a formal basis” [1]. This perceived need for formal ethics is often accompanied by the premise that (partially-)autonomous systems are

required for making the world better, and that imbuing them with a sense of ethics is an associated prerequisite.

In this article, I want to challenge this premise by advocating caution against a formalization of ethics. Other works have focused on the object formal ethics may actually be required for, i.e., artificial moral agents (AMAs), and whether those are desirable, cf., e.g., [6, 7]. Arguments against and for AMAs have primarily focused on questions of moral responsibility [8, 9], human safety [10], alignment with moral norms [11] or public trust [12, 13]. However, these arguments, though valuable in their own right, typically picture the end of a long process towards formally implementing an automatic evaluation of ethics. In contrast, I will develop an argument by looking at formal ethics itself. As such, I consider formal ethics understood as (i) an approach due to Gensler [22] that aims at employing logical formulae for adding verifiable rigor to moral statements, and

✉ Christian Herzog né Hoffmann
christian.herzog@uni-luebeck.de

¹ Ethical Innovation Hub, University of Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany

(ii) computational approaches to ethical decision-making in artificial intelligence that make use of so-called top-down, bottom-up and hybrid approaches [31], i.e., approaches that follow rule- or principle-based ethics, algorithms that are trained from moral norms captured by data, or combinations of these, respectively. My main argument will extend to both kinds of formal ethics. As I will attempt to show, actual implementations of ethical evaluation by some formalism or another will increase the significance and societal dangers of power asymmetries.

I will attempt to show that formal ethics produces unjustified power imbalances, disadvantaging those without a proper command of the formalisms, and those not in a position to decide on the formalisms' use. As such, I will argue that formal ethics presents an obstacle to inclusive and fair processes for arriving at a moral consensus or preliminary forms thereof. Also, this may impede a significant portion of society from acquiring a deeper understanding of moral issues. I will argue that particular proposed use-cases of formal ethics run contrary to their proclaimed promises of increasing the rigor of moral deliberation and even improving human morality on the whole. For this purpose, I will first analyze promises of formal ethics put forward in the literature in Sect. 2.

Researchers and technology companies are already implementing means for ethical evaluation into devices, such as self-driving vehicles, cf. [14, 15], even though this may happen only implicitly or tacitly. The fields of formal and machine ethics thus represent a systematized approach to developments that already transpire. In light of formal ethics—or more implicit variants thereof—as a tool to implement automated ethical evaluation, this article is primarily concerned with practically existing asymmetric relationships of power, how they may undermine laudable goals in engineering and machine ethics, and how formal approaches to ethics may contribute their perpetuation.

For establishing a notion of 'inclusive moral deliberation' in Sect. 3, I will refer to Jürgen Habermas' ideal of discourse ethics [16]. I will attempt to connect this to Michel Foucault's idea of relationships of power, cf., e.g., [17, 18], as ever-present practical challenges. As such, power imbalances may become particularly worrisome, as moral thought could be dominated by a moral perspective that is restricted by some practical limitations of formal ethics or by improperly powerful entities. Therefore, in Sect. 4, some of the challenges emerging from asymmetries in power relationships in the domain of the automation of moral deliberation will be elaborated on. In doing so, I hope to contribute to the field of machine ethics by highlighting practical moral challenges for the attainment of the Habermasian ideal.

The significance of this ideal is, of course, debatable. In defense, I will argue for the importance of means to inclusive moral deliberation as a critical contributor to moral

progress in Sect. 4 as well. I will defend the view that moral progress is mainly due to an increased moral understanding, a notion due to Hills [19], that denotes the ability to translate deeply understood moral beliefs into practice in both deed and discourse. I will argue for the role of moral understanding as making a critical difference between technological and moral progress: While formalisms have undoubtedly facilitated technological progress, moral progress may not significantly benefit from formal ethics.

I will further elaborate on implications for formal ethics by referring to obstacles imposed on inclusive moral deliberation in principle and practice. The latter hinges on the claim that formal ethics rather perpetuates existing power asymmetries than abolishes them. I will illustrate this via the example of autonomous vehicles by referring to current trends in their development, observable asymmetries in relationships of power, and associated discursive elements, or rather the lack of inclusive ones. I will also briefly outline, how there is a danger that power asymmetries and the limitations of formal ethical evaluation risk failure in advancing human morality by so-called artificial moral advisors, see, e.g., [20]—a visionary concept apparently sincerely dedicated to making artificial ethical evaluation work for humankind.

In conclusion, I maintain that the utility of formal ethics is limited to applications, in which moral debate has reached a point close to consensus or in which devices powered by formal ethics act as modest facilitators of moral debate. The limitations of formal ethics in implementing moral consensus, in turn, should be transparently and inclusively debated.

2 The promises of formal ethics

Approaches to formalizing ethics mainly seem to fall into two categories concerning their proclaimed goals. The first category, or goal, denotes the attempt to improve the ethicality of human decision-making through increasing circumspection and reducing the ambiguity of statements about ethics. For instance, formal approaches to ethics should make ethics "more precise" [21, p. 56], "as hard as logic" [22, p. viii], or "provide more than vague, general criteria" [23] and "disambiguate natural language, [...]to provide computable meaningful results" [24, p. 42]. Associated tools should "improve human moral decision-making" [20], e.g., by employing formal logic. Beyond that, some support the view that humankind's alleged irrational, egoistical, unreflected, and gratification-desiring behavior can be held at bay by machines capable of ethical decision-making that may even become superior to that of humans [25].

A second category denotes attempts to provide frameworks for ethically aligning technological artifacts with society's moral consensus [26, 27]. In this context, formal

ethics should sometimes “provide an operationalizable, and presumably quantitative, theory” [23], i.e., make ethical evaluations amenable to computational frameworks. Self-driving cars may arguably be the first type of robot that will enter society at mass scale [28]. Its associated ethical dilemmas are variants of the well-known trolley problem [29], or more subtle questions, cf. [30]. Eventually, there is a need for engineers to be able to implement consensus on these moral issues.

In the following, I will try to analyze more specifically how different approaches attempt to achieve the goals above and the associated broader promises. For this purpose, I will consider two facets of formal ethics: (i) An approach due to Gensler [22] aims at employing logical formulae for adding verifiable rigor to moral statements. (ii) Computational approaches to ethical decision-making in artificial intelligence make use of so-called top-down, bottom-up and hybrid approaches [31], i.e., approaches that follow rule- or principle-based ethics, algorithms that are trained from moral norms captured by data, or combinations of these, respectively. A third facet denoting formal ethics associated with processes aimed at certifying the ethical alignment of corporate or research conduct will not be considered here.

2.1 Harry J. Gensler’s formal ethics

A particular approach to formal ethics is due to Harry J. Gensler, 1996, who proposes that a formal part of ethics is “as hard as logic” [22, p. viii]. This part, Gensler claims, can contribute to any traditional ethical theory because it presupposes no foundational views [22, p. 180]. Gensler claims to have constructed formal ethics using symbolic logic [22, p. 166] and that a formal ethical principle may be defined “as one that is expressible using this symbolism” [22, p. 5], principle “of inference expressible using only variables and logical terms”, or constants [22, p. 2]. In formal ethical principles, the constants may consist of attitudes, such as “believe”, “desire”, or “ought”. The variables denote agents or items that the verbs refer to. Gensler claims that, due to this construction, formal ethical principles—if valid—are largely uncontroversial except if their underlying formal logical principles are. In turn, he admits that metaethical controversies about the meaning and justification of formal ethical principles exist. However, Gensler claims that disputes about the conclusions following from the premises are rare [22, pp. 2, 5].

Gensler distinguishes formal from material ethical principles—on which normative ethics focuses—merely on the basis that the latter “contain concrete terms”, i.e., variables are being replaced with a specific action. For instance, “Do not kill” is a material ethical principle not amenable to

the tools of Gensler’s formal ethics to check its validity.¹ Gensler’s formal ethics is a language system with mechanisms for formal verification of whether a statement’s implications match those implied by expressions in natural language. Gensler seems to imply that natural language is rife with inconsistencies and often vague and, consequently, proposes that his approach is (i) a critical framework to expose unjustified moral intuitions or avoid their becoming sentimentalized, (ii) a tool to clarify vague concepts [22, pp. 11–12], and (iii) a way of systematizing principles that support the application of ethical theories, such as utilitarianism, to concrete examples [22, pp. 154–157].

There is a range of compelling objections against Gensler’s project. For instance, Vorobej [32], is charitable to Gensler’s project’s goals but doubts that a formal approach to ethics can remain neutral about the content of morality. Vorobej illustrates this by highlighting that vivid criticism on the notion of universalizability exists. However, in this article, I would like to attempt to advocate caution to the very project of formalizing ethics by calling into question that the proposed benefits can be realized, as just processes would require more inclusive, and hence, less formal ways of deliberating. I will elaborate on the meaning of this in Sect. 4.1. First, however, I will turn to a related, but ultimately quite different strand of approaches to formalizing ethics.

2.2 Approaches to computational implementation of moral deliberation

While Gensler’s project of devising formal ethics mostly translates the machinery of formal logic onto ethics, more recent approaches take formalization as a prerequisite of making moral deliberation amenable to implementation in artificial intelligence (AI) systems. The underlying rationale most often seems to be that automating moral deliberation is unavoidable as systems become increasingly autonomous, cf. [26]. Even though there are compelling objections to this rationale, see, e.g., [6], I will not argue against it here, but rather focus on why the promises of formal ethics will likely not come to pass. The scope of this article cannot cover all possible and current approaches to formalizing—and, for that matter, implementing—ethics. The selection of references, however, will capture relevant views.

