AI and Ethics (2023) 3:329–336
https://doi.org/10.1007/s43681-022-00173-5

COMMENTARY

AI ethics and its pitfalls: not living up to its own standards?

Thilo Hagendorff[1]

Received: 24 February 2022 / Accepted: 9 May 2022 / Published online: 30 May 2022

Abstract
AI ethics is deemed to be an essential ingredient in the quest for trustworthy AI. Hence, demands for implementing AI ethics and ethicists into AI organizations, especially corporations, are ubiquitous. However, the assumption that AI ethicists have particular epistemological advantages compared to non-ethicists as well as the idea that AI ethics automatically decreases the likelihood of unethical outcomes are both flawed. Therefore, this comment lists risks that either originate from AI ethicists themselves or from the consequences their embedding in AI organizations has. The compilation of risks comprises psychological considerations concerning the cognitive biases of AI ethicists themselves as well as biased reactions to their work, subject-specific and knowledge constraints AI ethicists often succumb to, negative side effects of ethics audits for AI applications, and many more. Ultimately, the aim of this comment is not to diminish or deny the importance of the discipline of AI ethics, but rather to increase its capacities for self-reflection and, ultimately, effectiveness.

Keywords AI ethics · Bounded ethicality · Risk · Moral psychology · Artificial intelligence

1 Introduction

By claiming that AI research and development has to become an interdisciplinary field that should include the humanities and ethics in particular, the notion that AI ethicists are pivotal in the quest for "trustworthy", "value-based", "human-centered", "beneficial" AI has become mainstream, along with the demand to hire professional AI ethicists in corporations [1], using designations like chief AI ethics officer, chief trust officer, ethical AI lead, AI ethics and governance, or trust and safety policy advisor [2]. Such a hiring decision can have undeniable advantages for AI organizations. These range from proactively detecting non-compliance issues, unwanted technological consequences, fairness biases, accountability gaps, privacy violations, transparency issues, security vulnerabilities, strategic decisions concerning AI products, and the like. But apart from these advantages, embedding ethicists or AI ethics principles in AI organizations bears a range of specific risks that are hitherto disregarded in the field of AI ethics itself. Hence, this comment discusses and lists risks that either originate from AI ethicists themselves or from the consequences their embedding in AI organizations has. The compilation of risks comprises psychological considerations of thinking about ethics and acting (un)ethically, subject-specific and knowledge constraints AI ethicists often succumb to, the complicated professional role of AI ethicists, the missing impact of AI ethics guidelines, as well as negative side effects of ethics audits for AI products.

2 Risks

In the following, several types of risks are discussed that are tied specifically to the inclusion of professional AI ethicists, ethical considerations, or ethical audits in organizations researching and developing AI systems and applications. Itemizing these risks is rather unusual for the field of AI ethics and seems to contradict the field's very purpose. Accordingly, the AI ethics literature is nearly completely focused on stressing the positive effects and importance of ethics considerations, conceptualizing ethics as a service, and eventually putting ethical principles into practice. On the other hand, self-critical approaches are rather rare [3]. This comment does not aim to diminish the importance and necessity of ethical reflections in the AI field. Rather, it stresses that implementing AI ethics is not as straightforward

✉ Thilo Hagendorff
thilo.hagendorff@uni-tuebingen.de

[1] Cluster of Excellence "Machine Learning: New Perspectives for Science", University of Tuebingen, Tübingen, Germany

and self-explanatory as it seems to be. In fact, well-meant ethical intentions can have inadvertent, hidden side effects that, if evaluated on their own, would be deemed unethical. This comment wants to list these side effects. They are not necessary but, in many cases, likely concomitant with AI ethics. This means that whenever claims on AI ethics or AI ethicists are made in this paper, this does at no time comprise AI ethics in general or all AI ethicists. This comment cannot make any quantitative statements. However, what can be claimed is that instances of the risks that are mentioned in the following sections do exist to an extent that they are worth discussing.

2.1 Role-model risks

In general, one can uphold a "boost hypothesis" that assumes that ethical considerations have a positive effect on moral behavior and attitude–behavior consistency. However, this hypothesis does not bear closer examination. Empirical research has shown that professional ethicists—which include AI ethicists—do not appear to behave morally better than non-ethicists [4–10]. Studies on the moral behavior of professional ethicists in the context of theft, voting participation, membership in the Nazi party, littering, honesty, caring behavior, etc. found that, on average, ethicists do not behave differently compared to non-ethicists of a comparable socio-economic background. A possible explanation for this phenomenon is that professional ethicists mainly focus on very abstract, obscure thought experiments that are often detached from issues of everyday life. Another simple but plausible explanation is that mere knowledge about ethical issues and theories does not affect behavior. Indeed, research in moral psychology shows that ethical reasoning is mainly a post hoc rationalization of already entrenched opinions and behavioral routines, that it is the result of automatic mental processes that are relatively detached from conscious reflection on moral principles, and that factors other than moral reasoning significantly influence moral judgments and moral behavior [11–19]. Moreover, empirical research has shown that endorsing moral principles and possessing explicit philosophical expertise about ethics does not enhance the consistency of the application of these principles or the stability of moral judgments against the influence of biases [20]. Professional ethicists succumb to biases, inconsistencies, post hoc rationalizations, superficial differences, irrationality, or other psychological factors that they would not endorse upon reflection to the same degree as non-ethicists. This happens despite high levels of academic expertise and training and even though one would assume that expertise in ethics serves as some kind of protection against these factors [21, 22]. In addition to that, research shows that people typically have a biased perception of their own ethicality, deeming that they behave more ethically than they actually do [23]. Similar insights hold true for another meta-bias where people are able to identify others' susceptibility to biases but they see themselves as immune to these biases [24].

