AI & SOCIETY (2023) 38:1503–1505
https://doi.org/10.1007/s00146-023-01630-3

CURMUDGEON CORNER

Beta-testing the ethics plugin

Keith Begley[1]

Received: 30 May 2021 / Accepted: 9 January 2023 / Published online: 25 January 2023

The three main kinds of theory in normative ethics, namely, consequentialism, deontology, and virtue ethics, are often presented as the 'palette' from which we may choose, or use as a starting point for an investigation. However, this way of doing ethics and philosophy, by the palette, may be leading some of us astray. It has led some to believe that all that there is to ethics, and to ethics of AI, is given in terms of these already devised petrified categories of theory. It has also led others to abandon normative ethics and philosophy altogether and to resort to descriptive methods that are then used to justify action. I wish to argue that (1) we should not abandon traditional philosophical approaches, but (2a) this does not entail that the petrified palette should constitute the beginning of our philosophical investigations. Further, (2b) I recommend a non-methodological approach in which it is instead radical questions that spur these investigations,[1] which arise through consideration of the practical actions (potential or otherwise) of machines and their programmers.

It is prudent not to begin from, or by restricting the space of investigation to, the palette.[2] The results of beginning from this kind of petrified thinking can be seen, for example, in a recent attempt to avoid the inflexibility of the three "single-component theories" by 'combining' them in the descriptive Agent-Deed-Consequence model (ADC) (Dubljević and Racine 2014, as cited in Wernaart 2021; Dubljević et al. 2018, as cited in Aliman and Kester 2022), which is proposed for use in autonomous vehicles (Dubljević 2020, as cited in Wernaart 2021). However, the authors have overlooked the fact that each of these three can already acknowledge agents, deeds, and consequences, but in ways that are often incompatible.[3] The issue here is at root a methodological one caused by the petrified starting point. The authors begin from a perception of "deadlocked moral intuitions" elicited by the constituent theories of the palette that are "unsuccessful in both establishing their supremacy and in proving the moral judgements/intuitions invoked by opposing schools false" (Dubljević and Racine 2014, 5, 12 and 17, as

[1] What I mean by 'radical questions' ('radical' meaning 'root'), are those that articulate dilemmas that call into question even our ability to answer by pointing to examples, exemplars, or everyday standards. The undermining of this ability, by calling into question those standards, is what makes a question radical. Ethical questions are often prime instances, though not the only ones. Consider the question of whether or not some particular person, thing, or act is good. This cannot always be answered by pointing to another person, or this or that thing or act. Such cases must be answered at least in part by considering the further question 'What is Good?', otherwise we would be left attempting to compare particulars without a basis of comparison. Such radical questions are readily found in ancient Greek philosophy, especially in Plato and Aristotle, where they are called *aporiai*. These *aporiai* are amenable only to answers that are general, unitary, and explanatory standards for judgement. I am indebted to Vasilis Politis for his notion of *radical aporia*. For more on this topic, I recommend especially his two recent monographs on Plato.

[2] The analytical tradition itself is not much older than computer science and, at root, both arose from the same advances in formal logic and mathematics at the end of the nineteenth century. It is rarely mentioned that the tripartite division was invented during the twentieth century and may often be misleading. Anscombe coined the term 'consequentialism' in her article 'Modern Moral Philosophy' from 1958, which was also a locus for *virtue ethics*. The term 'deontology' was first introduced by Bentham, in his *Chrestomathia* from 1816, as a synonym for 'Dicastic Ethics', addressed to the will, as opposed to the 'Exegetic' or 'Expository', addressed to the understanding. The term was reintroduced by C. D. Broad, in his *Five Types of Ethical Theory* from 1930, and used in reference to types of action considered regardless of their consequences. I say all this not in simple-minded veneration of the past or aversion to recency, but merely to point out that beginning from such a palette restricts our view, has the potential to beg important questions, and sometimes commits the reverse of those fallacies.