¹ An example of a formal ethical principle central to Gensler’s project is the Golden Rule (or the principle of universalizability), expressed in prose as “Treat others as you want to be treated”. Gensler then proposes different formulations in terms of constants and variables, such as “If you want X to do A to you, then do A to X” [22, p. 12], only to reject them based on absurd consequences that derive from taking the formulation literally.

In contrast to Gensler, in formal ethics for automation, material ethical principles are mostly explicitly considered. The goal to reduce or even eliminate vagueness from ethical principles, however, is a common one. For instance, Conitzer et al. [23] write:

“To be useful in the development of AI, our moral theories must provide more than vague, general criteria. They must also provide an operationalizable, and presumably quantitative, theory that specifies which particular actions are morally right or wrong in a wide range of situations.”

Similarly, Polonski [33] goes further, claiming that ethics needs to become a fully quantitative theory. Bonnemains, Saurel and Tessier [24] set the goal in creating a new, unambiguous form of language amenable to automation.² More cautiously, Anderson, Anderson and Armen [21] restrict computational ethics to the domains, where experts are in consensus.

From Kantianism [34], or duty/principle-based ethics [35] to Utilitarianism [36], a range of approaches takes only particular moral theories into account. While these approaches may constitute valuable contributions to machine ethics research, they do not consider the problem of moral disagreement, i.e., the lack of agreement among moral philosophers (and society at large) about the correct moral theory (or moral norms). Any reliance on only a single moral theory implemented in computational frameworks is very likely to yield highly controversial decisions [37]. Due to this, I would like to focus on the issue of formalizing ethics more generally.³ Thus, for the remainder of this section, I will consider approaches of formal and computable ethics that aim at tackling the problem of moral disagreement. One such approach, due to Bogosian [38] explicitly addresses the problem of moral disagreement by taking into account moral uncertainty, i.e., the uncertainty of a particular theory to yield the “correct” moral decision.

In short, Bogosian’s approach uses the notion of Maximum Expected Choiceworthiness (MEC) [39], which averages the ethical evaluation of an action, i.e., the ‘choiceworthiness’, over different moral theories, weighting each theory by its ‘credence’, i.e., the likelihood of the theory to yield the correct action. Bogosian discusses both crowd and expert sourcing as approaches to acquire values for both the choiceworthiness and credence quantities. Similarly, Russel [11] states that to ensure AI remained beneficial to humanity,

formalisms of inference algorithms should be employed to estimate human objectives. He writes: “[...] the objectives we put into the machines have to match what we want, but we don’t know how to define human objectives completely and correctly.” [11, p. 170]. Russel’s underlying assumption is quite common to a range of approaches to implement ethics into machines: Humans are deemed inept at ethics or at accounting for the ethical implications of the goals they set. Machines will be less fallible.

I will argue that there is significant doubt that both categories of approaches to formal ethics may achieve their proclaimed goals. This may be due to significant practical and technical challenges, which would, however, only present inconclusive arguments. Therefore, I will not elaborate on these here. Nevertheless, a healthy skepticism with respect to technical aspects of the project of formal ethics may underscore arguments I will set out in Sect. 4. Instead, I will argue that if too powerful entities wield limited tools for automating moral deliberation, moral progress is in danger. For this purpose, I will first argue for inclusive moral deliberation as a requirement for moral progress.

3 Inclusive moral deliberation as a requirement for moral progress

In what follows, I will argue for the importance of means to inclusive moral deliberation as a pivotal contributor to moral progress, a notion which I will define later. This also hinges on the view that moral progress requires deeper moral understanding, a notion due to Hills [19] broadly construed as knowing *why*—instead of only knowing *that*—something is moral, which, in turn, is facilitated by taking part in moral deliberation. If correct, the utility of formal ethics for contributing to moral progress is constrained by its tendency to both disfavor inclusive moral deliberation and acquiring moral understanding.

3.1 Inclusive moral deliberation and power

Let ‘inclusion in moral deliberation’ denote the condition and processes required to provide the necessary opportunities and resources to practice moral deliberation in a nondiscriminatory and participatory way, i.e., open to all and inherently respectful of the diversity of all participants. Habermas developed the notion of the ‘ideal speech situation’ as an idealized model for arriving at consensus [17, p. 4]. In [16, p. 198] he writes.

“Argumentation insures that all concerned in principle take part, freely and equally, in a cooperative search for truth, where nothing coerces anyone except the force of the better argument.”

² Bonnemains, Saurel and Tessier, 2018 state that a “formal approach consists in defining a minimal set of concepts that is necessary to deal with ethical reasoning. A language is defined upon this set of concepts to compute ethical reasoning with automatic methods.

³ For an extensive overview on computational approaches to ethical decision-making, consider recent surveys, e.g., [3, 4].

Habermas connects his ideal to requirements that include that all parties affected take part, have equal possibility to contribute, and are capable of empathizing with each other. Importantly, power differences should be neutralized, and participants should engage with each other in transparent, rather than strategic or deceptive ways [16, pp. 65–66]. Obviously, these conditions do not come about by themselves, but require, e.g., meaningful regulation.

To defend this, it is perhaps illuminating to consider the difference between an inclusive and a kind of neutralist liberal notion of free participation in moral deliberation. A typical view of neutralist liberal citizenry is that one may freely pursue one's own interests constrained only by the principle that one may not infringe on the freedom of others. This rationale is often referred to as John Stuart Mill's, 'Harm Principle'.⁴ Taking this form of liberalism as an approach that "interprets democracy as a process of aggregating the preferences of citizens" [41, p. 19], proponents of liberalism would argue for the relatively unrestricted competition of individual moral norms and practices with processes similar to that of the marketplace. In this model, collective moral deliberation and, hence, implicit ways of arriving at moral consensus would result from the aggregation of individual moral deliberation turned into practice as well as a free expression of speech in robust debate guaranteed by neutral institutions. The metaphor often used for this is that of a 'marketplace of ideas', which, however, as argued by Ingber [42] is unrealistic, because it lacks awareness of perpetuated power asymmetries between opponents in discourse. Similarly, Awad [43, p. 49] argues that marketplace "procedures privilege the most powerful or 'competitive' interests" as "liberal freedom conflicts with policies to guarantee everyone's access to [...] democratic participation". Hence, liberalism following the doctrine of state neutrality typically rejects processes aimed at neutralizing power differences as demanded by Habermas.

The significance of power structures in determining, or at least influencing, the outcome of political debate has been showcased recently by revelations surrounding scandal associated with Cambridge Analytica.⁵ Irrespective of the fact that the company has been found guilty of acquiring data on U.S. citizens through unauthorized access to Facebook data, its business case rested on the promise that it can manufacture personalized advertisements to influence voter behavior [44]. The underlying technology relies on psychometric

analysis of social networking data to quantify personality trait parameters [45, 46] and aligning the message of online advertisements with these on an individual level. It remains doubtful how effective this kind of 'ad micro-targeting' is in influencing political votes [47]. However, the idea that money can buy political votes by producing personalized subliminal messages aimed at persuading unexpected citizens browsing the internet is clearly running contrary to the idea of inclusive debate, in which political opponents would meet at eye level. In this case, the enabling mechanism of economic fortune having the potential to decide debates resides in an unregulated market of personal data.

This real-world example illustrates that an approach to moral deliberation lacking any legitimate instrument to rebalance power may prevent challengers of moral norms established by a dominant party from having a fair chance of enacting change. In contrast, inclusive moral deliberation may be characterized by both extensive centralized and decentralized dialogic or multilogic exchange between individuals governed by just procedure and imbued with a sense of solidarity among participants [43, p. 50]. Details of such just procedures are non-trivial, and the present article does not attempt to propose any such framework. Rather, it attempts to illuminate the challenge to inclusive moral deliberation from formal ethics.

Habermas has been accused of being "overly naïve and idealistic" [17] in portraying an unattainable "ideal speech situation". Both his vision of discourse being decided based on "the force of the better argument" [16, p. 198] and his proposed ideal of power neutrality seem to be in conflict with reality, as his theory appears to lack a practical insight into the ways arguments may become powerful in virtue of being brought forth by powerful agents.

However, even though my attempt to deriving implications for formal ethics from inclusive moral deliberation is aimed at being practical, using the Habermasian concept as an ideal to strive for appears a good point of departure. In honoring this ideal, I intend to highlight practical obstacles, as both proponents and critics of formal ethics alike should not be solely tethered to high ideals and long-term visions but should also consider the pathways to achieving proclaimed goals. Therefore, it should be taken into account that consensus may, at times, be impossible and that moral norms may emerge from decentralized deliberation, involving also decentralized power relationships that may be extremely difficult to address procedurally. It also acknowledges the non-communicative forces that are at play. Due to limitations in space, however, I will mostly analyze practical impediments to formal ethics contributing to moral progress. Many of these require consideration of prevailing power structures.

