Digital Society (2022) 1: 3
https://doi.org/10.1007/s44206-022-00003-5

ORIGINAL PAPER

Experimental Machine Ethics and the Problem of Entrenchment

Christoph Merdes[1]

Received: 29 October 2021 / Accepted: 10 May 2022 / Published online: 6 July 2022

Abstract

The increasing prevalence of autonomously operating artificial agents has created the desire and arguably need to equip such agents with moral capabilities. A potential tool to morally sanction an artificial agent as admissible for its tasks is to apply a so-called moral Turing test (MTT) to the machine. The MTT can be supported by a pragmatist metaethics as an iteratively applied and modified procedure. However, this iterative, experimentalist procedure faces a dilemma due to the problem of technological entrenchment. I argue that, at least in certain important domains of application, the justification of artificial moral agents requires their deployment, which may entrench them and thereby undermine the justificatory process by hindering its further iteration.

Keywords Moral Turing test · Technological entrenchment · Pragmatist metaethics

1 Introduction

What moral code[1] ought a (semi)autonomous artificial agent to follow? And how should it be tested whether it *does*? This tandem of questions is one of the central problems of machine ethics. Autonomous, in this context, describes a machine that operates without effective[2] human control. The same machine may sometimes

[1] I use the term code somewhat loosely here to denote the set of procedures the machine follows that we recognize as explicitly moral; this does not necessarily require morality to be fully codifiable in the sense of stating a comprehensive set of determinative moral rules. This should not be simply assumed, as some philosophers call into question if morality is even codifiable in that manner (cf., Fricker, 2007, p.73ff.).

[2] What is considered effective control is context dependent. For a lethal drone, it could require a human-in-the-loop who always has to make the ultimate decision to discharge a weapon, whereas in an automated moderation system, it can be sufficient to have a human to address complaints to and correct decisions made by the machine.

✉ Christoph Merdes
christoph.merdes@fau.de

[1] Center for Applied Philosophy of Science and Key Qualifications, Friedrich-Alexander-Universität Erlangen-Nürnberg, Bismarckstrasse 12, 91054 Erlangen, Germany

operate under effective human control and thus non-autonomously and at other times operate autonomously. A machine may also operate without effective human control, but under certain conditions actively ask a human for input.

There are several obstacles from a philosophical point of view to give a clear and simple answer to the two questions: First, there is widespread disagreement in moral judgment, be that on the level of high theory between utilitarians and Kantians, or be it at the very concrete level of everyday moral discourse about obligations towards refugees. Second, there is a deep conceptual problem of what constitutes a moral agent and the ability to pass adequate — if fallible — moral judgment.

As an alternative to providing an ultimate and uncontroversial answer to these disagreements for the purpose of dealing with pressing problems of rapid automation, the moral Turing test (MTT) has been suggested[3]. There are several variations of the test (Allen et al. (2000), Wallach and Allen (2008); see also Arnold and Scheutz (2016) for a critical discussion of the MTT framework and extensions to it), but for a basic understanding, the most simple version suffices:

> *An observer textually[4] poses moral judgment problems to both a human and a machine. If, after a prolonged sequence of interrogation, the observer cannot distinguish between human and machine above chance level, the machine is considered capable of moral judgment.*

Of course there are several problems with this test — for instance, we have to assume that we can agree on the human subject and the observer both being moral agents of some adequacy — but it gives us a practical criterion to apply to a given machine. My focus in this paper is not on the issues with the test itself, but to give some philosophical grounding to it in the form of a pragmatist metaethics and explore one of the central problems exposed by this framework: the phenomenon of technological entrenchment as a threat to an experimental progression in the design of artificial moral agents (AMAs) and as a result, the undermining of the justification provided by that process.

To be clear, in the following, the MTT framework is considered to encompass not only the basic version stated above and similar variants under clinical conditions, but any framework of recognizing suggested AMA as sufficiently competent moral agents that is built on qualified human judgment in a comparative test under conditions where the mere fact that the machine is a machine and not a human is removed from the judgment procedure. The full protocol should likely consist of a combination of more idealized, highly controlled tests such as the one described above, and increasingly realistic test environments — the latter of which bring about the central problem discussed in this paper.

To note, while this is a great practical difficulty, in theory, we can clearly distinguish between properties that the machine has qua machine — certain computational

[3] In analogy to the Turing test for general intelligence (Turing, 1950), an attempt to deal with another elusive concept.

[4] The medium can be a different one; it is merely supposed to exclude irrelevant information, at its most simple the appearance of a human and a machine.

capacities, for instance — and the fact that it *is* a machine. The former can legitimately figure in the comparative judgment in the MTT framework, the latter cannot, as it is not by itself significant for the adequacy of an agent's moral judgment.[5]

In rough outline, the central problem is that developing a properly justified AMA requires its experimental deployment; however, a deployment that enables an effective assessment has to take on, in many domains, such a scale that it makes it difficult to revert if moral objections to the AMA arise. This is a problem structurally very similar to the so-called Collingridge dilemma (Liebert & Schmidt, 2010); this is the label for a type of problem of control in technology assessment. It states that the development and deployment of technology is difficult to control, because when it is still easy to change its path, knowledge is lacking, but the required knowledge is only acquired when it is already very difficult to swerve. It has become *entrenched*. The problem of AMA is similar, though differs in that it is ultimately grounded in an internal tension of the *moral* justification of an AMA implementation.