[3] Take utilitarianism, which, although recognising agents, is agent-neural. That is, it cannot be combined with a non-agent-neutral theory. It follows from such incompatibilities that, whatever the ADC approach is, it is strictly speaking not a combination of the three normative approaches. A full investigation would be beyond the scope of the present short article. However, it should be clear to see that this objection can be almost symmetrically extended to (at least) the other two main kinds of normative theory.

✉ Keith Begley
keith.begley@mu.ie; begleyk@tcd.ie

[1] Department of Philosophy, Maynooth University-National University of Ireland Maynooth, Maynooth, Co. Kildare, Ireland

cited in Wernaart 2021). Thereby, they treat the issue as one regarding the kinds of theory themselves and not, for example, a dilemma regarding a particular event or action that sets their investigation in motion.

Some researchers have even resorted to so-called 'non-normative' or descriptive ethics in an attempt to escape such perceived deadlocks, for example, the 'Augmented Utilitarianism' (AU) framework (Aliman and Kester 2019, as cited in Wernaart 2021, n. 92).[4] However, this is a myopic manoeuvre, because advancing beyond mere descriptions of actions, beliefs, and intuitions, to treating them as guides to, or standards or criteria for, what should happen, entails that these are *ipso facto* treated as normative criteria.[5] Such a method can only arrive at a description of what *is*, that is, the aggregate actions, beliefs, and intuitions of a particular population at a particular time, etc., but this need not inform what *ought* to be done or what is *Good*, i.e., the traditional subject matter of ethics.[6] Furthermore, making such a claim would carry a questionable commitment to a socially constructed nature of morality.[7]

A more recent expression of AU makes it clear that this allows for moral relativism because it is intended to be agnostic and ideally applicable to *most* ethical frameworks that might be selected by a society.[8] The authors also claim that AU does not have philosophical aspirations, yet it is said to be focused on deliberations about what morality *is* (Aliman and Kester 2022, 65), and is clearly intended to have a normative function with regard to AI, whether or not the framework itself embodies specific normative claims.

It is sometimes suggested that the problem with normative theories is their operationalisation. That is, that the problem consists of their not clearly being amenable to being put into terms interpretable by a computer. The assumption here is that we have been presented with at least a list of the names of the possible solutions. All we would have to do is either pick a team, combine the approaches (e.g., in ADC), or avoid them altogether by resorting to mere descriptive ethics (e.g., in AU). That is, to plug in the ethics and begin beta-testing.

There is certainly a shared responsibility to employ technology in an ethical manner. However, we should not pretend that the compulsory questions regarding whether it is possible, practicable, and ethical, to mathematicize ethical reasoning, have already been answered. Designing machines that perform operations that are functionally equivalent to an idealized ethical machine, is certainly a reasonable intermediate technological goal, especially in view of the fact that we are already implementing machines in ethically significant contexts and so have no choice but to improve them.[9] However, it would be both ethically questionable and philosophically suspect to consider them to be ethical reasoners (cf. Lokhorst 2011, as cited in Wernaart 2021) or moral agents (cf. Wernaart 2021), on that basis.

The approaches that we have discussed either begin from the palette or attempt to avoid normative ethics altogether. The antidote to this methodology is to begin instead from the practical actions of machines and their programmers (potential or otherwise). Theoretical distinctions or posits should only be proposed in service of answering specific questions that arise in the course of the investigation.[10] For example, asking whether it is ethical to teach an AI to reason ethically will immediately involve the further question of whether

[4] Wernaart (2021, 9) notes that "they move away from the debate on what a machine *should* do, and instead focus on *what we want* it to do" (My emphasis added). This is, I believe, a misunderstanding of the situation caused by the notion that because one is avoiding explicitly using the normative palette, one is doing non-normative ethics.

[5] Aliman and Kester say that "Instead of specifying what an agent ought to do, AU helps to identify what the current society *should want* an (artificial or human) agent to do if this society wants to maximize expected utility" (Aliman and Kester 2019, 4, as cited in Wernaart 2021, n. 92). There are clearly two normative elements already involved here. The first is that AU is intended to help to identify what *should* happen. The second is that it relies upon prior normative conceptions of utility.