Authors, such as Michel Foucault, have made relationships of power central to their ideas. In an interview, Foucault states that "[...] power is always present: I mean the

⁴ The 'Harm Principle' is often traced back to the following quote: "That the only purpose for which power can be rightfully exercised over any member of a civilised community, against his will, is to prevent harm to others." [40, p. 13].

⁵ For a brief introduction to the events that transpired, see <https://www.wired.com/amp-stories/cambridge-analytica-explainer/>, accessed July 29th, 2020.

relationships in which one wishes to direct the behavior of another.” [18, p. 122]. In Foucault’s terms, normalizing pressures on individuals derive from (moral) codes that contribute to asymmetrical power structures [48]. The presence of power per se, in turn, is unavoidable. The key seems to be that “universals must be questioned” [17] to counter the construction of potentially oppressive systems.

Thus, if we accept that defining what is right and what is wrong is a way (perhaps the ultimate way) of exerting power, we should eliminate power imbalances in the processes that potentially generate new ones. These are precisely the processes that should govern moral deliberation. However, it may be objected that there can be no complete equity of power, simply because power is also a necessary component for protecting a particular status quo. This may be as explicit as police having the actual physical power to enforce the law, but it might also entail the formation of political groups venturing to maintain a strong following among citizens. However, in the defense of this objection to the Habermasian ‘ideal speech situation’, I maintain that by upholding inclusion in moral debate, we would not deny that power may be a necessary element. Instead, promoting inclusive moral debate is a plea to abolish power asymmetries as much as possible primarily when in debate about moral values, and it is not denying that consensus about agreed-upon moral norms may require some way of enforcing them, though, of course, through just means. True danger from power asymmetries incurs, when power is utilized to circumvent or shortcut moral debate. Thus, at the very least, moral issues should be debated with only a minimum of assertions about what constitute foundational moral truths to aid in allowing more far-reaching premises to be challenged. Hence, and as elaborated on above, inclusion does not merely mean participation but also requires a willingness to engage with others in an open-minded way.

Accordingly, the notion of inclusive moral deliberation must extend beyond moral debate and must also include consideration for relationships of power on both a macro and micro scale. Consequently, it should also be referring to the nondiscriminatory and fair possibilities of moral agents to participate also in decentralized processes of deliberation.⁶

⁶ A starting point for decentralized deliberation is moral deliberation that happens in dialogue. For instance, Gardiner [48, p. 31] maintains that from the idea of ‘ethical dialogism’ as argued for by Bakhtin [76] follows a distinctly respectful and empathic idea of ethics:

“Dialogue as it occurs within the everyday lifeworld establishes a relation of mutuality, shared responsibility (or answerability), and unsolicited concern between human beings that supersedes the dictates of a systematic or formalized morality.”

Philosophical ideas similar to such ‘ethical dialogism’ [77] have already transpired into the mainstream. For instance, the idea of ‘non-violent communication’ [78] is a method focusing on compassion and collaboration in conflict resolution that respects differing views.

This process entails not only dialogic and multilogic debate, but also the freedom for individual deliberation as well as acting upon the conclusions. As seen from the Cambridge Analytica scandal, however, to guarantee just procedure in moral deliberation may require regulation, especially given the possibilities of novel technology. Further, contextualization is relevant to inclusive moral deliberation, which means a reevaluation of agreed-upon moral norms and a recommencement of deliberative processes should circumstances change.⁷

Having now discussed inclusive moral deliberation as an ideal close to Habermas’ ‘ideal speech situation’ but imbued with a Foucauldian sense for the relevance of relationships of power in practice, I turn to ‘moral understanding’ as a valuable precept for later analyzing implications to formal ethics.

3.2 Moral understanding as a distinct way of progressing

In this article, I will assume that a notion of moral progress, such as the one proposed by Moody-Adams [49] exists. She writes:

“Moral progress in belief involves deepening our grasp of existing moral concepts, while moral progress in practices involves realizing deepened moral understandings in behavior or social institutions.”

Also, Buchanan [50] offers a concept of moral progress that acknowledges plurality in moral concepts, a realistic consideration of human fallibility in compliance with moral norms, and a need to revise an understanding of what constitutes an improvement over time. Such a perspective aligns well with the notion of inclusive moral deliberation, as it

⁷ In this article, I am addressing formal ethics and its proclaimed promises also in current contexts of application, such as autonomous driving. As it stands, the ability to decide on an autonomous vehicle’s behavior a priori even in dilemmatic situations by virtue of applying formal ethical frameworks in automation is challenging both legal and moral norms. A good example for contextualization is provided by the German ethics commission’s recommendations for handling situations in which a driving algorithm would have to choose between harming different pedestrians [65]. In German constitutional law, offsetting the lives of the innocent against other potential victims is impermissible. However, the ethics commission challenges this principle in light of a preprogrammed algorithm, because it may be in the interest of everyone to minimize the risks of road injuries, before becoming an actual subject considered in a potential offsetting. The commission admits to not having been able to provide a final recommendation. This is creditable as the commission acknowledges the need for further moral deliberation by refraining from declaring these questions to be settled by virtue of their power from being perceived a moral authority. Even though one may desire clearer answers, moral issues may require broader, more inclusive deliberation in various contexts.

aims at facilitating the correction of errors by being open to challenges. As such, means to inclusive moral deliberation promotes what Buchanan, denotes by “meta-moral progress”, i.e., “moral progress in the means by which moral progress is achieved.” [50]

Thus, a society’s moral progress may be understood as the dissemination of both a deeper understanding of *why* a certain moral standard is superior to a previously held moral belief as well as respective practices among a significant portion of members of that society. However, we are not only interested in the consequences of improved moral norms—which may, e.g., consist of reduced suffering—, but also in the mechanisms that perpetuate this progress.

Alison Hills [19] has elaborated on the notion of moral understanding. She attempts to distinguish knowing *that* from knowing *why* an act is moral or immoral. Denote the former moral knowledge. The latter constitutes moral understanding. While simple testimony about facts, e.g., offered by experts, may lead to an increase in knowledge, both moral and non-moral, Hills contends that it may not yield understanding in the moral domain. Hills also characterizes moral understanding as the ability to generate new true moral beliefs, which is an achievement “that can be credited to you” [19, p. 102], but can be done collectively through discourse, or even by accepting advice.⁸

There is debate about whether moral understanding is simply a form of descriptive knowledge, see, e.g., [51]. A core argument rests on conceiving of understanding merely as knowing why a certain reason really is a reason. Hills evades by arguing that moral understanding has value irrespective of whether it is conceived of as a form of knowledge. For instance, even if moral understanding would just be a form of *knowing how*, simple testimony about the *how* does not typically put the recipient into the position to exercise the testimony’s subject matter. For instance, you can tell someone how to drive a car, but it will require several attempts to put this knowledge into practice as a habitual skill. Thus, even if moral understanding is a species of knowledge, for the purposes of this article, it is sufficient to take moral understanding as a way of putting this knowledge into effect, e.g., by actively engaging in moral discourse. Similarly, Hills [19, p. 121], posits that developing moral understanding, even from moral testimony, requires practice—a concept not unlike Aristotle’s notion of habituation,

cf. [59, NE II.1, 1103a14–19]. Thus, moral understanding also puts knowledge into effect by aiding in internalizing and forming moral habits.

Hills is not opposed to accepting moral testimony *per se*. Instead, she distinguishes between simply believing or taking it as advice to ponder. In other words, moral understanding can be practiced and developed when taking moral testimony as advice, but simply accepting it—even if correct—incurs the danger of acting for the wrong moral reasons later on, even if the action itself is right.

Concerning the difference between non-moral and moral testimony, Hills argues that “it is far from obvious whom to trust about moral matters.” [19, p. 125] This allusion to moral disagreement highlights that there exists a difference between using formalisms in sciences, such as mathematics or physics, and ethics for achieving progress for at least three reasons: (i) It is much easier to know whom to trust concerning formalisms in STEM⁹ research, (ii) applying wrong methods onto novel applications will typically yield more obviously wrong results in technology, and (iii) evaluating technology in a non-moral sense is a task involving only a few, while just about everyone is required to do moral evaluations.

Reason (i) simply follows from the unsettled controversies about moral expertise, e.g., about what it is and whether it exists, as well as the problem of moral disagreement itself, cf. [53]. Reason (ii) indicates that disagreement about moral reasoning is a more significant issue than disagreement about technological methods and explanations. Evaluating approaches in technology based on the achieved effect is much more appropriate, while moral consequentialism is much less widely accepted. In technology, it is also considered superior to achieve some effect with the right method in the sense that this method may involve a model that constitutes a good approximation of the truth and is well understood. However, it will be more evident if a wrong method does not translate well to other applications and therefore fails to achieve the intended effect. Moral judgments, however, will not be so obviously wrong, requiring greater scrutiny of the reasons or principles by those that deliver the judgments. Finally, (iii) again links the idea of moral progress to an ideally pervasive adoption of the right moral reasons for arriving at the right moral judgments. While the science and methods behind technology are only rarely carried out by the consumers or users of that technology, everyone is involved in moral matters.

In summary, I have argued that moral understanding, even if taken merely as the ability to put moral knowledge into practice, is a prerequisite for, or is at least facilitating, moral progress, both on an individual and a society-wide level. In

⁸ More formally, take p to be a moral belief and q as its reason. According to Hills, moral understanding involves having several abilities to a certain extent that together amount to “treat q as the reason why p , not merely believe or know that q is the reason why p ” [19, p. 102]. This includes being able to explain p in your own words, conclude p from q , or infer q from the information that p . Crucially, it also involves the ability to conclude p' from q' , where p' and q' denote similar beliefs or reasons, respectively.