One must suppose that all the mentioned factors do hold true for AI ethicists as well. Hence, they are no different than non-ethicists, likely to have a flawed self-perception, deeming themselves more ethical and bias-sensitive than they actually are. In addition, they may very well not live up to the expectations regarding their own moral integrity when it comes to their actual behavior. Hence, while one may already be skeptical about ethicists' participation in AI research and development contexts due to their limited technical expertise (see Sect. 2.3), one may be also skeptical when AI ethicists are depicted as individuals with particularly wise conduct, as role models, or even as personifications of morality and virtues. However, it is obvious that one should not exclude AI ethicists from organizational structures in research and academia. Rather, AI ethicists have to become knowledgeable in moral psychology and able to reflect on the psychological forces that impinge on them. One cannot separate personal belief systems, biases, or behavioral traits from one's professional work for instance on AI policies or assessments. The former always influences the latter [25, 26]. AI ethicists, due to the nature of their profession, must be particularly self-aware of this correlation. If they do not keep it in mind, non-ethicists like AI practitioners are in no way less competent or versed in reaching morally right and ethically reflective decisions in their field compared to professional AI ethicists. This is, as research on moral psychology suggests, due to the fact that other factors than ethical expertise and knowledge predominantly determine moral behavior and decision-making. First and foremost, these factors comprise peer influences, situational forces, as well as intrapersonal factors of bounded ethicality [27]. AI ethicists should learn to deal with these factors that act upon their very own considerations and decision-making.

2.2 Bounded ethicality risks

When using a naïve ethical action theory, one can assume that individuals go through three phases. First, they perceive a confrontation with a moral decision they have to make. Secondly, they reflect on ethical principles and come up with a moral judgment. And finally, they act accordingly to these judgments—and, therefore, morally right. However, such a naïve action theory is flawed in several regards. As mentioned before, moral judgments are, in most cases, not influenced by moral reasoning [11]. Moral judgments are made intuitively, and moral reasoning is used in hindsight to justify one's initial reaction. In short, typically, moral action precedes moral judgment [28]. This leads to specific consequences for AI ethics. Typically, AI ethicists may be tempted to mainly convey mere knowledge (instead of practical skills

or motivations) about ethical issues that are related to AI. However, up to now, they do not consider the many hidden psychological forces like powerful unconscious biases, blind spots, situational forces, and the like that are likely to thwart ethical decision-making not just in themselves, but also in AI practitioners [27, 29]. In research on moral psychology, those factors are subsumed under the umbrella term "bounded ethicality" [27, 29]. Factors of bounded ethicality can, as long as AI ethicists do not acknowledge and address them in AI practitioners, increase the likelihood of unethical decision-making, rationalize it, or reframe it in a misleading way. Hence, to effectively improve ethical decision-making in AI practitioners, AI ethicists must go beyond simply conveying knowledge about ethical issues. Instead, they should engage in initiating measures for ethics training that foster virtues, motivations, or character dispositions that are potent to protect against factors of bounded ethicality and help "debiasing" ethical decision-making [30]. In practice, AI ethicists who are embedded in academia or industry settings are mostly supposed to ensure legislation requirements, to evaluate AI systems along ethical criteria, to embed ethical principles into the business use-case, AI design, and deployment, or to facilitate public debates [2]. Hence, providing behavioral training is not a common practice for AI ethicists, besides its importance in increasing the likelihood of ethical decision-making. In this context, especially business ethics resources describe plenty of measures for ethics training in organizations [31–35]. This training, as soon as it is specialized on AI-related topics, clearly falls into the range of responsibilities of AI ethicists. If they only engage in composing and discussing high-level frameworks of principles and refrain from also putting a focus on cognitive biases, value–action gaps, overconfidence, implicit biases, in-group favoritism, moral disengagement, self-serving biases as well as many others, they will not contribute to the full extend to the quest for reducing the likelihood of unethical behavior in AI organizations [36–43]. However, up to know, AI ethics neither discusses bounded ethicality in AI ethicists themselves (Sect. 2.1) nor how to address it in AI practitioners despite the decisive role methods to circumvent effects of bounded ethicality would have.