A bit more formal, the argument can be represented in the following progression of reasoning:[6]

1. Some version of the MTT is the best method to justify an AMA.
2. The MTT framework is best justified by reference to a pragmatist metaethics.
3. Pragmatist metaethics requires that the technology remains reversible, as it recognizes that any choice is fallible, and committing irreversibly to a particular version of AMA technology would be tantamount to assert certainty about its adequacy.
4. However, the practical implementation of the MTT requires not only laboratory experiment, but ultimately a practical deployment.
5. That practical deployment leads to technological entrenchment.
6. Technological entrenchment limits, constrains or even precludes reversibility.
7. (from 6 and 3) The technology cannot be justified on the basis of a pragmatist metaethics.
8. (from 7 and 2) The technology cannot be justified via an MTT.
9. (from 8 and 1) The technology cannot be justified.

The paper proceeds as follows in explicating the central concepts of this line of reasoning and to lend support to the premises: In Sect. 2, the required concepts and theoretical elements of a pragmatist metaethics are assembled and used to reconstruct a justification of the moral Turing test (MTT) as a device for the assessment of AMAs. Next, the phenomenon of entrenchment and some of its causes are analysed (Sect. 3). Section 4 offers a full statement and analysis of the dilemma using

[5] There are, of course, a number of properties that are quite difficult to ascribe to machines, such as qualitative experience, consciousness and similar characteristics that we liberally ascribe to humans. Depending on the theory of moral agency, these properties *are* significant, and thus contribute to the practical difficulty of creating appropriate test conditions. Theoretically, however, they are still separate and there are plausibly non-human agents that have or could have those same properties.

[6] To note, in particular, the first two premises of this argument are certainly controversial. In particular, one might object to the assessment that the disadvantages of an MTT approach outweigh the problems resulting from disagreement for theory-driven approaches.

autonomous vehicles as an illustration. The concluding remarks summarize the argument and address mitigating measures to approach the dilemma.

2 Pragmatist Machine Ethics

2.1 A Pragmatist Stance

The central methodological assumption of pragmatist philosophy is that philosophical analysis is to be construed from action and practice. For instance, as James (1975) put it, to reconstruct a concept or choose between various possible reconstructions of a concept, one needs to ask for the *practical* difference it makes. Of course this is not the place to give a full reconstruction of the historically varied view summarized under the label pragmatism; rather, I will offer an interpretation of pragmatism as it applies to metaethics, in particular, the justification of ethical judgment.[7, 8]

The reconstruction of the test's justificatory power is built around three consequences of the pragmatist outlook, denoted here as *homo mensura*, *iterative experimentalism* and *pluralism*.

Homo Mensura One important consequence of taking the pragmatist stance is to rely on human judgment as the arbiter of justification. This follows from the rejection of any claim to a transcendent truth. Such a claim would in particular not be fallible and hence not revisable; but there is no method to achieve such infallible insights, as human action is involved in all our judgments, and its consequences are always contingent. Our reasoning, hence, may always need revision; rather than an absolute standard of infallible truth, we therefore rely on a fallible act of interim acceptance.

This analysis applies to the acceptance of moral judgments as to the acceptance of factual claims alike. It does not imply, though, that anything can be accepted at any time by a given agent. The pragmatist view still allows us to set intersubjectively valid standards of acceptance that become more demanding and sophisticated over time, when our insight into the subject matter improves. For instance, if someone wishes to assert a claim that should be accepted by the community of physicists, they will have to comply with the standards for acceptance required by that group. These standards of acceptance are therefore not eternal — in fact they ought to be changing over time — nor are they universally valid across all domains of life. These are indeed the other two main aspects of my interpretation of pragmatism to be laid out for a reconstruction of the MTT.

[7] The foundation of my analysis is Dewey (1929), ch. X and Dewey (1939), but it is adapted to the context of the MTT, a task that Dewey presumably did not foresee in his writings. I do not claim to fully capture Dewey's overall ethical view, but merely to draw on his arguments in the cited writings.

[8] For a more general reconstruction of the pragmatists' metaethics, cf. Sepielli (2017).

It is worth noting that this view of justification as grounded in qualified acceptance contradicts both an intuitionistic theory that views raw, unsophisticated moral intuitions as the foundation of justification and views that strive to justify moral judgments with some method entirely transcending human subjectivity. The latter, according to the pragmatist, is impossible, whereas the former is insufficient. Intuitions may serve as a starting point for a process of refinement, but they should not be taken as its unquestionable horizon.[9]

Iterative Experimentation The pragmatist stance suggests to understand moral judgment by reference to action; specifically, to ascribe moral agency to an agent, an operational procedure has to be given to determine its status. This would not quite justify calling it experimental, but the procedure needs to be reproducible and, like an experiment, be accepted as a stand-in or model of, in our case, the application of moral judgment in real circumstances.