⁹ Science, technology, engineering and mathematics.

turn, technological progress is more likely not to require a kind of understanding on an individual level. This will turn out to be relevant in Sect. 4.3, in which I will portray formal ethics as not being supportive of moral understanding and hence moral progress.

Even if a distinction between moral and technological understanding would not withstand further scrutiny,¹⁰ evidence points to the importance of understanding formalisms in the technological domain as well. Consider the financial crisis of 2008. Financial trading algorithms based on the formalisms of machine learning and big data were opaque to scrutiny for at least two reasons: The algorithms were proprietary and hence held intentionally secret, and the utilized formalisms are considered complex insofar as even experts fail to understand the causality behind them [54]. Algorithms of the latter sort—e.g., deep learning and statistical data-based approaches typically dubbed “black-box algorithms”—are also powering many of the bottom-up approaches in machine ethics¹¹ [31], trying to formalize processes of imbuing autonomous systems with a sense of moral norms.

We are now in a position to recapitulate: Inclusion in moral deliberation considers approaching the ideal of Habermas’ discourse ethics in both actual debate and decentralized moral deliberation by addressing relationships of power. Moral understanding, in turn, seems to be valuable in acting against undue power imbalances in moral deliberation by endowing participants with the ability required to challenge moral consensus. Arguably, inclusive moral deliberation and moral understanding promote moral progress. In what follows, I will draw implications for formal ethics.

4 Implications for formal ethics

In this section, I will first propose the view that formal ethics poses an obstacle to the ideals of inclusive moral deliberation. Second, I will attempt to go beyond this principled analysis and show that formal ethics is promoting and perpetuating power asymmetries existing in current contexts associated with autonomous systems. Third, I will argue that formal ethics does not support the development of moral understanding and may not present a viable path forward to equalize asymmetric relationships of power.

4.1 Formal ethics as an obstacle to the ideals of inclusive moral deliberation

I will first illustrate ways formal approaches to ethics can present an obstacle to inclusive moral deliberation in principle. Recall that the ideal can be outlined as requiring that all parties affected (i) take part, (ii) have equal possibility to contribute, (iii) are capable of empathizing with each other, as well as that (iv) power differences should be neutralized and (v) participants should engage with each other in transparent ways, cf. [16, pp. 65–66].

There are at least two different facets to formal ethics as an obstacle for inclusion in moral debate relating to the above criteria: Formalisms in ethics (i)–(ii) may present a technical, as well as a social barrier to participation in moral discourse with an equal possibility to contribute, (iii), and often have as their stated goal to “remove sentimentalism” and may thus dissuade participants from empathizing. I will turn to (iv)–(v) in Sect. 4.2, where I will show that formal ethics may perpetuate power asymmetries that hinder scrutiny and transparent debate.

4.1.1 Technical and social barriers to participation

Concerning the criteria (i)–(ii), formal ethics presents a barrier to entering moral discourse to those without a proper command of the formalisms. Formalisms are challenging because they involve a significant knowledge of elements not directly associated with ethics, such as symbolisms or mathematics.

Requiring certain educational attainments may seem unavoidable, however. Proper command of the natural language is also a prerequisite for, e.g., moral discourse.¹² Asymmetry in discussion techniques also exists between those well-educated in the humanities and those that have problems articulating their contributions. Formalisms or a group membership associated with particular skills may also present social barriers to participation. The notion of ‘moral power’ as the “degree to which an actor, by virtue of his or her perceived moral stature, is able to persuade others” [55] is an important one to consider when striving for the ideals of inclusive moral deliberation. Consequently, I maintain that neither a lack of technical abilities in formal ethics nor of education in moral philosophy should present barriers to moral

¹⁰ Especially, since this article deals with technological approaches to moral evaluations.

¹¹ Cf. to Sect. 2.2 for a brief explanation of top-down, bottom-up and hybrid approaches to machine ethics.

¹² Inclusion may, of course, also refer to the promotion of ways in which people with disabilities can participate. In this case, natural language in spoken or written form may also constitute a barrier and ways to decrease this barrier are required. I do not want to exclude that this is a highly relevant issue. However, in the following, I will not explicitly address the requirements of inclusion for letting mentally or physically challenged citizens participate. I consider the need for methods for inclusion of this kind self-evident.

discourse. The point then is that, even as a trained ethicist, moral arguments need to be advanced with meekness. This means that moral discourse demands to allocate time and patience, including an articulation of refutations in ways the respective addressees will comprehend. It also requires an awareness of the power emerging from authority as tacitly perceived by laypeople.¹³ Expertise in ethics, hence, should not be about moral authority but about being committed and able to deliver a transparent, intelligible, and well-balanced exposition of arguments. These should be typically conveyed in accessible natural language. Thus, even if formal ethics achieve to increase the rigor of some approach to ethics, see, e.g., [22, p. 180], inclusion in moral debate demands to re-translate the results into comprehensible and inclusive natural language.

In moral argument, formalisms rather seem to obscure the content of the proposed solution. A case in point is the formalism proposed by Bogosian, see [38]. The parameters that enter his computational framework to balance competing moral theories would at best function as translating moral consensus into operational computer code. Contrary to this, Bogosian argues by utilizing his formal framework directly, claiming that “[t]his procedure, however ad hoc, seems to be the best possible way of approaching this particular case”.

4.1.2 Barriers to empathy

A common conception in formal approaches to ethics often seems that human emotions lead moral judgments astray. For instance, Gensler writes that “[e]ven when the logic is clear to us, we might be unable to be consistent because of psychological defects or strong emotions.” [22, p. 19] Anderson and Anderson claim that “[h]umans are prone to getting ‘carried away’ by their emotions to the point where they are incapable of following moral principles.”¹⁴ [56]

The notion of inclusive moral deliberation, which borrows from Habermas’ discourse ethics [16], stresses a need for enabling participants to engage in discourse committed to non-deceitful and open-minded exchange of arguments. Concerning criterion (iii) above, participants should also be “capable of empathizing”, meaning that they should both have the capacity and resources to develop empathic

reactions, e.g., time and information. Thus, if the ideals of discourse ethics are acknowledged, this requires *humans* to take part in moral deliberation, as these capacities seem to be uniquely human (as opposed to artificial beings). Irrespective of humanity’s ability to design artificial moral agents that may eventually turn out to be in some way superior to humans in terms of their capabilities for moral deliberation or even their capacity for empathy, such artificial agents would face the same demands for letting those affected, including humans, participate inclusively.

However, current formal approaches to ethics of the top-down, rule-based type,¹⁵ such as Gensler’s, at least, do not appear to facilitate human empathy or to allow for space and time in which participants may relate to each other. Visionary concepts, such as the artificial moral advisor due to Giubilini and Savulescu [20] e.g., instead promise to facilitate alignment of one’s behavior to one’s own moral standards even under the pressures of time. The focus, hence, seems to rather consist in increasing the time efficiency of moral deliberation, rather than its quality. Similarly, bottom-up approaches to computational ethics do not promote that participants may empathize with each other in inclusive moral debate as well. This is because bottom-up approaches do not formalize ethics itself, but rather focus on formalizing the process of arriving at a machine capable of producing moral evaluations from data. The purpose, even if the outputs of the machine act only as suggestions to humans, amounts to the automation of part of the deliberation process to increase the time efficiency of moral deliberation, even though algorithms may take data on moral judgments that involved emotions and empathizing as inputs. In light of the ideals of discourse ethics, such increases in efficiency need to be well-argued for, if the typical casualty of this kind of rationalization is the time and effort put into empathizing with others.

Modern technology indeed raises the issue that decisions with salient ethical implications can be—and thus potentially also have to be—taken in fractions of a second. Autonomous driving once again illustrates this. There is little controversy that self-driving cars are a technology that, in principle, should be developed, since the amount of accidents is typically assumed to decrease significantly with its adoption [57].¹⁶ Still, the debate on how self-driving cars should operate in dilemmatic or even mundane situations remains far from settled [30]. However, novel arguments on ascribing responsibilities bring in notions of subjective

¹³ In fact, open online consultations on ethical guidelines, such as those on the ethics of AI by the UNESCO or the European Commission, are common practice. See <https://en.unesco.org/news/unesco-launches-worldwide-online-public-consultation-ethics-artificial-intelligence>, accessed July 29th, 2020, and see <https://en.unesco.org/news/unesco-launches-worldwide-online-public-consultation-ethics-artificial-intelligence>, accessed July 29th, 2020.

¹⁴ Researchers on artificial moral agents have already begun to question the usefulness of emotions also in the context of artificial moral reasoning, typically with the result that it may not be necessary to mimic the function of emotions in humans, see, e.g. [79, 80].

¹⁵ For a brief explanation of top-down, bottom-up and hybrid approaches to formal ethics cf. Sect. 2.2.

¹⁶ There remains doubts and speculation about the actual extent of decreases in road fatalities and accidents, though, cf. [81, 82].

user experience [58], which seems challenging to do using existing formal approaches to ethics.