2.3 Non-expert risks

Whereas AI research, or the computer sciences in general, is driven by numbers and data by nature, ethicists sometimes claim to have special access to non-quantifiable areas of human life, to areas that cannot be "translated" into numbers [44]. Hence, ethics seems to have an epistemological advantage compared to the former since it can, so the argument goes, grasp emotions, beliefs, desires, or values that are deemed unquantifiable. But as long as ethics fosters such self-concepts, its discourse lacks a decisive enrichment, namely empirical, and with that, quantitative underpinnings. This does not mean that quantitative data should bear more weight than value-centered, qualitative arguments. However, to produce sound scientific insights, both should be combined, instead of just relying on the latter. However, philosophy, and hence philosophical, non-interdisciplinary ethics, has a long tradition of embracing the naturalistic fallacy or the is-ought problem [45], meaning the tenet that one cannot base normative statements on observations about what is. This tenet led to a severe neglect of empiricism in practical philosophy and an amnesia of the value and importance of empirical results for ethical reasoning [46, 47]. Similarly, AI ethics research works stemming from philosophy often put together arguments without delving into empirical evidence, meaning technical details, about particular AI systems. This has in many cases led to redundant, abstract discussions on topics like trust, autonomy, existential risks, fairness, explanations, moral robots, etc. However, if practical philosophy intervening in AI-related ethical questions would also possess a certain understanding of engineering or respective areas of computer science and be able to empirically investigate AI systems first-hand, this could be a significant enrichment for AI ethics.

In theory, the requirements for a professional AI ethicist who does not only focus on traditions of practical philosophy must include an overview of computer science, psychology, sociology, law, and many others. AI ethicists cannot be pigeonholed into one of these disciplines [2]. Traditionally, applied ethics is a strict subfield of practical philosophy, though. Philosophy, in turn, has a reputation for being a rather conservative subject that engages in research mainly by exegesis of classical philosophical texts but does very seldom yield knowledge that is not reliably determinable by other sciences. Philosophy intellectually grasps current societal and technological phenomena, but when it does, it often focuses on speculative scenarios that, by its very nature, do not require detailed empirical knowledge. Therefore, one can assume that philosophers typically rely too much on introspection and thought associations in their arguments, rather than on empirically observable facts. In the AI field, this circumstance manifests in the popularity of fictitious topics like machine consciousness, artificial moral agents, trolley problems, robot rights, AI "narratives", superintelligence, or other futuristic or speculative scenarios. For interdisciplinary AI research settings that possess tangible real-world implications, these are rather bad premises, though. To the best of my knowledge, hitherto there has not been a study that gathers data on the actual level of technical knowledge about machine learning among AI ethicists. Hence, only anecdotal evidence can support the tentative claim that, in many cases, AI experts with a background in computer science are better ethicists than genuine AI ethicists themselves. This is because the most pressing problems that are

discussed in AI ethics, like fairness, robustness, privacy, explainability, etc., can be ameliorated by technical means. In the end, putting ethics into practice happens to be an issue in AI developers' everyday world. Despite that, AI ethicists often work with assessment lists, principled frameworks, and other deontologically contextualized codes. However, linear auditing mechanisms or abstract principles are not compatible with agile technology development [48]. In fact, motivational settings or character dispositions regarding particular "AI virtues" [30] as well as detailed how-to knowledge on bias mitigation, robustness gains, differential privacy, explainable AI, etc. are decisive in achieving "ethical AI". However, both aspects only make sense when internalized by AI practitioners, whereas AI ethicists are mostly bystanders.

2.4 Effectiveness risks

One of the main results of the field of AI ethics is the surge in publications of AI ethics guidelines [49, 50]. Lists of abstract principles that are at the core of many guidelines are not the right ingredient to bring about change, though. This argument can be further underpinned by empirical research. In a controlled study by McNamara [51], researchers critically reviewed the idea that ethical guidelines serve as a basis for ethical decision-making for software engineers. In the survey, test subjects were 63 software engineering students and 105 professional software developers. They were presented with eleven software-related ethical decision scenarios, testing whether the influence of the ethics guideline of the Association for Computing Machinery (ACM) [52] influences ethical decision-making in six vignettes, ranging from responsibility to report, user data collection, intellectual property, code quality, honesty to the customer to time and personnel management. The results are disillusioning. The main finding was that the effectiveness of codes of AI ethics is almost zero and that they do not change the behavior of practitioners. No statistically significant differences in the responses were found across test subjects who did and did not read through the ethics guideline [51]. All in all, the study shows that knowledge about ethical issues does not affect behavior. It is a lesson on the ineffectiveness of applied AI ethics, or, to be more precise, the current methods ethicists are using to propel their field.