The procedure should be iterative, that is, repeated and improved consecutively in the face of problems recognized in the agents sanctioned as moral agents. This demand is grounded in the recognition of fallibility and contingency. To be more precise, our very capacity for moral judgment requires an iterative procedure to progress, as the experiences required for further progression become, at least in many cases, only accessible once we reached a certain level of moral sophistication already. For instance, the formal implementation of certain rights in the law has enabled us to understand that it is insufficient to realize human freedom.[10]

Testing for artificial morality with a literal experiment provides an operationalization of this idea of moral reasoning and progress; both in that the experimental design can be improved upon, but also the tested machines can be improved to accord better and better with the standard set by the experiment. This projects the vision of moral machines and MTTs proceeding in tandem, being revised iteratively to advance the capabilities of those machines while also being open for human moral judgment to adapt and improve in the process.

Pluralism Both the founders of pragmatism and its more contemporary disciples (cf., Rorty, 2000, ch. 7 for an example) argue that there are legitimate grounds for pluralism across our various domains of knowledge in accordance with their varying practices of justification. With regard to ethics, several dimensions of such pluralism are recognizable: between ethics and the empirical sciences, between domains of application (warfare, elder care, stock trading) and between individuals relying on different, though rational, practices of justification.

The MTT is not able to *deal* with this pluralism in the sense of resolving it; when different observers performing the test come to different conclusions, the MTT does

[9] This may remind the reader somewhat of constructivism and the notion of reflective equilibrium. There are indeed structural similarities, but also some important distinctions; cf. Proulx (2016) for a discussion of the relationship between the two stances.

[10] I do not claim that it was impossible to come to this conclusion before those rights were implemented; however, the contingent empirical reality would have been crucial to confirm those speculative insights.

not offer a resolution to this conflict itself. But it captures the disagreement, makes it explicit and therefore pins down what has to be decided in some other fashion. It thus acknowledges that artificial morality is not a problem that can be outsourced to engineers and philosophers — even if we assumed, quite optimistically, that those could agree — but needs to be dealt with by those affected by the machines in question.

Furthermore, pluralism regarding domains of ethics implies that a positive MTT result should not be assumed to be transferable to any domain. The sanction provided by the MTT is *domain specific*, and thus should not be taken as a universal attribution of moral competence. This point is well in line what would normally be assumed for humans when it comes to questions of applied ethics in highly complex and ethically sensitive domains. It cannot simply be assumed that someone who is very competent when it comes to questions of biomedical ethics is automatically to be considered competent in environmental ethics. While it seems plausible that certain aspects of moral reasoning transfer well from one domain to another, competency for moral judgment is not exhausted by those, and thus its transferability cannot be assumed.

2.2 Experimentalist Machine Ethics

The moral Turing test presents, within this pragmatist framework, an operationalization of our notion of a moral agent. The basic test presented in the introduction is certainly insufficient to capture all aspects of this concept, and there exists a number of variants capturing various aspects better, such as the comparative MTT (Allen et al., 2000)[11] or a demand to provide not only judgments, but also justifications. It is not the purpose of this paper to systematically analyse these variants, but it is instructive to take note of the reality of the systematic variation of the procedure. With these remarks in mind, we can turn to the reconstruction of a justification of an MTT.

First, it is rather obvious that the MTT satisfies the demand that human acceptance be the relevant arbiter. Human agents function both as a subject for a comparison and more directly as the observer and judge in the comparison task. The pragmatist reconstruction suggests that conditions ought to be imposed on those subjects, namely that they are recognized prior to the test as competent members of their moral community, and need to adapt to the changes in that community; in our context, that concerns in particular their interaction with and judgment of the artificial agents sanctioned by prior instances of a procedure in the MTT framework.

There are a number of quite obvious problems surrounding the notions of competence and moral community; for instance, it is controversial whether moral competency can be taught, whether it refers to sound moral reasoning or adequate judgment and various other issues. Similarly, the notion of moral community is

[11] See Arnold and Scheutz (2016) for a critical discussion.

highly problematic (even more so than the community of physicists used as an example before). Who is a member of one's moral community, how universal is it, what different kinds of moral status may be represented within a moral community, and so forth. These problems are serious, but I do not intend to solve them in this paper. These are problems that have to come up in any serious approach to justify the deployment of autonomous artificial agents, and at the very least, an MTT grounded in the pragmatist stance offers a clear map of the problems at hand.