Still, facilitating empathy in moral discourse by formal approaches may be possible. Currently, however, even in approaches that are not directed mainly towards enabling computational implementations, empathy does not seem to play a significant role.

Approaches to automating moral deliberation, in turn, will need to accommodate implementing moral consensus. Perhaps, such a consensus can be implemented as a kind of law without regarding the empathic actions that occurred during the inclusive moral deliberation. In that case, this law probably needs to be quite specific. If, however, such a level of specificity should turn out infeasible, moral consensus needs to be formulated in ways that can be extrapolated to a range of circumstances. This, in turn, might require the formal ethics underlying the implementation to incorporate empathy, which some researchers are already working on, see, e.g., [59, 60].

Extrapolations of agreed-upon moral norms onto other contexts incur the danger of being opaque, effectively concealing the moral saliency in potentially highly relevant applications. This leads to aspects (iv–v), by which formal ethics may hinder a challenging of moral norms. Such a possibility to challenge within a framework of inclusive moral deliberation, however, is necessary for not impeding moral progress.

4.2 Formal ethics as a promoter of power asymmetries

Having shown that formal ethics pose obstacles to ideals of a Habermasian notion of inclusive moral deliberation in terms of impeding opportunity to equally participate and empathize in moral debate, I will continue to consider the further requirement to neutralize power differences and to engage in transparent discourse. I will frequently draw on the example of autonomous driving to highlight practical considerations when addressing power asymmetries. Autonomous driving is a salient use-case, as the debate on its ethical implications persists simultaneously to efforts of large multinational companies to enter markets as early as 2025 [61].

In Sect. 4.2.1, I will argue that there already is evidence for asymmetry in power relations regarding automation technology, as companies can dominate the introduction of ethical principles due to their largely unrivaled position to produce both the machines and the code for automation. Even before society-wide moral discourse could have settled on moral consensus and regulations, precedents are made that constitute a *de facto* moral decision. The mechanisms that guide success in these cases are not necessarily those of the most convincing moral argument but may instead reside in the economic benefit.

Second, in Sect. 4.2.2, I will advance the view that technological products employing formal ethics could result in a central ethical standard perpetuated by decentralized devices. The formalisms undergirding these standards will likely be limited in how they can reflect the variety of stances in moral discourse. Accordingly, these limitations should be explicitly addressed in inclusive moral deliberation.

4.2.1 Accumulation of power in the technology industry

The autonomous driving industry is already beginning to implement features without important ethical questions being settled. For instance, Elon Musk, writes that “it would [...] be morally reprehensible to delay release [of the autopilot functionality] simply for fear of bad press or some mercantile calculation of legal liability” [62]. Still, according to Mider, “Musk described Autopilot as a kind of rough draft, one that would gradually grow more versatile and reliable until true autonomy was achieved.” [63] Musk’s statement came only months before the first publicly reported fatality due to the self-driving system and only a few days before a software update that Tesla claimed would have prevented the accident [64]. Accidents can never be averted with complete certainty. However, irrespective of whether one disagrees with Tesla’s trial and error approach, Tesla is an example of an entity that uses its power over both production and formalisms to pursue a particular ethical stance on autonomous driving that, in essence, amounts to simple utilitarian calculus. Instruments of power that the people can wield against such a unilateral perpetuation of moral views are limited. They mostly rest in the market or by taking the opportunity to vote for stricter regulation in upcoming elections. Both these instruments, however, are only of a post-hoc nature, as it is currently within the company’s power to set the pace by bypassing actual moral discourse.

For instance, Tesla has announced that from 2016, all cars will have the hardware necessary to implement autonomous driving [14]. This announcement amounts to one of the most important technological proponents of autonomous driving unequivocally deciding that, in terms of sensory input and computational capacities, nothing more is needed to allow for ethically aligned autonomous driving. Note that this happened even before probably the first preliminary set of ethical guidelines on autonomous driving had been published by a national governmental agency, cf. [65]. The hardware and software in their cars put Tesla in a position to determine the self-driving algorithms at will via “over-the-air” software updates [64], almost simultaneously affecting its current global fleet of more than a million cars [66]. As Mider puts it, “Musk’s decision to put Autopilot in the hands of as many people as possible amounts to an enormous experiment, playing out on freeways all over the world.” [63]

While the capabilities of current autonomous driving systems may not yet have reached a level that allows certain controversial moral evaluations to be made both in dilemmatic or more mundane situations, likely, this may soon be the case. Despite this, moving forward in implementing autopilots without having the hard- and software capacity to address ethical quandaries rigorously can be viewed as a particular stance on the moral debate as well: detailed ethical considerations are considered of low importance, while the company controls the message that safety is key.

Even if one agrees in principle, the details of Tesla's autopilot are a company secret, which excludes the public from scrutinizing an important moral matter that affects large parts of society [67, 68]. Instead of merely presenting a technical or social barrier, as argued in Sect. 4.1.1, here, the application of formalisms to ethically salient domains can even be protected as a trade secret [69].¹⁷

In defense of the companies pursuing automation in morally salient domains, e.g., market pressures may be the reason why there is a lack of incentives to engage in inclusive moral discourse first. Perhaps inclusive discourse would not even require a full revelation of trade secrets. One might argue that competition drives innovation, and the earlier autonomous vehicles are marketable, the more lives and money will be saved due to reduced accidents. Instead, it might suffice to lay open qualitative explanations, e. g., on how an autonomous vehicle would act presented in a widely comprehensible form. However, it may only be hoped that the company with the best package in terms of value and ethical considerations would emerge victoriously. In light of my earlier attempts to contrast a liberal marketplace against an inclusive approach to moral deliberation in Sect. 3.1, this would at least be closer to the demands of the latter, as the details on the moral debate would be made more transparent. However, a free market would still lack just procedures that guarantee that moral matters can be deliberated inclusively without the interference of existing power structures. On the market, such power structures could emerge from the simple fact that capital is accumulated based on economic success with products that are quite remote from being associated with the moral matters to be deliberated. This route to economic power is exactly what Musk laid out in his agenda, cf. [62].

¹⁷ In fact, such practice is not limited to the autonomous driving industry. Large multinational companies and smaller enterprises are engaging in similar endeavors of pushing forward automation in morally salient domains. As with autonomous driving, this may often happen with laudable intentions. For instance, IBM's Watson for Oncology is a recommender system aimed at improving outcomes for cancer patients but notorious for having been trained on too little and fabricated data, which led to the automated suggestion of contraindicated drugs [1, 83].

In these matters, it is not formal ethics per se that creates or perpetuates asymmetric relations of power, which are already strong in the technology industry [70, Ch. 1]. However, with formal ethics as discussed in Sect. 2.2 being highly reliant on the technology industry, and as a presumed facilitator of the trend of autonomous systems that offer sweeping promises of increased comfort and safety, it may be complicit. As I have tried to illustrate above, powerful technology companies venture into automation in morally salient domains without adhering to inclusive deliberation. By virtue of their access to the resources to implement ethical frameworks, they unilaterally push their moral agenda relatively uninhibited by challenges.

New ideas are needed to decouple economic power from the power to determine how moral issues will be addressed in the implementations of products. As long as this is not the case, advancing formal approaches to implement moral deliberation or even a mere alignment to moral norms, presents the danger that existing relationships of power (economic or otherwise) may dominate the processes that determine moral norms within society. For instance, impartial bodies would need to provide evidence that the qualitative explanations do indeed match the implementation based on formal ethical frameworks.

4.2.2 Implications of perpetuating moral norms through limited formalisms

As considered in the previous section, formal ethics are an instrument of power that can be exploited to perpetuate particular moral norms. In this section, I will consider ways in which formal ethics may perhaps unwittingly limit the breadth of the moral norms in practice. Instead of arguing against implementing formal ethical evaluations per se, I will argue for an explicit consideration of these effects and, ultimately, of the limits imposed on moral norms that autonomous systems may be aligned to. If a particular moral norm is tacitly favored by some formalism, even if it would qualify as “good” by some measure, centralized implementations may constitute an accumulation of power that contradicts the notion of inclusive moral deliberation.

To illustrate, consider that even though today's mobile devices are highly decentralized, software updates may change the behavior of millions of devices almost simultaneously, creating a centralized system in effect. Himmelreich denotes the emergence of significant patterns from the aggregation of identical programming code in millions of devices as the ‘challenge of scale’ [30]. Analogously, identical ethical formalisms in many devices may replicate a particular moral view. They also replicate the limitations of the formalisms required to program the devices. These limitations may turn out to be unavoidable, as it is likely that no formal approach to ethics can encompass all types of ethical

considerations. There may also be a mismatch between society's moral consensus and what an ethical formalism may allow to implement.

Consider the case proposed by Himmelreich in which both an autonomous vehicle and a pedestrian are approaching a pedestrian crossing [30]. Such situations may be imbued with high uncertainty. Humans intuitively judge whether they are being noticed, what others might be about to do, and the likelihood of particular events, such as coming across pedestrians at a specific daytime. In autonomous cars, both limitations in sensor hardware and the implementation of ethical evaluations restrict what can be taken into account. These limited capabilities of both formal ethics and hardware in an autonomous car should be accounted for in an inclusive moral deliberation of how an autonomous car should react in given situations, precisely because they will yield systematic implications in the aggregate.