This ineffectiveness of the intended effects of codes of ethics may be accompanied by another effectiveness risk resulting from moral communication, or, in other words, attempts to persuade others of the importance of ethical principles. In theory, AI ethicists should possess significant influence in all departments of organizations researching and developing AI applications. However, the decisive question is how this influence takes place. By training, ethicists do not acquire persuasion techniques or knowledge of methods for strategic communication [53, 54]. However, the only means ethicists have to decrease the distance between actual and target states towards the latter are not money, power, or other "symbolically generalized communication media" [55], but persuasion techniques. That means that ethicists provide normative orientations, engage people's motivation to head towards normative goals, and shape the path towards them [56]. The respective persuasion attempts can vary in their nature, though. The most obvious way of making persuasion attempts is through the use of moral language and vocabulary, meaning to stress good what is good and bad. However, this will often trigger confrontational situations, especially when contradictions between ethical and business requirements arise. And these situations, then, are likely to trigger reactance phenomena, backfire, or boomerang effects [57–60]. Reactance occurs when pressure is applied to individuals regarding a desired change in their normative attitudes or when individuals feel that their behavioral freedoms are to be diminished. In addition to that, reactance implies a reinforcement of attitudes that are contrary to what is actually desired by ethicists [61]. Hence, well-intended ethical interventions can, in cases where strong and confronting moral language or other suboptimal persuasion techniques are used, result in the exact opposite of what was intended. In such a case, having AI ethicists in place bears the risk of actually strengthening unethical tendencies in organizations. To avoid that, AI ethicists should be trained in the psychology of ethically motivated persuasion and strategic communication. However, under normal circumstances, such training is neither part of ethicists' academic education nor their daily work.

2.5 Ethics washing risks

Since involving ethical assessments in AI research and development is not mandatory, it is likely that AI organizations will only accept overhead expenses and install ethicists in cases where they already expect their AI products to be well rated. Organizations, however, that fear significant reputational losses as a result of serious ethical scrutiny are unlikely to initiate it in the first place. If that hypothesis is correct, AI ethicists are most needed where they are least likely to be called upon. Hence, ethical oversight and associated consequential costs are likely to disadvantage those organizations that are already economically disadvantaged due to their more ruthless competitors. A simple test to see whether ethical considerations are taken seriously in corporations is to look for cases where they refrain from seizing business opportunities even though they would be completely legal. Relinquishing monetary advantages, however, goes against the fundamental system conditions of the company in question, which leads to ethics becoming a "natural" opponent of economic success, at least as long as it is deemed to be some kind of supplementary external auditing

entity. However, since incorporating ethics into corporations can at least boost their reputation, a twisted version of ethical assessments can be embraced.

In this regard, AI ethicists and ethics discourses can be deployed at the forefront of companies and AI organizations, while business as usual continues behind the scenes [62]. In an empirical study by Vakkuri et al. [63] about applied AI ethics in industrial settings, the researchers conducted semi-structured qualitative interviews in various companies. When asked whether their AI development practices take into account ethical considerations or ethics tools, all respondents answered "No" [63]. However, according to the study, AI developers consider the physical harm potential of AI systems, and responsibilities are approached pragmatically from a financial, customer relations, or legislative stance. What is salient here, though, is that these opinions correspond to established moral attitudes that most people have internalized anyhow, but that are detached from what is actually discussed in AI ethics, like, for instance, the evaluation and governance of structural, social, or emotional effects of AI systems. In general, papers on AI meta-ethics regularly ascertain that AI ethics is "toothless" [64], has a fig leaf function [65], is a mere invention of the AI industry [66], a puppet of industry stakeholders [67], used for whitewashing purposes [68], an escape from binding legal norms and regulation [69], and the like. All in all, AI ethicists may be willing tools for interests that are actually situated outside of ethical settings. The existence of AI ethicists may allow the successful and persuasive conduct of actual malpractices like choosing ethical principles that retrofit pre-existing behaviors, making misleading or unsubstantial claims about ethical measures, exploiting ethics to delay or avoid necessary regulation, etc. [62] in the first place. Without AI ethicists or the embedding of ethics discourses into AI organizations, such malpractices and the associated abuse of trust would not play out since referring to "ethics" would not work. Notwithstanding, one may be tempted to accuse companies of ethics washing when they hire ethicists and to accuse them of ignorance if they do not hire ethicists. Such contradictoriness is obviously not productive. Ultimately, professional AI ethicists are always dependent on their employers, and that raises the likelihood of them not objecting to participating in ethics washing practices. However, it is true that such a restriction is not a necessity that a priori allegations without any underpinning through actual cases are counterproductive, and that AI ethicists can bring about real and authentic changes towards beneficial and ethical AI applications.

2.6 AI authority risks

A common tenet when discussing the use of AI systems in high-stake areas is to deem these systems as mere decision support tools, but not as decision makers themselves. However, this is too short-sighted since AI systems have a particular kind of authority, a semblance of an objectively functioning, mathematically precise, and neutral tool—despite the fact that they will under any circumstances contain errors and unfairness [70]. Especially in situations with high stress or time pressure, human decision makers who are supposed to merely consult AI recommendations may be prone to suspend this tenet and simply take the AI decision for granted, showing trust in the "truth-telling" power of technology. Algorithmic decision-making promotes defensive human decision-making [71], meaning that humans tend to offload the strain of decision-making onto computers. Medical, military, or other practitioners in high-stake areas must be aware of AI authority.