Second, while the MTT is not inherently iterative, it is framed as an experiment, and it clearly can be iterated and improved upon over time. The changes can logically be separated into modifications of the test structure and changes to the human subjects involved. In the former category are the variants of the MTT mentioned above. While they are often informed by insight from prior actions, these can in principle be explored theoretically and in advance. The latter type of iterative change is a product at least in part of the consequences of prior tests and experiences with the machines deployed on the basis of those tests. It is to be expected that human perception of artificial agents and expectations towards their (moral) behaviour become more sophisticated over time, requiring new instances of the MTT.[12]

The MTT as an operational definition is thus supported by the pragmatist framework, though that framework demands an implementation that has built into it the further modification and iteration of the operationalization. It suggests that an AMA, even if it is at one point sanctioned by an acceptable MTT, may lose this status in the future. While this might seem quite obvious, it will become a key problem when we turn to the problem of entrenchment.

Finally, the MTT offers some resources to reflect pluralism. Primarily, unlike, for instance, a verificationist approach, it allows human agents who are not technical experts to assess the capabilities of the machine. Also, there is no temptation to transfer the same, supposedly universal set of ethical principles from one domain to the other; acceptance in one domain does not imply acceptability in another one. These are notably rather weak points of support. The realization of pluralism depends more on the implementation of MTTs and the customization capabilities of the machines produced.

Modifications of the test are reconstructible in the pragmatist framework as suggestions to understand the concept of moral agency or related concepts in different ways. As they change the way the test is executed in quite concrete ways, such modifications are automatically recognized as significant changes by the pragmatist view. To give just one illustrative example, Arnold and Scheutz (2016) discuss the ramifications of a test variant that has the agents not offer hypothetical judgments, but indeed perform morally sensitive actions. A way of reconstructing this

[12] There are already, though rather particular, insights into the human moral evaluation of machines (Awad et al., 2018; Voiklis et al., 2016). While these are quite interesting and may offer a useful starting point, there is no reason to expect that human judgment will not adapt to real experiences with supposedly moral machines.

modification is that it assumes that actual *agency* is quite different from capacity for mere judgment.[13]

To be clear, this is a key argument against the verificationist approach preferred by Arnold and Scheutz (2016), not because we may have practical concerns about its feasibility[14], but because it does not have an answer to the problem of theoretical disagreement in normative ethics. If it is already assumed that, for instance, a certain version of utilitarianism is the correct system of morality, verifying that the machine operates according to that theory would, combined with information about its sensors, actuators and how the theory is embedded in the rest of its architecture, provide a justification for its behaviour. But the assumption is entirely unwarranted. As an engineering approach to construct a moral machine, verification may still be useful, and its utility should not be dismissed. However, when it comes to the justification of machine deployment in a world where no standard can be laid out to verify against, it cannot be the ultimate measure for such justification.[15]

As noted before, there are several other variants already in the conversation, and the sanction of an AMA may require passing multiple different variants to establish different features of moral agency. However exactly the procedure may look in detail, at this point, we have the outline of a process to sanction AMAs on pragmatist grounds: To design a modular, modifiable MTT, execute it involving the relevant human population, and repeat and modify the process conditional on the results of the machines' deployment in terms of qualified acceptance. The logic of this justification relies critically on the reference to the possibility of reverting choices that turn out or become unacceptable.[16] This process, as will become clear over the next sections, runs into the problem of technological entrenchment.

3 Technological Entrenchment

3.1 What Is Entrenchment?

The notion of entrenchment is not new in philosophy of science and technology. It describes a certain type of path-dependent development, where at some point in

[13] In fact, this difference is one of the major issues that their criticism of MTTs and test frameworks in general builds on, namely the gap between judgment and action. They are correct in recognizing this gap; however, if, as I argue, verification cannot deliver due to the problem of deep moral disagreement, this gap is a shortcoming that might be inevitable unless we choose to reject the deployment of machines that would require moral capabilities entirely. I discuss this point in more detail in the conclusions.

[14] Compare, however, the argument by Wallach (2017) that the complexity of a robotic agent operating in an open environment will remain substantially unpredictable.

[15] While Arnold and Scheutz (2016) acknowledge the problem of disagreement, they do not offer any solution. This seems to be further evidence for the severity of the issue.

[16] This is not dissimilar to common inductive reasoning; an epistemic agent's justification for accepting a hypothesis based on available evidence is contingent on their willingness to revise their acceptance on the basis of future countervailing evidence. Without this revisability, even if the agent's acceptance is firmly grounded in available evidence, it is not justified because the inclusion of evidence is arbitrarily cut off at one point in time.

history, a choice is made to implement a certain technology, which, due to one or more of a variety of factors, becomes very difficult to revert.

Entrenchment is not limited to technologies; it is a common phenomenon of the human condition. For instance, certain beliefs tend to entrench themselves as it becomes difficult to revise them for a variety of logical and psychological reasons. A belief that is situated at the core of an agent's network of beliefs is hard to revise. Similarly, social norms and institutions tend to entrench themselves as a matter of practical necessity. A social norm that is easily revisable cannot be strongly habituated and hence is unstable and will tend to fail to serve the purpose of a social norm.

While these types of entrenchment are not of immediate relevance to my argument, they highlight an important question: Is there a genuine alternative to entrenchment? In some sense, the status quo will have a tendency to be entrenched, be that with regard to thought, social practice or technology. That much is trivial.