Objections to this demand might again stem from the argument that advancing autonomous cars immediately will save numerous lives.¹⁸ The benefit of saving many lives may ultimately outweigh certain reservations. However, by advancing without far-reaching testing and inclusive debate about the results, autonomous cars may also accidentally but systematically be putting particular traffic participants at risk that, previously, were under a far lower risk [71]. In some moral views, this may amount to offsetting some lives against others, at least statistically [65]. Despite the proclaimed benefits of automated driving, I maintain that this inherent trade-off would need to be decided based on inclusive deliberation, rather than ad hoc based solely on the current technical limits to ethical evaluations. Simulations, cf. [72], could shed some light on these issues and may be a valuable instrument to guide more inclusive discussions. In any case, the mere unproven prediction of fewer accidents should not be the sole enabler for obtaining a license to deciding on the formal ethical evaluations employed in autonomous driving.

Indeed, inclusive moral deliberation could avert the danger that people might stop caring. They might stop being interested in the circumstances under which their self-driving car will swerve, slow down cautiously, or overtake daringly. Even when each individual may not notice, in the aggregate, subtle changes in the self-driving cars' behaviors will be significant. The danger lies in this significance eluding public scrutiny and inclusive moral debate if powerful entities, such as the companies may unilaterally decide on these matters. However, given current power structures, heavy reliance on mobility, and barriers for the general public to scrutinize the formal aspects of autonomous driving

software in both ethical and non-ethical means, this may need to be enforced by means of regulation.

Power issues may be relevant in autonomous driving but are even more salient in the proposal of an artificial moral advisor [20]. According to Giubilini and Savulescu [20] such a device would aid an individual to better comply with her or his own moral beliefs by presenting advice and relevant data. Even if the aim of such replicated devices is to facilitate alignment with the user's individual moral norms, they may turn out to perpetuate the problems that they try to solve. This will be the focus of the next section.

4.3 Formal ethics as an obstacle to moral understanding

In the previous sections, I have attempted to argue that formal ethics is both an obstacle to inclusive moral deliberation as well as that it perpetuates existing power asymmetries. In the following, I would like to refute an objection that may be raised against working towards inclusive moral deliberation, especially under power neutrality from a practical point of view: It may be argued that power asymmetries and a lack of inclusion may be tolerated to promote progress. It may be objected that the ideal of inclusive moral deliberation may ultimately impede both technological as well as moral progress¹⁹ because it slows down innovation. Progress lies at the core of the promises put forward by both formal and computational approaches to ethics, cf. Sect. 2. Formal approaches to ethics sit at the intersection of technology and moral philosophy. Hence, it may be suggested that similar to how formalisms have advanced technology, formal approaches to ethics will promote moral progress.

However, as argued in Sect. 3.2, there is a difference between technological and moral progress, which mainly resides in the latter requiring moral understanding to permeate society. Hence, in the following, I will first summarize the argument that formal ethics do not facilitate moral understanding because it presents an obstacle to inclusive moral deliberation. Second, I will caution against formal ethics that proclaim ways of improving human morality by elaborating on the notion that genuine moral progress on both an individual or a society-wide level demands that moral agents understand *why* something is morally right or wrong.

¹⁸ This may well be true, even though actual statistical evidence on this issue remains limited [2–5].

¹⁹ In the following, I will again adopt the definition by Moody-Adams [1] and Buchanan [6–9] as presented in Sect. 3.2, and paraphrased here for the convenience of the reader: In brief, moral progress according to Moody-Adams [6–9], involves deepened moral understanding and following up on this by appropriate behavior. Likewise, Buchanan [2–5] stresses that what amounts as moral progress should allow for a pluralistic view on moral norms.

4.3.1 Formal ethics do not promote moral progress

In Sect. 4.1, I have argued that formal ethics, as a method proclaimed to add rigor to moral deliberation, presents both technical and social barriers to participating in moral deliberation. I have continued to argue that computational approaches to moral deliberation do not facilitate empathizing with those concerned in the deliberative process. Further, I have argued that formalisms present instruments of power. In Sect. 4.2, I have tried to elucidate asymmetric relationships of power as a practically relevant obstacle.

I believe these aspects together show that formal ethics is both not supportive of inclusive moral deliberation as well as presents actual instruments to perpetuate existing power asymmetries. Then, in turn, if the concepts elaborated on in Sect. 3 hold some truth, formal ethics have the potential to impede moral progress by preventing moral understanding as a consequence of a lack of inclusive moral deliberation.

Note that I intend inclusive moral deliberation to not only concern actual debate but also to consist in the aggregation of decentralized dialogic or even individual acts of deliberation. As per Hills [19], the continuous exercise of such deliberation—turned into new moral beliefs and practices—amounts to proper moral understanding. If moral deliberation is not inclusive, a significant portion of society will be bereft of developing moral understanding. As a consequence, this portion will not contribute to moral progress as defined by Moody-Adams [49] who considers moral progress to follow precisely from deepened moral understanding and turning this into practice.

It may be objected that it is sufficient that citizens simply abide by superior moral norms to result in progress and that this may be facilitated by devices. Thus, it may be claimed that this issue is simply definitional. A full defense of defining moral progress as something more than “better compliance (not mere conformity) with valid moral norms” [50] is beyond the scope of this article. However, it seems plausible that a proper definition should account for the plurality of moral views and the constant change in what we regard as valid moral norms. Thus, particularly in moral reform, it appears more essential to make moral progress reliant on our ability to argue for the validity of some moral norm in terms of the currently most defensible arguments. The notion of moral understanding as per Hills incorporates this ability to explain and exchange moral arguments. Moral progress may hence be thought of being facilitated mostly by decentralized dialogic and multilogic exchange that both proliferate and challenge moral norms, rather than by coercing individuals into behavior of a higher moral standard.

I concede that formal ethics can spark a fruitful academic debate, whose results may eventually enter public discourse as accessible popular science. However, actual moral progress requires the wide-spread adoption and understanding

of potentially superior moral norms. In contrast, wide-spread technological understanding may largely remain related to its application, not its development and proliferation, to result in progress.

It may be an overly crude analogy, but one may hardly expect one’s sense of orientation to improve, while one is busy reading magazines in a self-driving vehicle. Likewise, I suggest that moral deliberation must be practiced both individually and in inclusive exchange with others. A good sense of direction may be a dispensable skill. Moral deliberation, however, is not.

Still, there are proposals to use formal approaches pursuing the laudable goal of attaining overall moral progress. If there is a use-case for formal ethics that can rebut these worries, it would be a proposal for systems that promote moral understanding. However, I would like to argue that artificial moral advisors, as proposed by, e.g., Giubilini and Savulescu [20] and Anderson [73], do not promote moral understanding and, hence, moral progress.

4.3.2 Artificial moral advisors do not promote moral progress

Next, I will argue that the notion of moral understanding contrasts with the idea of employing formal ethics to devise technological means for improving compliance of individuals concerning their own moral standards or society’s moral norms.

For instance, the idea of an artificial moral advisor is proposed to guide its user on the morally right action [20]. The authors state that such a system “would assist us in many ethical choices where—because of our cognitive limitations—we are likely to fall short of our own moral standards.” In essence, a kind of artificial moral enhancement is proposed, which, according to Anderson [73], should lead to humans being more reflexive, less egoistic, and following good role models. Giubilini and Savulescu suggest to achieve this by improving or even substituting the function that human emotions play in supporting quick moral judgments by technological means. They envision a system that processes information and compiles it into moral advice more thoroughly than a human could. In doing so, the system should approach the ideal observer as per Firth [74], i.e., be “(1) omniscient with respect to non-ethical facts, (2) omnipercipient ([...] capable of [...] using all the information simultaneously), (3) disinterested, (4) dispassionate, (5) consistent, and (6) normal in all other respects.” As the moral advice would be tailored to the moral beliefs of its user, such a system would not be subject to criticism based on the objection that a single moral theory is being promoted [75]. However, I have tried to argue that dispassionateness may not be an ideal worth pursuing in every circumstance in Sect. 4.1.2. Empathy may hold notions about the situation

of a particular individual that mere facts and data collection cannot cover, but which may still be relevant for moral decisions.

Regardless of this objection, a moral advisor would need formal tools both for implementing the data collection as well as proposing its ethical evaluation in alignment with the user's moral values. Technical challenges can plausibly call into question the potential for success of arriving at a moral advisor that truly fulfills the criteria of an ideal observer. For instance, how can it be ascertained whether all data relevant for an ethical evaluation has been acquired? Both hard- and software constraints will restrict data acquisition and evaluation, meaning that it is required to prioritize the items to consider. However, like Giubilini and Savulescu [20, p. 170] one may object that humans are “suboptimal information processors”, too. I concur, but at the very least, humans may question which information is relevant and which is not from time to time. Posing and answering such questions may, in fact, be a highly relevant aspect in forming moral understanding. A formalized system would be much more rigid and encourage to stop questioning the origins and relevancy of the data used for ethical evaluation. In addition, the data collection processes may be tacitly determined by the entity that supplies the moral advisors, which is likely to have a non-negligible effect on the moral evaluations. Even if single evaluations may be relatively robust, biases may appear in the aggregate. A significant asymmetry in relationships of power ensues. I have tried to illustrate similar tendencies in Sect. 4.2, which can already be observed with the autonomous driving industry today. It stands to reason that the potential for abusing such power is even more pronounced in general everyday moral advice than in self-driving cars.