Ethics-based audits, AI impact assessments, ethics consulting, AI fairness tools, ethics-as-a-service offers, ethics labels for AI systems, and the like [48, 72–74] are supposed to make AI systems trustworthy, but in this way, they can actually exacerbate the problem of defensive decision-making. The more trustworthy an AI system is supposed to be, the more following the respective algorithmic decisions becomes the most defensible decision one can make, though perhaps not the best. And the more AI systems are implemented in high-stakes areas, the more comprehensive the auditing requirements become. And with that, the trust in the system and thus the defensive decision-making increases, although making decisions that one believes are the best for clients or organizations, even when it entails deviating from second-best algorithmic decisions, becomes all the more important. In fact, defensive decision-making is mainly a problem in the medical field, where practitioners tend to show "assurance behavior," meaning ordering unnecessary tests, medical imaging procedures, drugs, or surgery since this decreases the likelihood of malpractice litigation [75]. But, on the other hand, the respective decisions can very often deviate from what is best for patients. AI systems exacerbate this situation in two ways. First, when fed with training data on past medical decision-making routines, they perpetuate patterns of these routines and therefore mechanize suboptimal decision-making procedures. And second, as mentioned above, especially medical AI systems are subject to ethical scrutiny [76], hence trust in the outputs of these systems is likely to be increased, too. And with that, users are even more prone to AI-induced defensive decision-making.

Obviously, dynamics leading to defensive decision-making cannot and should not be avoided by refraining from auditing AI systems, by not increasing their trustworthiness via impact assessments, labels, etc. However, AI ethics should at least account for the potential that well-meant or mandatory audits of AI systems can exacerbate the problem of defensive decision-making. The constitutive dynamic

described here is neither caused nor a necessary accompanying effect of ethics-based influences on AI research and development. However, it is a potential, undesired side effect that in and off itself should be considered in a self-reflective impact assessment of AI ethics itself.

2.7 AI avoidance risks

AI authority is one side of the coin. The other side is AI skepticism [77]. While trusting AI recommendations too much, being skeptical of them can also lead to suboptimal outcomes. Skepticism towards AI systems can, among other things, result from a lack of value alignment. This lack can also be fostered by AI ethicists. Usually, and as described in the last chapter, it is argued that trust in AI applications and hence their adaption can be fostered when ethicists play a part in auditing them. Otherwise, severe social opportunity costs associated with AI system underuse may occur [48]. However, one can also argue that it is exactly the other way around. Following Bonnefon et al. [78, 79], one could use the example of self-driving cars. Ethicists would very likely opt for a utilitarian regulation that states that autonomous vehicles should strive to minimize casualties, even if this means sacrificing their own passengers for the greater good. Alternatively, they opt for a complete prohibition of determining different values of human life [80]. However, such an implementation of ethical principles or prohibitions in AI systems bears the chance of people opting out of using the very systems, which nullifies all their potential benefits. When AI ethicists demand a limit on the overall number of accidents and hence the greatest safety benefits for all, meaning passengers and pedestrians, this could paradoxically result in thwarting these very goals since consumers would likely be inclined to ditch such cars and hence the safety benefits that are connected to them [79]. This scenario could also hold true for other applications. Bonnefon et al. [78] mention sentencing [81] and organ allocation algorithms [82]. Here, making these systems more "ethical" in the sense of minimizing rearrests or giving younger people priority for organ transplants could again cause paradoxical effects that induce people to ditch these systems despite their likely advantages. Other examples may involve all AI systems that promise certain benefits but offer easy ways to opt out in case of a disagreement with amendments encouraged by AI ethicists. In the end, AI ethics audits should always be two steps ahead, meaning that ethicists do not just consider immediate, but also secondary technical consequences. This, in turn, may mean discarding obvious ethical decisions in favor of non-obvious, seemingly unethical ones. And furthermore, this means that AI ethics has to question its fundamental negativity bias. This bias, which causes the field to concentrate on negative issues and events rather than positive ones (which also holds true for this paper), goes hand in hand with or even triggers the negativity bias of media reports on AI topics. This way and by not stressing the positive sides of AI technologies, AI ethics assists in diminishing the reputation these technologies possess for the wider public. However, this can again lead to severe opportunity costs that occur due to the underuse of potentially life-saving or life-improving AI technologies. Hence, AI ethics—besides its strong preference to shed light on fairness issues, black box problems, privacy breaches, etc.—should learn to also have an eye on the positive events and positive sides of AI to prevent a situation where benefits are given away due to (unsubstantiated) fears regarding the likelihood of machine biases, accidents, or other harms.

3 Conclusion

Intuitively, it seems that AI ethics is a discipline that necessarily reduces the likelihood of unethical outcomes in the AI field. But this intuition does not hold true under scrutiny. In fact, one can identify several risks that either originate from ethicists themselves or from the consequences their embedding in AI organizations has. These risks relate to psychological considerations about bounded ethicality in ethicists themselves, to considerations regarding the individuals who react to (or ignore) ethical principles and advice, to the complicated professional role of AI ethicists, the lack of impact of AI ethics guidelines, or the potential negative side effects of ethics audits for AI products. With that, this comment does not aim to diminish the importance of AI ethics. Quite the contrary, its objective is to increase self-reflection and thus the effectiveness of the very discipline. However, the comment also stresses that implementing AI ethics is not as straightforward and self-explanatory as it seems to be. Well-intentioned ethical considerations can have unintended, hidden side effects that, if evaluated on their own, would be deemed unethical. These side effects should be avoided to turn AI ethics into a discipline that can live up to its own standards.