However, there is a distinction to be made; while we can choose not to pursue a given goal via technological means (and hence not entrench any technology in particular), as long as there is a society, there will be some set of social norms. Depending on our world view, more revertible social norms may also be seen as favourable, but the presence of somewhat entrenched social norms seems sociologically inevitable. In that regard, even though we may also speak of entrenchment regarding other aspects of human life, we should be precise as to what the alternatives are to entrenching one choice over another. I discuss this point in further detail regarding technological entrenchment below.

Next, before assembling the key factors of entrenchment, it is worth clarifying our terms.

First, not much hinges on the exact definition of technology. It is assumed to refer to a stable, socio-technical system aimed at an identifiable set of purposes[17]. The deployment of a technology in that sense involves the production of technological artefacts, an infrastructure to enable the deployment of those artefacts and training of humans in skills and social norms to use and maintain the technology.

Second, the term *difficult* is intentionally ambiguous and vague. It is vague to express that difficulty is a matter of degree, even though when it comes to decision-making, it has to be determined *how* difficult is *too* difficult. For instance, the replacement of an operating system in a corporate bureaucracy is quite expensive, though whether it is considered difficult so as to consider the technology as entrenched depends on the context. The term is intentionally ambiguous because it allows for a variety of notions of difficulty.

Those include, for instance, economic cost, sustainability, technical or scientific feasibility or moral admissibility. To some extent, these kinds of difficulty translate into each other; a technical difficulty, e.g. might be addressed by a larger investment of resources; an unsustainable choice may be morally inadmissible, and so forth. It seems, however, that these translations tend to be imperfect[18] and there is

[17] For a more detailed account of the concept of technology, cf. Radder (2009).

[18] In the case of worries of technical feasibility, more resources may solve the problem, but it is likely uncertain if they will.

no reason to attempt to unify them into a uniform concept. Ultimately, various kinds and degrees of difficulty are expressed and aggregated in the decision-making process, and the characteristics of the process — e.g. the goals of the decision makers — have to inform the weighing and combining of difficulties.

Finally, a remark on the notion of reversibility. It can be rather straightforwardly interpreted as the removal of the technological artefacts and their infrastructure accompanied by either the choice of a different technology to achieve its purposes or an adjustment of the goals that led to the adoption of the technology. This is an ideal-typical characterization of reverting a technological choice, and in practice, the reversal can be an extended process that ultimately leaves significant remnants. While these technological residuals may be quite serious[19], it is still useful to understand such cases as reverting a technology choice.

It is also worth pointing out that reversibility is not a fact independent from human actions that we encounter in the world. In some respects, it is: It depends on the laws of physics or the amounts of certain resources that happen to be present in our reach. But the actual reversibility of any given technology is also heavily influenced by its design, that is, by the actions of its engineers as well as a number of broader human-dependent factors.

For instance, all else being equal, an algorithmic procedure implemented as software in a re-programmable computer will be more reversible than the same procedure implemented directly in hardware. Of course, there are other design goals that may favour the hardware implementation overall, and there are cases where software has become entrenched deeply; my point is simply that the same goal of calculating the result of an algorithmic procedure can be achieved in a more or less revertible manner, and creative engineering can therefore increase reversibility. This observation will be important when it comes to any recommendations regarding the problem resulting from entrenchment.

3.2 Factors of Entrenchment

While it has to remain rather general, a brief inquiry into some of the major factors of entrenchment will provide a useful background for the further analysis.

- Humans are adapted to the technology (acquired the relevant skills, habituated the social norms) and it would be costly (in work hours, personal effort, etc.) to move them to a different technology.
- The technological artefacts and necessary infrastructure for the new technology has to be redesigned and rebuilt, whereas no additional value can be drawn from the already developed, existing technology. In fact, disposing of the old technology may impose a vast additional cost. This implies a potentially enormous use of resources, causing not only economic cost but also raising issues of sustainability and distributive justice.

[19] Think, for instance, of nuclear waste that does not become harmless for a long time, even if nuclear power is abandoned as a technology.

- The switch to a new technology comes with a great deal of risk and uncertainty; a new technology may have a variety of unforeseen and unintended consequences. This point is dealt with at some length by Liebert and Schmidt (2010) and presents an integral part of Collingridge's dilemma as mentioned in the introduction.

One may also consider the lack of feasible alternative a factor of entrenchment, but that seems not to be very useful. The conceit of the concept is to analyse and evaluate path-dependent *choices*, and if there is no alternative technology, the question is more one of the prioritization of goals. My intention here is not to provide an empirical account of the factors in detail, but to illustrate for what reasons entrenchment occurs at all.[20] In that vein, consider the following two examples.

First, consider automotive technology in a country such as Germany. The technology encompasses a massive amount of artefacts — vehicles, the required road system, etc. — and extensive background infrastructure — car manufacturing, fuel production, etc. — and human training and adaption of social norms to govern behaviour in a world full of cars. If the German government decided that it wishes to replace the mode of transportation in some key dimension, much of the artefacts, the infrastructure and the human capacities would have to be replaced or modified. I shall discuss this case at some length with respect to autonomous vehicles in the following section, but problems of a similar kind arise with electric cars or an expansion of public transportation already.