However, my main point in this section is that such technological means powered by formal ethics will not contribute to moral progress as defined, e.g., in Sect. 3.2. In this conception, moral progress follows from moral understanding, i.e., an active attempt to develop a comprehension of moral reasons translated into moral practice. I maintain that formal tools for programming devices that deliver moral advice would prevent such an active formation of understanding.

Still, moral advisors may be proposed to shortcut only everyday decisions, especially those concerning which individuals may be said to have, or at least may feel to have, reached a high level of moral understanding already. However, as moral deliberation is not only a solitary but also a collaborative endeavor, such advisors do not facilitate exchange. Instead, they are aimed at automation for saving on time. In a few instances, people will still exchange about the moral advice they followed, scrutinize, and perhaps challenge it. However, a system, as proposed by Giubilini and Savulescu [20], is designed to counteract our purportedly fallible intuitions and emotions. The system expects

us to follow suit unquestioningly because we “do not have the time and the mental resources to gather and process all the information” [20, p. 170]. Arguably, moral understanding would not ensue. In light of my proposition that moral understanding is paramount to moral progress in Sect. 3.2, Giubilini's and Savulescu's proposal of an artificial moral advisor would then be likely to fail their own proclaimed goal of promoting moral progress.

In summary, I suggest that we should be very cautious about employing formal ethics as a potentially barrier-laden, opaque, and, ultimately, non-inclusive means to arrive at “more rigorous” moral judgments, regardless of whether this may be in an automated or expert-guided way.

5 Conclusion

In this article, I advocate caution against formal ethics. I have argued that it tends to produce unjustified power imbalances and is not supportive of inclusive moral deliberation. I have argued, however, that inclusive moral deliberation encourages increased and society-wide moral understanding. Increased moral understanding, taken as a heightened grasp of moral reasons, however, leads to moral progress. In effect, formal ethics tends to not support moral progress. However, for trying to ethically align technological artifacts, formal ethics may be necessary. I assert that a positive vision of use-cases of formal ethics thus entails approaches that (i) try to facilitate inclusive moral deliberation, (ii) provide interactive tools for training ethical reflection, or (iii) are used to implement a well-laid out moral consensus obtained and continuously challenged by inclusive moral debate for ethically aligning technology that needs to be endowed with some autonomy.

Thus, the perspective that I have presented suggests limited use of formal ethics for moral deliberation because we are best advised to favor inclusive and dialogical ways to avoid power imbalances instead. These are particularly worrisome if formal ethics are limited in reflecting actual moral consensus. If some restricted formal frameworks are widely adopted in automated moral decision-making, decisions could be dominated by a limited moral perspective. In conclusion, we might be best advised not to be talking of imbuing autonomous vehicles with ethics if all that we can and should do is implement the rules, principles, or case-based choices we have deliberated through fair decentralized and society-wide deliberation. This means that we should deliberate transparently the technical limitations of formal approaches to ethics and the restrictions that they impose on implementable moral consensus. The question then is, what moral consensus we can find, considering both the technical restrictions, societal constraints, and the promises a new technology holds. In practice, this could mean that we must

openly and honestly debate, what risks and disadvantages we are willing to accept for realizing the promises. This would lead to a new honesty about artificial moral agency. It would reduce the risk that those dominating technology and formal approaches can misappropriate the notion of “ethics”.

Acknowledgements The author is grateful to Graham Bex-Priestley for his valuable comments and pointers.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Oosterheld, C.: Formalizing preference utilitarianism in physical world models. *Synthese* **193**(9), 2747–2759 (2016). <https://doi.org/10.1007/s11229-015-0883-1>
- Fisher, M., List, C., Slavkovik, M., Winfield, A.F.T.: Engineering moral agents - from human morality to artificial morality (Dagstuhl Seminar 16222). *Dagstuhl Reports* **6**(5), 114–137 (2016). <https://doi.org/10.4230/DagRep.6.5.114>
- Cervantes, J.A., López, S., Rodríguez, L.F., Cervantes, S., Cervantes, F., Ramos, F.: *Artificial moral agents: a survey of the current status*, vol. 26. Springer (2020)
- Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., Bernstein, A.: Implementations in machine ethics: a survey, 1–37 (2020) [Online]. <http://arxiv.org/abs/2001.07573>
- Wallach, W., Allen, C.: *Moral machines: teaching robots right from wrong*. University Press, Oxford (2009)
- van Wynsberghe, A., Robbins, S.: Critiquing the reasons for making artificial moral agents. *Sci. Eng. Ethics* **25**(3), 719–735 (2019). <https://doi.org/10.1007/s11948-018-0030-8>
- Poulsen, A., et al.: Responses to a critique of artificial moral agents (2019) [Online]. Available: <http://arxiv.org/abs/1903.07021>
- Beck, S.: The problem of ascribing legal responsibility in the case of robotics. *AI Soc.* **31**(4), 473–481 (2016). <https://doi.org/10.1007/s00146-015-0624-5>
- Sharkey, A.: Can robots be responsible moral agents? And why should we care? *Conn. Sci.* **29**(3), 210–216 (2017). <https://doi.org/10.1080/09540091.2017.1313815>
- Scheutz, M.: The need for moral competency in autonomous agent architectures. In: Müller, V.C. (ed.) *Fundamental issues of artificial intelligence*, pp. 517–527. Springer International Publishing, Cham (2016)
- Russell, S.: *Human compatible—artificial intelligence and the problem of control*. Penguin Random House (2019)
- Wiegel, V.: Building blocks for artificial moral agents. In: *Artificial Life X: Proceedings of the Tenth International Conference on the Simulation and Synthesis of Living Systems* (2006)
- Simon, J.: The entanglement of trust and knowledge on the Web. *Ethics Inf. Technol.* **12**(4), 343–355 (2010). <https://doi.org/10.1007/s10676-010-9243-5>
- The Tesla Team: All tesla cars being produced now have full self-driving hardware,” Tesla.com Blog (2016). <https://www.tesla.com/blog/all-tesla-cars-being-produced-now-have-full-self-driving-hardware> (Accessed Jul. 14, 2020)
- Hern, A.: Self-driving cars don’t care about your moral dilemmas (2016)
- Habermas, J.: *Moral consciousness and communicative action*. Polity Press, Cambridge (1983)
- Flyvbjerg, B.: Ideal theory, real rationality: habermas versus foucault and nietzsche (2000). <https://doi.org/10.2139/ssrn.2278421>
- Fornet-Betancourt, R., Becker, H., Gomez-Müller, A., Gauthier, J.D.: The ethic of care for the self as a practice of freedom. *Philos. Soc. Crit.* **12**(2–3), 112–131 (1987). <https://doi.org/10.1177/019145378701200202>
- Hills, A.: Moral testimony and moral understanding. *Ethics* **120**(1), 94–127 (2009)
- Giubilini, A., Savulescu, J.: The artificial moral advisor—the ‘ideal observer’ meets artificial intelligence. *Philos. Technol.* **31**(2), 169–188 (2018)
- Anderson, M., Anderson, S.L., Armen, C.: An approach to computing ethics. *IEEE Intell. Syst.* **21**(4), 56–63 (2006). <https://doi.org/10.1109/MIS.2006.64>
- Gensler, H. J.: *Formal Ethics* (1996)
- Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., Kramer, M.: *Moral decision making frameworks for artificial intelligence* (2017) [Online]. Available: www.aaai.org
- Bonnemains, V., Saurel, C., Tessier, C.: Embedded ethics: some technical and ethical challenges. *Ethics Inf. Technol.* **20**(1), 41–58 (2018). <https://doi.org/10.1007/s10676-018-9444-x>
- Klein, W.E.J.: Robots make ethics honest. *ACM SIGCAS Comput. Soc.* **45**(3), 261–269 (2016). <https://doi.org/10.1145/2874239.2874276>
- Cave, S.J., Nyrup, R., Vold, K., Weller, A.: Motivations and risks of machine ethics. *IEEE Proc* (2018). <https://doi.org/10.1109/JPROC.2018.2865996>
- IEEE Global Initiative on Ethics of A/IS, “Ethically Aligned Design - Version 2 (2018)
- Lin, P.: The moral gray space of AI decisions (2018). Accessed: Aug. 02, 2020. [Online]. Available: <https://ai.shorensteincenter.org>
- Foot, P.: The problem of abortion and the doctrine of the double effect. *Oxford Rev.* **5** (1967)
- Himmelreich, J.: Never mind the trolley: the ethics of autonomous vehicles in mundane situations. *Ethical Theory Moral Pract.* **21**(3), 669–684 (2018). <https://doi.org/10.1007/s10677-018-9896-4>
- Allen, C., Smit, I., Wallach, W.: Artificial morality: top-down, bottom-up, and hybrid approaches. *Ethics Inf. Technol.* **7**(3), 149–155 (2005). <https://doi.org/10.1007/s10676-006-0004-4>
- Vorobej, M.: Review of ‘Formal Ethics’ by Harry Gensler. *Dialogue Can. Philos. Assoc.* **38**(2), 449–450 (1999). <https://doi.org/10.4324/9780203288245>
- Polonski, V.: Can we teach morality to machines? Three perspectives on ethics for artificial intelligence. *Medium* (2017)
- Powers, T.M.: Prospects for a Kantian Machine. *IEEE Intell. Syst.* **21**(4), 46–51 (2006). <https://doi.org/10.1109/MIS.2006.77>
- Anderson, M., Anderson, S.L.: “GenEth: a general ethical dilemma analyzer. *Paladyn. J. Behav. Robot.* **9**(1), 337–357 (2018). <https://doi.org/10.1515/pjbr-2018-0024>

