A criticism of AI ethics guidelines

This paper investigates the current wave of Artificial Intelligence Ethics GUidelines (AIGUs). The goal is not to provide a broad survey of the details of such efforts; instead, the reasons for the proliferation of such guidelines is investigated. Two main research questions are pursued. First, what is the justification for the proliferation of AIGUs, and what are the reasonable goals and limitations of such projects? Second, what are the specific concerns of AI that are so unique that general technology regulation cannot cover them? The paper reveals that the development of AI guidelines is part of a decades-long trend of an ever-increasing express need for stronger social control of technology, and that many of the concerns of the AIGUs are not specific to the technology itself, but are rather about *transparency* and *human oversight*. Nevertheless, the positive potential of the situation is that the intense worldwide focus on AIGUs will yield such profound guidelines that the regulation of other technologies may want to follow suite.

Keywords: *Artificial Intelligence Ethics; Applied Ethics; Ethics Guidelines; Social Control of Technology*

Author Information

Mihály Héder, Budapest University of Technology and Economics; SZTAKI Institute for Computer Science and Control
https://orcid.org/0000-0002-9979-9101

How to cite this article:
Mihály, Héder. "A criticism of AI ethics guidelines."
Információs Társadalom XX, no. 4 (2020): 57–73.
https://dx.doi.org/10.22503/inftars.XX.2020.4.5

Introduction

As the Artificial Intelligence (AI) industry has gain increasing prominence and achieved mainstream breakthroughs – yet again, after periods of progress interrupted by AI "Winters" (Crevier 1993; Hendler 2008) – there has been a proliferation in the number of guidelines, codes of ethics and manifestos created concerning how to address the moral concerns arising from the development of AI. This paper provides a critical analysis of the approaches and effectiveness of Artificial Intelligence Ethics GUidelines in general (AIGUs from this point on), from the perspectives of the philosophy of technology and applied ethics.

The first point of investigation of this paper is the existential question of AIGUs themselves. The need for and benefit of creating AIGUs may seem to be self-evident at first glance, but it should not be beyond questioning. A refined version of this question is: *What kind of AIGUs* are likely to reach the goals they set out to achieve, and in particular, what are the promising methodologies? As will become evident, the author of this paper does not believe that the efforts to create AIGUs aren't worth pursuing. Yet, their success is greatly dependent on the hidden premises and assumptions behind them, which make these assumptions valid subjects of investigation.

The second angle of investigation is about the specificity of AIGUs. In simple terms, the questions are: What are the elements of these guidelines that are not really about AI but are relevant for any novel technology and it just so happens that they are raised in the context of AI, and what are the considerations that truly only arise in the context of AI? The importance of this line of investigation is that it may advance the development of AIGUs by focusing them properly, instead of them trying to fight on too many fronts. In short, what are the unique differentiating factors of the field of AI that need to be accounted for?

Evidently, AIGUs fit into the field of applied ethics as the most recent domain-specific effort after work on bioethics, nano-ethics, information ethics and the like. Yet, AIGUs have a unique opportunity because of the unprecedented brightness of the spotlight that has been shining on this field of technology since at least the mid-2010s. Given earlier efforts, what can this field learn from other fields of professional ethics in which we have more experience now?

In order to narrow down the primary sources examined by this paper to a manageable amount yet still remain relevant, we define the subject of the investigation as documents that are written with a prescriptive intent for practitioners and decision-makers involved in AI development projects, focusing specifically on the moral dimensions of their work; hence the name AI ethics guidelines or AIGUs. Throughout this paper, the terms "regulation" and "guideline" are used somewhat interchangeably. Obviously, there are significant legal and therefore practical distinctions between them; however, from

the perspective of this article, the differences are orthogonal, since the focus is not on enforcement but rather on what is rational. Moreover, many of the AIGUs are written with the intent of being a basis for regulation, so their normative status may change in the future.

After scoping down the investigation this way, we still ended up with a quite a high number of documents to consider. Therefore, a second way of narrowing down was employed – that is, the potential outreach of such works. In this regard, the manifestos of large political entities (EU, US, China) and professional organizations (IEEE, OECD, big corporations) are prioritized. Finally, an unfortunate, but necessary limitation is that only AIGUs available in English are considered. The goal is not to provide a broad, quantitative survey like the one published in *Nature Machine Learning* of over 84 sources (Jobin et al. 2019) or another in *Minds and Machines* of over 22 AIGUs (Hagendorff 2020); instead, the motivation of this paper was to arrive at a qualitative understanding of what is a rational and consistent approach for creating an AIGU and what are the limitations of such an endeavour.