If, for instance, Germany were to decide, as political currents suggest, that in the future, only electrical vehicles will be admitted. This requires the replacement of an enormous fleet of cars, but also significant changes in the infrastructure, e.g. to provide enough charging stations. Certain engineering skills will depreciate in the process, requiring at least some retraining and job displacement, and social norms may have to adapt in subtle ways. This is, of course, not to be read as an argument against the technology switch, towards which I take no position in this paper. It is merely to point out in which ways the specific automotive technology is entrenched; and this example is a relatively benign case, as we merely assumed a change in the engine technology, whereas the road infrastructure, city planning and many other issues that would come up in a move to more public transportation are not even touched upon yet.

The example of individual transportation by means of the combustion engine highlights an important point on entrenchment: Even if, as recent developments suggest, decisions are made to revert and move to a different technology, the cost of entrenchment is real. First, entrenchment has the technology being used beyond the point where it would be rational to enact change; second, the costs and other negative side-effects of entrenchment remain a fact even if the technology is ultimately replaced.

[20] According to Liebert and Schmidt (2010) and their interpretation of Collingridge (1980), one goal of technology assessment is to reduce or even eliminate entrenchment. While it seems plausible enough that entrenchment can be mitigated in certain cases, on a fundamental level, it seems inevitable; resources go into the development and implementation of a technology, and if the path is to be reverted, those resources are essentially lost.

The second example, taken from Winsberg (2010, ch. 6), is quite different. He discusses the entrenchment of certain model families in climate science and the path-dependent and at least sometimes value-laden choices involved. This case is different, in that the population of agents using the technology is much smaller and the relevant artefacts exist primarily in software. Whereas in the previous example, the primary issues seem to be the massive economic cost and the sustainability problem of replacing a massive amount of hardware, the key problem here appears to be human cognitive limitations and the development cost.

What adds to the problem in Winsberg's analysis is the fact that the technology — the set of existing climate models — has to be used at present to make decisions of enormous consequence. If the community of astrophysicists came to the conclusion that their currently existing models of the beginning of the universe are unsatisfactory — or even just that there might be a different model family to be investigated — they may spend a couple of decades developing those alternative models and not much will be lost in comparison. If the new model family in fact turns out to be inferior, its developers can more or less simply return to the old ones and further work on them from there. If the community of climate scientists did the same with their models, it may turn out that they wasted valuable time they could have spent improving their already rather advanced models further. This is a particular type of uncertainty combined with time pressure, which in this case acts as a factor to entrench current models.

To summarize, entrenchment takes on quite different forms for quite different reasons. What they have in common is that they impact decision-making on technology choices conditional on prior decisions. This is exactly what generates a problem for the iterative development of artificial moral agents suggested by the pragmatist stance, to which we can turn to complete the argument.

4 Application

4.1 A Pragmatist Dilemma of Entrenchment

With the concepts of technological entrenchment and a pragmatist machine ethics methodology, the dilemma can be assembled: To justify the deployment of a class of machines, they are required to be part of an iterative process of development, deployment and revision; in particular, this implies that they need to be put into productive use for an effective evaluation. According to the pragmatist stance, justification for the present use of the technology has to refer, among other things, to the future possibility of revision.

But, at least for many technologies, putting the machines into practical use entrenches them in the various ways described. Therefore, the very procedure argued for as a justificatory device has an inherent tension: Serious justification demands a test under real circumstances, but that test in an open environment may interrupt the iterative process via entrenchment.

While this is a rather simple argument on its face, there are a few subtleties and potential sources of misunderstanding. First, it should be understood that the issue

is primarily construed here as a *moral* objection grounded in a metaethical analysis of justification, as opposed, for instance, one of economic efficiency. The urgency of equipping autonomous systems with moral capabilities is a moral one. Entrenchment itself is not generally a moral problem — and if it is, then in an indirect manner — whereas in this circumstance, there is a direct moral problem, as it is constituted by AMAs' behaviour being recognized as morally inadequate. The reasons it is difficult to *overcome* entrenchment are still only indirectly moral, for instance by wasting resources that are direly needed elsewhere or imposing undue burdens on workers interacting with the machines.

Second, one might question the pragmatist justification procedure and instead demand a process that ensures in advance that no inadequate AMAs become deployed and therefore possibly entrenched in the first place. The main objections to this suggestion have already been laid out, in that it does not acknowledge the reality of moral disagreements and human fallibility in moral judgment.

Third, the argument is not based on a particular theory of normative ethics. It is based in a pragmatist account of justification the satisfaction of which is threatened by entrenchment. This argumentative strategy is thus quite different, for instance, than the argument for banning autonomous lethal robots due to their violation of principles of international humanitarian law such as discrimination between combatants and non-combatants (cf. Sharkey (2012)). The implications of the arguments may ultimately coincide or diverge, but the logic of the argument and what it assumes is quite different.