Acknowledgements This research was supported by the Cluster of Excellence "Machine Learning—New Perspectives for Science" funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—Reference Number EXC 2064/1—Project ID 390727645.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Fisher, B.: Top 5 AI hires companies need to succeed in 2019, 2019. https://info.kpmg.us/news-perspectives/technology-innovation/top-5-ai-hires-companies-need-to-succeed-in-2019.html. Accessed 12 Oct 2021
2. Hickok, M.: What does an AI Ethicist do? A guide for the why, the what and the how. 2020. https://medium.com/@MerveHickok/what-does-an-ai-ethicist-do-a-guide-for-the-why-the-what-and-the-how-643e1bfab2e9. Accessed 12 Oct 2021
3. T. Hagendorff, Blind spots in AI ethics, AI and Ethics (2021) 1–17.
4. Schwitzgebel, E., Rust, J.: The moral behavior of ethics professors: relationships among self-reported behavior, expressed normative attitude, and directly observed behavior. Philos. Psychol. **27**, 293–327 (2014)
5. Schwitzgebel, E., Rust, J.: The behavior of ethicists. In: Sytsma, J., Buckwalter, W. (eds.) A companion to experimental philosophy, pp. 225–233. Wiley, Chichester (2016)
6. Schönegger, P., Wagner, J.: The moral behavior of ethics professors: a replication-extension in German-speaking countries. Philos. Psychol. **32**, 532–559 (2019)
7. Schwitzgebel, E.: Do ethicists steal more books? Philos. Psychol. **22**, 711–725 (2009)
8. Schwitzgebel, E., Rust, J.: Do ethicists and political philosophers vote more often than other professors? Rev. Philos. Psychol. **1**, 189–199 (2010)
9. Rust, J., Schwitzgebel, E.: Ethicists' and nonethicists' responsiveness to student e-mails: relationships among expressed normative attitude, self-described behavior, and empirically observed behavior. Metaphilosophy **44**, 350–371 (2013)
10. Schwitzgebel, E.: The moral behavior of ethicists and the role of the philosopher. In: Luetge, C., Rusch, H., Uhl, M. (eds.) Experimental ethics, pp. 59–64. Palgrave Macmillan UK, London (2014)
11. Haidt, J.: The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. Psychology Review **108**, 814–834 (2001)
12. Mathews, K.E., Canon, L.K.: Environmental noise level as a determinant of helping behavior. J. Pers. Soc. Psychol. **32**, 571–577 (1975)
13. Milgram, S.: Behavioral study of obedience. J. Abnorm. Psychol. **67**, 371–378 (1963)
14. Isen, A.M., Levin, P.F.: Effect of feeling good on helping: cookies and kindness. J. Pers. Soc. Psychol. **21**, 384–388 (1972)
15. Darley, J.M., Batson, C.D.: "From Jerusalem to Jericho": a study of situational and dispositional variables in helping behavior. J. Pers. Soc. Psychol. **27**, 100–108 (1973)
16. Inbar, Y., Pizarro, D.A., Bloom, P.: Disgusting smells cause decreased liking of gay men. Emotion **12**, 23–27 (2012)
17. Latané, B., Darley, J.M.: Group inhibition of bystander intervention in emergences. J. Pers. Soc. Psychol. **10**, 215–221 (1968)
18. Danziger, S., Levav, J., Avnaim-Pesso, L.: Extraneous factors in judicial decisions. Proc. Natl. Acad. Sci. U.S.A. **108**, 6889–6892 (2011)
19. Williams, L.E., Bargh, J.A.: Experiencing physical warmth promotes interpersonal warmth. Science **322**, 606–607 (2008)
20. Schwitzgebel, E., Cushman, F.: Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. Mind Lang. **27**, 135–153 (2012)