AI ethics guidelines in focus

In this article, we work with seven sources that are directly quoted and more closely investigated. These are OECD's (2019) *Recommendation of the Council of Artificial Intelligence*; IEEE's (2019) *Ethically Aligned Design* "Vision", EU's *Ethics Guidelines for Trustworthy AI* (AI HLEG 2019), *Beijing AI Principles* (2019), *Artificial Intelligence at Google* (2018) manifesto, Microsoft's (2019) *AI principles* and the *Report on the Future of Artificial Intelligence* (Holdren et al. 2016). The general principles behind selecting these sources are detailed in the introductions section: i.e. they need to be prescriptive, aimed at practitioners and decision-makers, deal with moral questions and have a high potential impact. While the appraisal of their potential impact is ultimately an inexact science, there are good arguments for the inclusion of these seven documents.

The OECD guideline is included because of OECD's global scope and because it includes – at least notionally – the most diverse set of countries and political entities. Previous OECD guidelines in different fields, like the Frascati Manual, have become successful common denominators in their subject areas through with a narrow-enough scope and broad global acceptance.

The IEEE guideline is arguably the most comprehensive in terms of topics. Hagendorff (2020) reports that it covers most (18) of the common AIGU topics he identified, while the next best AIGU in this metric covers only 14. Another significance of this guide is that it is explicitly stated to be an input for the upcoming P7000 series of IEEE standards on the ethics of AI. Since in the field of the ICT industry, the IEEE counts as one of, if not the most crucial source of standards and recommendations, practitioners in AI are bound to anticipate

and incorporate the P7000 series, like the P7001, which is promised to deliver a measurable, certifiable standard on the transparency of autonomous systems.

The EU Commission's effort through its AI high-level expert group is important because of the breadth of topics covered, the international (but intra-EU) collaboration manifested in it, and also because it is expected to serve as input to an upcoming regulatory framework (European Commission 2020), or simply put, legislation on AI. The reason for including a report by the US Government (Holdren et al. 2016) is similar. While there is less clarity about whether this Obama-era report will ever serve as a direct basis for legislation, the document is quite comprehensive.

The Beijing AI Principles, while nowhere near as comprehensive as the previous three AIGUs, can be seen as the position of the Chinese state on the issue of AI ethics, and hence warrants our attention.

Finally, Microsoft (2019) and Google (2018) AIGUs are included because they represent the position of two powerful industrial actors. That is not to say that these companies will turn out to be the most important players in the field, but at least these two have published guidelines.

All of these and many more AIGUs have been analyzed quite extensively already. As already mentioned, we include reviews in *Minds and Machines* in which Hagendorff (2020) studied 22 AIGUs, and in *Nature*, in which Jobin et al. (2019) investigated 84 AIGU sources.

These reviews reveal a remarkable similarity in the key concerns these sources identify. While the terminology differs, we can still identify some key ideas. Specifically, in the reviewed AIGUs, the leading concerns are *transparency* (sometimes coupled with explainability); *justice and fairness*; *responsibility* and *accountability*; *privacy*; a tendency to *promote good* (beneficence or facilitation of well-being); and some provisions to maintain *human autonomy*, and related to that (and to accountability) *human oversight*.

As Hagendorff (2020a; 2020b) establishes, there are plenty of omissions, too. Not only is it that the AIGUs may be lacking in scope, but it is also unclear how much difference they will make and what the chances of compliance with them are. In this paper, however, we start from a step further back: while the question of compliance is an important one, our working premise in this paper concerns the case of users of an AIGU who actively want to do the right thing *and* are ready to subsume their decisions as needed *and* to dedicate resources as required to this end. In other words, we presuppose, albeit with an exaggerated level of optimism, the best of intents and attitudes, because even with this assumption, the compilation of an AIGU is a challenging regulation and philosophical problem.

As we will see later, it is far from self-evident that all of the concerns above are novel ones, specifically brought forward by AI. But before getting to the concerns they cover, we investigate the motivation in general for creating AIGUs and then the reasoning behind our seven sources in particular.

Why make guidelines?

At first glance, the existential question of why make AIGUs is not dissimilar from the justification debates about technology regulation in general.

It is widely believed that one of the first regulated technologies – in the modern sense, with exact measures and gauges – was the steam engine boiler (Green 1953). This regulation, devised by a US engineering association in 1884 was a significant milestone as it was unprecedentedly enacted in legal code, including all of the details in 1907. We may say that with this event, the social control of technology was attempted at an entirely new level.

But why is the social control of technology (Collingridge 1981) necessary? In the case of the steam engine, the aim to avoid disasters was the main reason. Before regulation, explosions were common, claiming anywhere between a handful to over a thousand souls in a single accident. Technological disasters were seen as the result of chasing profits recklessly, hence cutting costs at the expense of safety, and of sheer carelessness. The boiler code was a success story, as it put a floor under the more dangerous forms of cost-cutting and enforced sound design and testing. The members of the American Society of Mechanical Engineers involved in the project could retire with the reasonably plausible belief that they had saved lives by their tiresome committee attendance and contribution. Unlike medical professionals, they could not point to the actual individuals they saved, but from the statistics, they knew there must be thousands of them. And all the while, the progress of technology was not seriously hindered, as some of the opponents of regulation feared prior to the enactment of the regulation.