Finally, rather similar to the proposed measures against Collingridge's dilemma, one might suggest to monitor the technology and build it in such a way that it is robust in the face of problems and relatively easy to change if necessary (Liebert & Schmidt, 2010). I will address the potential of such measures in my concluding remarks; the main response for why this does not resolve the basic dilemma — even though such measure can have desirable mitigating effects — is that it does not take entrenchment serious enough.

4.2 Illustration: Autonomous Vehicles

The introduction of self-driving cars serves as an interesting example. One may agree with Himmelreich (2018) that the issue of self-driving vehicles and their decision procedures should be construed as a political rather than a moral one, but we can certainly imagine, as a thought experiment, the introduction of autonomous vehicles for individual transportation equipped with an explicit moral reasoning module. We need not make too many assumptions about the technology; it could be a bottom-up system generalizing from previous judgments.[21] or a top-down implementation of rules in the form of constraints on admissible plans[22, 23]

[21] Similar, for example, to the approach Anderson et al. (2006) employ to train their advisory system.

[22] Along the lines of the implementation of the laws of war suggested by Arkin (2008).

[23] Furthermore, one could use a combination of both, in what Allen et al. (2005) call a hybrid approach. The method applied in the design context is secondary for our purposes.

A first reason for entrenchment problems is that it is unclear whether the full benefits of self-driving vehicles could be achieved without replacing non-autonomous vehicles altogether. If that was the case, we could assume there will be substantial pressure to replace non-autonomous vehicles entirely and in as short a span of time as possible. As a consequence, reverting entirely to the previous technology would require the replacement of a vast fleet of cars.

But even if the technology of self-driving vehicles does not have to be rolled back in its entirety, hardware replacements on a large scale may be necessitated by the iterative process. For instance, if a certain type of sensor is insufficient to provide the required information to the moral module, it had to be replaced across the fleet.

We can also expect infrastructure entrenchment of various kinds, in manufacturing, repair, but also possibly in the road infrastructure. In particular with respect to the last, problems similar to the ones arising from the vehicle hardware may come up. For instance, roads may need to be equipped with new guidance systems.

Furthermore, while it might be, compared to the hardware, a lesser issue, the moral control software would have to be updated potentially rather frequently. While the updating itself may appear not as a huge issue, the interactions with other software components and the hardware will be difficult to control. Besides these technical concerns, the frequent need for updating may come with serious problems on the human side.

Humans adapt to the characteristics of traffic, but not necessarily very fast. For instance, it has been noted that the lack of noise in electrical vehicles has them threaten other traffic participants, who are used to audible cues. Indeed, manufacturers sometimes equip their products with additional noise output to avoid those problems. There is no reason to expect that humans will be faster at adapting to autonomous vehicles, and of course potentially frequent updates heighten this problem dramatically: Every time a major update is rolled out, the vehicles' behaviour instantaneously changes across all the vehicles running the relevant moral module.

This last point is the most stark one in front of our pragmatist background: If humans cannot adapt fast enough — in terms of social norms, individual skills and habits — the whole iterative process is threatened. It presupposes that human agents are able to process their experience and improve on their prior judgments on that basis; but if they fail to effectively process the new experiences, the result of that interaction may not be a sophistication of judgment, but rather an irrational, impulsive response. If that were the case, the process of justification would be invalidated.

An example for such dysfunctional interactions is found in the discussion on strategic responses in the literature. In particular, the case of a vehicle differentiating between helmet-wearing bicyclists and those without in its assessment of whom to expose to the risk of an accident (Goodall, 2014). In a vacuum, it makes sense to rather hit someone with a helmet, as they are more likely to survive. But a plausible strategic response to that is for bicyclists not to wear helmets to move into the protected category, reducing overall safety in traffic.

To be clear, this is all thought experiment right now, but hopefully, it is sufficiently vivid and plausible to illustrate the point. The moral capabilities of a machine may become part of an entrenched technology, and thus undermine their own justification as part of a progression of moral evaluations.

5 Concluding Remarks

In this paper, it has been argued that there is, due to the phenomenon of technological entrenchment, a strong internal tension in the pragmatist process to justify a class of AMA. Fallibilism and the consequent demand for revisability clash with the need to experience the new technology in productive use and real interaction to form the next level of moral judgment to sanction the next generation of AMA.

As noted above, general mitigation strategies against entrenchment may help reduce the tension somewhat; monitoring is required anyway, and the iterative setup of the pragmatist MTT approach also suggests an implementation that is open to future changes. As I noted above, engineers have the capacity to design technology in more reversible ways, and the argument I presented adds to the existing reason to value their efforts and ingenuity in that regard.

However, the reality of technology, in particular as it is physically implemented and interacted with by humans, imposes limitations on any such strategy to solve the problem simply through technological advancement.