[6] The authors of this approach do not address this well-known fallacy of the derivation of an *ought* from an *is*, which goes back to Hume (*Treatise* III.1.1). In light of this, their approach would, at the very least, require a sustained attempt at philosophical justification in that regard. For example, Steven Kraaijeveld is more careful in noting that although empirical findings may be useful for informing some areas of ethical enquiry, the role is a supportive one that does not generate normative conclusions in the absence of prior normative premisses.

[7] This is wisely recognised by the authors of the Moral Machine experiment, which is "a multilingual online 'serious game' for collecting large-scale data on how citizens would want autonomous vehicles to solve moral dilemmas in the context of unavoidable accidents" (Awad et al. 2018, 59, as cited in Wernaart 2021). When reflecting on their study, Bonnefon said that "Considering our results to be normative would amount to saying that there is no objective definition of what is moral or immoral; morality would therefore be a social construction, limited in space and time. Consequently, morality would be whatever the population thinks is moral, here and now. Moral Machine would therefore be the arbiter and standard of morality. Personally, I think this interpretation is insane, but the philosophical debate is beyond me. All that my coauthors and I could do was say that we never had any such ambitions!" (Bonnefon 2021, p. 132).

[8] The palette is also mentioned in this context due to the salience in the debate of these so-called "classical" normative frameworks; see footnote 2. There is a potential here that the framework itself will rule out certain views, but this is not determinable on the present characterizations that are available.

[9] I am grateful to an anonymous reviewer for pushing me to clarify this point, especially in view of the fact that some researchers may be unaware of it.

[10] It could be objected that, because we are concerned with ethical issues, we as matter of course need to appeal to prior ethical theories. However, ethical theories are the results of prior philosophical investigations, not the beginning of them. Such objections would not pass muster in historical contexts in which there were no explicit prior theories, or they were not named, etc.

such matters can be taught, and so what it is to reason ethically and what is *Good* are also in question.[11] This avoids such investigations becoming embroiled in issues regarding how to choose between theoretical approaches, or how to 'combine', sublate, or avoid them, etc. These purely theoretical tangles are not what should motivate our enquiries as we further develop the ethics of AI. Instead, it should be considered that what is being opened up is an entirely new field of practical action that will eventually surpass that of humans in many respects. Hence, it will be necessary to consider the ethics of actions beyond those hitherto considered, and perhaps even radical questions not previously encountered in considerations of human action.

Acknowledgements The author wishes to thank a number of anonymous reviewers for their helpful and constructive comments, and Steven Kraaijeveld for his helpful comments and kind assistance, including for sharing his article 'Experimental philosophy of technology'. The author also wishes to thank Leon Kester for his kind assistance in obtaining a copy of some of his work.

Curmudgeon Corner Curmudgeon Corner is a short opinionated column on trends in technology, arts, science and society, commenting on issues of concern to the research community and wider society. Whilst the drive for super-human intelligence promotes potential benefits to wider society, it also raises deep concerns of existential risk, thereby highlighting the need for an ongoing conversation between technology and society. At the core of Curmudgeon concern is the question: What is it to be human in the age of the AI machine? -Editor.

Funding N/A.

Data availability N/A.

Declarations

Conflicts of interest N/A.

Ethics approval N/A.

Consent to participate The sole author grants their consent.

Consent for publication The sole author grants their consent.

References

Aliman N-M, Kester L (2022) Moral Programming. In: Wernaart B (ed) Moral Design and Technology. Wageningen Academic Publishers, Wageningen

Bonnefon J-F (2021) The car that knew too much: can a machine be moral? The MIT Press, Cambridge, MA

Wernaart B (2021) Developing a roadmap for the moral programming of smart technology. Technol Soc 64:101466

[11] Thereby, the ancient *aporiai* re-emerge. See footnote 1 on *radical questions*.