36. Anderson, M., Anderson, S. L., Armen, C.: Towards machine ethics: Implementing two action-based ethical theories. AAAI Fall Symp. Tech. Rep. **FS-05-06**, 1–7 (2005)
37. Shulman, C., Jonsson, H., Tarleton, N.: Which consequentialism? Machine ethics and moral divergence. In: AP-CAP 2009: the fifth asia-pacific computing and philosophy conference, pp. 23–25 (2009)
38. Bogosian, K.: Implementation of moral uncertainty in intelligent machines. *Minds Mach.* **27**(4), 591–608 (2017). <https://doi.org/10.1007/s11023-017-9448-z>
39. MacAskill, W., Ord, T.: Why maximize expected choice-worthiness? *Noûs* **54**(2), 327–353 (2020). <https://doi.org/10.1111/nous.12264>
40. Mill, J.S.: On liberty, 2001st edn. Batoche Books Limited, Kitchener (1859)
41. Young, I. M.: Inclusion and democracy. Oxford University Press (2002)
42. Ingber, S.: The marketplace of ideas: a legitimizing Myth. *Duke Law J.* **1984**(1), 1 (1984). <https://doi.org/10.2307/1372344>
43. Awad, I.: Critical multiculturalism and deliberative democracy: Opening spaces for more inclusive communication. *Javnost* **18**(3), 39–54 (2011). <https://doi.org/10.1080/13183222.2011.11009061>
44. Granville K.: Facebook and cambridge analytica: what you need to know as fallout widens. *The New York Times* (2018)
45. Staiano, J., Lepri, B., Aharony, N., Pianesi, F., Sebe, N., Pentland, A.: Friends don't Lie—inferring personality traits from social network structure. In: UbiComp'12 - Proc. 2012 ACM Conf. Ubiquitous Comput. no. February 2014, pp. 321–330 (2012) <https://doi.org/10.1145/2370216.2370266>
46. Markovikj, D., Gievska, S., Kosinski, M., Stillwell, D.: Mining facebook data for predictive personality modeling. AAAI Work. Tech. Rep. **WS-13-01**, 23–26 (2013)
47. González, R.J.: Hacking the citizenry?: Personality profiling, 'big data' and the election of Donald Trump. *Anthropol. Today* **33**(3), 9–12 (2017). <https://doi.org/10.1111/1467-8322.12348>
48. Gardiner, M.: Foucault, ethics and dialogue. *Hist. Human Sci.* **9**(3), 27–46 (1996). <https://doi.org/10.1177/095269519600900302>
49. Moody-Adams, M.M.: The idea of moral progress. *Metaphilosophy* **30**(3), 168–185 (1999). <https://doi.org/10.1111/1467-9973.00120>
50. Buchanan, A.: A pluralistic, dynamic conception of moral progress, vol. 1. Oxford University Press (2018)
51. Riaz, A.: Moral understanding and knowledge. *Philos. Stud. An Int. J. Philos. Anal. Tradit.* **172**(1): 112–128 (2015) <https://doi.org/10.1007/s1098-014-0328-6>
52. Aristotle: The nicomachean ethics. Oxford World's Classics (2009)
53. McGrath, S.: Moral disagreement and moral expertise. *Oxford Stud. Metaethics* **III**, 87–108 (2008)
54. Pasquale, F.: The black box society—the secret algorithms that control money and information. Harvard University Press, Cambridge; London (2015)
55. Mehta, J., Winship, C.: Moral power. *Handb. Sociol. Moral.* **41**(19), 425–438 (2010)
56. Anderson, M., Anderson, S.L.: Machine ethics: creating an ethical intelligent agent. *AI Mag.* **28**(4), 15–26 (2007)
57. Bertonecello M., Wee, D.: Ten ways autonomous driving could redefine the automotive world. McKinsey & Company, 1–6, (2015)
58. Coeckelbergh, M.: Responsibility and the moral phenomenology of using self-driving cars. *Appl. Artif. Intell.* **30**(8), 748–757 (2016). <https://doi.org/10.1080/08839514.2016.1229759>
59. Arkin, R.C., Ulam, P., Wagner, A.R.: Moral decision making in autonomous systems: enforcement, moral emotions, dignity, trust, and deception. *Proc. IEEE* **100**(3), 571–589 (2012). <https://doi.org/10.1109/JPROC.2011.2173265>
60. Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., Yang, Q.: Building ethics into artificial intelligence. *IJCAI Int. Jt. Conf. Artif. Intell.*, **2018**, 5527–5533 (2018)
61. Felber E., Langendorf, M.: Volkswagen plans to make autonomous driving market-ready. *Volkswagen Gr. News*, **370**, 1–4 (2019). [Online]. Available: <https://www.volkswagen-newsroom.com/en/press-releases/volkswagen-plans-to-make-autonomous-driving-market-ready-5498>
62. Musk, E.: Master plan, part deux. *Tesla.com* (2016). <https://www.tesla.com/blog/master-plan-part-deux> (Accessed Jul. 14, 2020)
63. Mider, Z.: Tesla's autopilot could save the lives of millions, but it will kill some people first. *Bloomberg Businessweek* (2020)
64. Brown, J.: Tesla announces update to self-driving system after fatality in May. *The Guardian*, San Francisco (2016)
65. Ethics Commission: Automated and connected driving—a report commissioned by the federal ministry of transport and digital infrastructure, Germany (2017). [Online]. Available: https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile
66. Tesla, Inc.: Wikipedia (2020)
67. Burrell, J.: How the machine 'thinks': understanding opacity in machine learning algorithms. *Big Data Soc.* **3**(1), 1–12 (2016). <https://doi.org/10.2139/ssrn.2660674>
68. Surden, H., Williams, M.-A.: Technological opacity, predictability, and self-driving cars. *Cardozo Law Rev.* **38**(121), 121–181 (2016). [Online]. Available: <http://scholar.law.colorado.edu/articles/24>
69. Meyers, J. M.: Artificial intelligence and trade secrets. *Landslide*, **11**(3) (2019)
70. Jasanoff, S.: The Ethics of Invention: Technology and the Human Future. W.W. Norton (2016)
71. Birnbacher D., Birnbacher, W.: Automatisiertes Fahren. *Inf. Philos.* 8–15, (2016)
72. Sovani, S.: Simulation accelerates development of autonomous driving. *ATZ Worldw.* **119**(9), 24–29 (2017). <https://doi.org/10.1007/s38311-017-0088-y>
73. Anderson, S. L.: How machines might help us achieve breakthroughs in ethical theory and inspire us to behave better. In: Anderson M., Anderson, S. L. (eds) *Machine Ethics*, pp. 524–533 (2011)
74. Firth, R.: Ethical absolutism and the ideal observer. *Philos. Phenomenol. Res.* **12**(3), 317–345 (1952)
75. Savulescu J., Maslen, H.: Moral enhancement and artificial intelligence: moral AI? In: Romportl, J., Zackova, E., Kelemen, J. (eds) *Beyond artificial intelligence. Topics in intelligent engineering and informatics*. Springer, Cham, pp. 79–95 (2015)
76. Bakhtin, M.: Art and answerability: early philosophical essays by M. M. Bakhtin. University of Texas Press, Austin (1990)
77. Brown, V.: The moral self and ethical dialogism: three genres. *Philos. Rhetor.* **28**(4), 276–299 (1995)
78. Rosenberg, M. B.: Nonviolent communication: a language of life, 2nd ed. Puddle Dancer Press (2003)
79. Wallach, W.: Artificial morality: bounded rationality, bounded morality and emotions. *Cogn. Emot. Ethical Aspects Decis. Making Hum. Artif. Intell.* **III**, 1–6 (2005)
80. Wallach, W.: Robot minds and human ethics: the need for a comprehensive model of moral decision making. *Ethics Inf. Technol.* **12**(3), 243–250 (2010). <https://doi.org/10.1007/s10676-010-9232-8>
81. Gilbert, B.: Self-driving cars still won't prevent the most common car accidents, according to a new study. *Business Insider* (2020)
82. Dixit, V.V., Chand, S., Nair, D.J.: Autonomous vehicles: Disengagements, accidents and reaction times. *PLoS ONE* **11**(12), 1–14 (2016). <https://doi.org/10.1371/journal.pone.0168054>
83. Strickland, E.: IBM Watson, heal thyself: how IBM overpromised and underdelivered on AI health care. *IEEE Spectr.* **56**(4), 24–31 (2019). <https://doi.org/10.1109/MSPEC.2019.8678513>

84. Ross C., Swetlitz, I.: IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. Stat News (2018). <https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/>
85. Teoh, E.R., Kidd, D.G.: Rage against the machine? Google's self-driving cars versus human drivers. J. Safety Res. **63**, 57–60 (2017). <https://doi.org/10.1016/j.jsr.2017.08.008>
86. Buchanan A., Powell, R.: The Evolution of Moral Progress, vol. 1. Oxford University Press (2018)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.