To be clear, the argument relies on two propositions that may turn out to be wrong: A better framework to reconstruct the metaethics of an MTT and justify its application could be found, opening up argumentative space such that the argument could be undermined. Similarly, though less plausible, a compelling process that does not resemble the iterative-experimental procedure of the MTT could be found, again defeating one of the premises of the central argument. I presented what I consider strong reasons for the acceptance of these premises, but ultimately, they are as fallible as any product of human inquiry.

Finally, one may suggest an entirely different route to avoid the dilemma: One may reject the deployment of explicit moral agents. This move would obviate the need for a justification of the moral judgment algorithm of any artificial agent. But machine ethicists generally argue that for agents of a certain level of autonomy, operating in an open environment, it would be morally inadmissible to deploy an artificial agent incapable of any moral reasoning and judgment. If this argument is valid, a rejection of explicit AMA[24] implies a rejection of autonomously operating artificial agents.

It is instructive to compare this argument with those commonly provided by proponents of a ban on autonomous lethal robots, summarized eloquently by Wallach (2017). His concern and that of many like-minded experts regards the compliance

[24] For a discussion of various levels of moral agency in artificial systems, see Moor (2006). Note, that my argument is not about full moral agency, but, as far as this distinction can be held up, merely explicit moral agent.

with principles of normative ethics, such as international humanitarian law or protections of human dignity. If those principles are not assumed, the argument falls through. My argument, on the other hand, is not built upon a substantive view of normative ethics, but on the structure of justification and tensions therein. This difference goes back to the objection to verificationist approaches above, namely that no agreement on these principles exists.[25]

I do not wish to make the call on which resolution to the dilemma is ultimately to be preferred here, but the discovery of a serious dilemma facing what I argue to be the most promising approach to justify a particular AMA's deployment certainly lends some support to a ban or moratorium.

Author Contribution Not applicable.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability Not applicable.

Materials and/or Code Availability Not applicable.

Declarations

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Competing Interests The author declares no competing interests.

References

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology, 7*(3), 149–155.

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence, 12*(3), 251–261.

[25] Wallach is also involved, as he reports in the paper, in an effort to bring about the ban he is advocating. From a pragmatic point of view (rather than a pragmatist one), going through established regulatory and legal processes makes sense. I do not advocate here for any particular method of preventing the deployment of autonomous artificial agents, for military or other purposes; I simply argue that this option is a solution to the dilemma that could be taken if mitigation strategies are considered insufficient.

Anderson, M., Anderson, S. L., & Armen, C. (2006). MedEthEx: A prototype medical ethics advisor. In *Proceedings of the National Conference on Artificial Intelligence* (vol. 21, p. 1759). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Arkin, R. C. (2008). Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction* (pp. 121–128).

Arnold, T., & Scheutz, M. (2016). Against the moral Turing test: Accountable design and the moral reasoning of autonomous systems. *Ethics and Information Technology, 18*(2), 103–115.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature, 563*(7729), 59–64.

Collingridge, D. (1980). *The social control of technology*. Martin, New York: St.

Dewey, J. (1929). *The quest for certainty*. New York: Capricorn Publishing.

Dewey, J. (1939). Theory of valuation. *International Encyclopedia of Unified Science*.

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Goodall, N. J. (2014). Machine ethics and automated vehicles. In *Road Vehicle Automation* (pp. 93–102). Springer.

Himmelreich, J. (2018). Never mind the trolley: The ethics of autonomous vehicles in mundane situations. *Ethical Theory and Moral Practice, 21*(3), 669–684.

James, W. (1975). *Pragmatism* (vol. 1). Harvard University Press.

Liebert, W., & Schmidt, J. C. (2010). Collingridge's dilemma and technoscience. *Poiesis & Praxis, 7*(1–2), 55–71.

Moor, J. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems, 21*.

Proulx, P. L. D. (2016). Early forms of metaethical constructivism in John Dewey's pragmatism. *Journal for the History of Analytical Philosophy, 4*(9).

Radder, H. (2009). Why technologies are inherently normative. In *Philosophy of Technology and Engineering Sciences* (pp. 887–921). Elsevier.

Rorty, R. (2000). *Der Spiegel der Natur: Eine Kritik der Philosophie*. Suhrkamp.

Sharkey, N. E. (2012). The evitability of autonomous robot warfare. *International Review of the Red Cross, 94*(886), 787–799.

Sepielli, A. (2017). Pragmatism and metaethics.

Turing, A. (1950). Computing machinery and intelligence-AM Turing. *Mind, 59*(236), 433.

Voiklis, J., Kim, B., Cusimano, C., & Malle, B. F. (2016). Moral judgments of human vs. robot agents. In *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)* (pp. 775–780). IEEE.

Wallach, W., & Allen, C. (2008). Moral machines: Teaching robots right from wrong. Oxford University Press

Wallach, W. (2017). Toward a ban on lethal autonomous weapons: Surmounting the obstacles. *Communications of the ACM, 60*(5), 28–34.

Winsberg, E. (2010). *Science in the age of computer simulation*. University of Chicago Press.