21. Schwitzgebel, E., Cushman, F.: Philosophers' biased judgments persist despite training, expertise and reflection. Cognition **141**, 127–137 (2015)
22. Tobia, K., Buckwalter, W., Stich, S.: Moral intuitions: are philosophers experts? Philos. Psychol. **26**, 629–638 (2013)
23. Tenbrunsel, A.E., Diekmann, K.A., Wade-Benzoni, K.A., Bazerman, M.H.: The ethical mirage: a temporal explanation as to why we are not as ethical as we think we are. Res. Organ. Behav. **30**, 153–173 (2010)
24. Pronin, E., Gilovich, T., Ross, L.: Objectivity in the eye of the beholder: divergent perceptions of bias in self versus others. Psychol. Rev. **111**, 781–799 (2004)
25. Sudha, K.S., Khan, W.: Personality and motivational traits as correlates of workplace deviance among public and private sector employees. J. Psychol. **4**, 25–32 (2013)
26. Milam, A.C., Spitzmueller, C., Penney, L.M.: Investigating individual differences among targets of workplace incivility. J. Occup. Health Psychol. **14**, 58–69 (2009)
27. Bazerman, M.H., Tenbrunsel, A.E.: Blind spots: why we fail to do what's right and what to do about it. Princeton University Press, Princeton (2011)
28. Drumwright, M., Prentice, R., Biasucci, C.: Behavioral ethics and teaching ethical decision making. Decis. Sci. J. Innov. Educ. **13**, 431–458 (2015)
29. Tenbrunsel, A.E., Messick, D.M.: Ethical fading: the role of self-deception in unethical behavior. Soc Justice Res. **17**, 223–236 (2004)
30. Hagendorff, T.: AI virtues: the missing link in putting AI ethics into practice, arXiv 1–20 (2020)
31. Loe, T.W., Ferrell, L., Mansfield, P.: A review of empirical studies assessing ethical decision making in business. In: Michalos, A.C., Poff, D.C. (eds.) Citation classics from the journal of business ethics, pp. 279–301. Springer Netherlands, Dordrecht (2013)
32. Treviño, L.K., Weaver, G.R., Reynolds, S.J.: Behavioral ethics in organizations: a review. J. Manag. **32**, 951–990 (2006)
33. Treviño, L.K., den Nieuwenboer, N.A., Kish-Gephart, J.J.: (Un) ethical behavior in organizations. Annu. Rev. Psychol. **65**, 635–660 (2014)
34. Weaver, G.R., Treviño, L.K.: Compliance and values oriented ethics programs. Bus Ethics Q. **9**, 315–335 (1999)
35. Kish-Gephart, J.J., Harrison, D.A., Treviño, L.K.: Bad apples, bad cases, and bad barrels: meta-analytic evidence about sources of unethical decisions at work. J. Appl. Psychol. **95**, 1–31 (2010)
36. Zhang, Y., Pan, Z., Li, K., Guo, Y.: Self-serving bias in memories. Exp. Psychol. **65**, 236–244 (2018)
37. Libby, R., Rennekamp, K.: Self-serving attribution bias, overconfidence, and the issuance of management forecasts. J. Account. Res. **50**, 197–231 (2012)
38. Merritt, A.C., Effron, D.A., Monin, B.: Moral self-licensing: when being good frees us to be bad, social and personality psychology. Compass **4**, 344–357 (2010)
39. Tversky, A., Kahneman, D.: Judgment under uncertainty: heuristics and biases. Science **185**, 1124–1131 (1974)
40. Cain, D.M., Detsky, A.S.: Everyone's a little bit biased (even physicians). JAMA **299**, 2893–2895 (2008)
41. Banaji, M.R., Greenwald, A.G.: Blindspot: hidden biases of good people. Delacorte Press, New York (2013)
42. Godin, G., Conner, M., Sheeran, P.: Bridging the intention-behaviour "gap": the role of moral norm. Br. J. Soc. Psychol. **44**, 497–512 (2005)
43. Bandura, A.: Moral disengagement in the perpetration of inhumanities. Pers. Soc. Psychol. Rev. **3**, 193–209 (1999)

44. Ammicht-Quinn, R.: Artificial intelligence and the role of ethics. SJI **37**, 75–77 (2021)
45. Moore, G.E.: Principia ethica. Dover Publications, Mineola (2004)
46. Molewijk, B., Stiggelbout, A.M., Otten, W., Dupuis, H.M., Kievit, J.: Empirical data and moral theory: a plea for integrated empirical ethics. Med. Health Care Philos. **7**, 55–69 (2004)
47. Doris, J.M., Stich, S.P.: As a matter of fact: empirical perspectives on ethics. In: Jackson, F., Smith, M. (eds.) The Oxford handbook of contemporary philosophy, pp. 114–154. Oxford University Press, New York (2005)
48. Mökander, J., Morley, J., Taddeo, M., Floridi, L.: Ethics-based auditing of automated decision-making systems: nature, scope, and limitations. Sci. Eng. Ethics **27**, 1–30 (2021)
49. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat Mach. Intell. **1**, 389–399 (2019)
50. Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., Srikumar, M.: Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication No. 2020-1, SSRN Journal 1–39 (2020)
51. McNamara, A., Smith, J., Murphy-Hill, E.: Does ACM's code of ethics change ethical decision making in software development? In: Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering—ESEC/FSE 2018, ACM Press, New York, pp 1–7 (2018)
52. Gotterbarn, D., Brinkman, B., Flick, C., Kirkpatrick, M.S., Miller, K., Vazansky, K., Wolf, M.J.: ACM code of ethics and professional conduct: affirming our obligation to use our skills to benefit society, 2018, pp. 1–28. https://www.acm.org/binaries/content/assets/about/acm-code-of-ethics-booklet.pdf Accessed 1 Feb 2019
53. Cialdini, R.B.: Influence: the psychology of persuasion. HarperCollins Publishers, New York (1984)
54. Petty, R.E., Cacioppo, J.T.: Communication and persuasion. Springer, New York (1986)
55. Luhmann, N.: Social systems. Stanford University Press, Redwood City (1995)
56. Heath, C., Heath, D.: Switch: how to change things when change is hard. Broadway Books, New York (2010)
57. Brehm, S.S., Brehm, J.W.: Psychological reactance: a theory of freedom and control. Academic Press, New York (2013)
58. Dowd, E.T., Milne, C.R., Wise, S.L.: The therapeutic reactance scale: a measure of psychological reactance. J. Couns. Dev. **69**, 541–545 (1991)
59. Hong, S.-M.: Hong's psychological reactance scale: a further factor analytic validation. Psychol. Rep. **70**, 512–514 (1992)
60. Nyhan, B., Reifler, J.: When corrections fail: the persistence of political misperceptions. Polit Behav **32**, 303–330 (2010)
61. Kaplan, J.T., Gimbel, S.I., Harris, S.: Neural correlates of maintaining one's political beliefs in the face of counterevidence. Sci. Rep. **6**, 1–11 (2016)
62. Floridi, L.: Translating principles into practices of digital ethics: five risks of being unethical. Philos. Technol. **32**, 185–193 (2019)
63. Vakkuri, V., Kemell, K.-K., Kultanen, J., Siponen, M., Abrahamsson, P.: Ethically aligned design of autonomous systems: industry viewpoint and an empirical study. arXiv 1–17 (2019)
64. Rességuier, A., Rodrigues, R.: AI ethics should not remain toothless! A call to bring back the teeth of ethics. Big Data Soc **7**, 1–5 (2020)
65. Hagendorff, T.: The ethics of AI ethics: an evaluation of guidelines. Mind. Mach. **30**, 457–461 (2020)
66. R. Ochigame, The Invention of "Ethical AI": How Big Tech Manipulates Academia to Avoid Regulation, 2019. https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/ Accessed 7 Jan 2020
67. Abdalla, M., Abdalla, M.: The grey hoodie project: big tobacco, big tech, and the threat on academic integrity. arXiv 1–9 (2020)
68. Hao, K.: In 2020, let's stop AI ethics-washing and actually do something, 2019. https://www.technologyreview.com/s/614992/ai-ethics-washing-time-to-act/ Accessed 7 Jan 2020
69. Wagner, B.: Ethics as an escape from regulation: from ethics-washing to ethics-shopping? In: Hildebrandt, M. (ed.) Bein profiled: cogitas ergo sum, pp. 84–89. Amsterdam University Press, Amsterdam (2018)
70. Kleinberg, J.M., Mullainathan, S., Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. arXiv 1–23 (2016)
71. Gigerenzer, G.: Risk savvy: how to make good decisions. Viking, New York (2014)
72. Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mokander, J., Floridi, L.: Ethics as a service: a pragmatic operationalisation of AI ethics. Mind. Mach. **31**, 239–256 (2021)
73. Hallensleben, S., Hustedt, C., Fetic, L., Fleischer, T., Grünke, P., Hagendorff, T., Hauer, M., Hauschke, A., Heesen, J., Herrmann, M., Hillerbrand, R., Hubig, C., Kaminski, A., Krafft, T.D., Loh, W., Otto, P., Puntschuh, M.: From principles to practice: an interdisciplinary framework to operationalise AI ethics, pp. 1–56. Bertelsmann Stiftung, Gütersloh (2020)
74. Zicari, R.V.: Z-inspection: A holistic and analytic process to assess Ethical AI. Mindful Use of AI, 2020. http://z-inspection.org/wp-content/uploads/2020/10/Zicari.Lecture.October15.2020.pdf accessed 24 Nov 2020
75. Studdert, D.M., Mello, M.M., Sage, W.M., DesRoches, C.M., Peugh, J., Zapert, K., Brennan, T.A.: Defensive medicine among high-risk specialist physicians in a volatile malpractice environment. JAMA **293**, 2609–2617 (2005)
76. Vayena, E., Blasimme, A., Cohen, I.G.: Machine learning in medicine: addressing ethical challenges. PLoS Med. **15**, 1–4 (2018)
77. Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S.J., Lermer, E., Coughlin, J.F., Guttag, J.V., Colak, E., Ghassemi, M.: Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digit. Med. **4**, 1–8 (2021)
78. Bonnefon, J.-F., Shariff, A., Rahwan, I.: The moral psychology of AI and the ethical opt-out problem. In: Liao, S.M. (ed.) Ethics of artificial intelligence, pp. 109–126. Oxford University Press, Oxford (2020)
79. Bonnefon, J.-F., Shariff, A., Rahwan, I.: The social dilemma of autonomous vehicles. Science **352**, 1573–1576 (2016)
80. Di Fabio, U., Broy, M., Brüngger, J., Eichhorn,U., Grunwald, A., Heckmann, D., Hilgendorf, E., Kagermann, H., Losinger, A., Lutz-Bachmann, M., Lütge, C., Markl, A., Müller, K., Nehm, K.: Ethik-Kommission Automatisiertes und Vernetztes Fahren, 2017.
81. Corbett-Davies, S., Goel, S.: The measure and mismeasure of fairness: a critical review of fair machine learning. arXiv 1–25 (2018)
82. Briceño, J.: Artificial intelligence and organ transplantation: challenges and expectations. Curr. Opin. Organ. Transplant. **25**, 393–398 (2020)