Technology guidelines, whether mandated by law or recommended by the peers of the profession, have proliferated ever since. There seems to be a general public understanding of just how big a factor technology is in our everyday lives. While the sociologists debate what the exact nature of our technological dependence is (some more widely held positions are those of Marcuse 1964; Feenberg 2002; Gerrie 2008), three things are beyond debate: 1) the extent to which technology plays a role in our lives is enormous; 2) these effects are not necessarily positive or desirable and 3) technology is *not* beyond the possibility of control. These three beliefs constitute the preconditions of guideline-making: that it is both important and possible to control technology. At least since the end of the second world war, public attention hasn't been lacking either, as evidenced by civil activism and political action. Weapon tests, the chemicals used in agriculture, city buildings and other technomaterial concerns were among the first to be regulated, but media content also was not far behind (here, we distinguish from political censorship, a much older practice). From the environment to biotechnology, net neutrality, nanotechnology, etc., it is now the case that any emerging new technology is a natural subject of some form of soft law or actual legislation.

On the other side of the equation are the alleged costs of regulation. Regulation, even when written with the best intentions, may have unintended and unwanted side effects, which could possibly be worse than the negative events and states they arguably prevent in the first place. For instance, they may be used by state actors, companies and individuals as inexpensive means of merely *appearing* virtuous (Hagendorff 2020b); they may unnecessarily elevate the barriers of entry to a market; and they may be used as one of many tools in ideological or political clashes. Another criticism is that regulation may just be a simple means of "capturing" a market (Posner 1974), in which a group attempts to maximize its own profits by stifling competition, investing in lobbying instead of innovation.

Moreover, we can safely postulate that some AI applications bring serious, life-and-death improvements to their area – like the very probable proposition that autonomous driving will become orders of magnitude safer than human drivers or that AI lab assistants will identify maleficent tendencies in blood composition or on X-Rays with much better accuracy. If that assumption holds, delaying adoption by putting the burden of too much compliance and red tape on developers has a cost that is perhaps measurable in lives lost even.

Furthermore, the control of technology faces an inherent, unavoidable epistemic challenge, one formulation of which is the Collingridge (1981) dilemma. That is, if we attempt to come up with regulation in a timely manner – early in the development process – we will not have enough information and experience with the technology as to where to concentrate our efforts. In later stages, we will have learned what would have been the key decisions, but by that time, it is too late, since established, ubiquitous technologies are hard to change.

The level of abstraction

Note that the various early regulations mentioned above, starting with the boiler code, did not rely on the terminology of ethics and moral theory. There existed codes of ethics, but they operated with rather general terms, and they were quite short, like in the IEEE Code of Ethics, that can be seen as some sort of code of chivalry for engineers. As we move on to modern applied ethics (e.g. nano-; bio-; information ethics, ethics of reproduction technologies) and arrive at AIGUs, there is a perception that there is never enough time for the regulators to catch up with technology. This is why the emphasis has shifted from regulating the artefacts that are the outcomes of the development projects directly, to try and instruct the developers. While it was possible to regulate the steam boiler's maximal pressure in exact pounds per square inch values in the legal text, in more complex and quickly changing technologies, beginning with biotechnology, the strategy became to induce self-regulation by only providing more abstract guidelines to be interpreted to the problem at hand

and also to mandate the involvement of ethicists in the project. The merger between professional ethics and technology regulation is now complete, for instance, in the EU Regulation of AI, currently in preparation (Cohen 2020). This more complex approach, however, does not mean that the Collingridge dilemma or the capture problem is prevented.

However in the case of the AI, there is yet another new level of added complexity. The distinctive nature of AI as a technology is the unprecedented autonomy of the resulting artefacts, compounded with a high level of intelligence. It appears to some extent that AI is about the development of artificial persons (person stands here in a limited sense). And since ethical codes should guide a person's behaviour, there is now the possibility for interference and confusion. Are AIGUs meant to govern the behaviour of the human developers or the behaviour of the artificial agent? This confusion is real, and we find that AIGUs contain normative elements that we can either see as guidance for the developers or for the artificial agent. One such example is the recommendation (present in almost all AIGUs) to avoid bias – sometimes understood to refer to the conduct of the AI, at other times to the conduct of the developers and in yet another occasions it is both or is too hard to tell.

Moreover, the nature of the AI agent seems to pose a unique challenge with regards to precondition 3) for regulation (see above). That precondition stated the almost trivial fact that the possibility of controlling technology needs to exist for any regulation attempt to be rational. Industrial AI is in the business of letting an artificial agent do the tiresome intellectual work of controlling various situations. Autonomous driving is a prime example. The challenge is that we do not want to prescribe *exactly* what the AI should do, as that would defeat the purpose of having an AI – in this sense, the goal is to relegate control, thus working against precondition 3).

On the other hand, we *do* want to control the overall situation, in the sense of avoiding unwanted or unpleasant outcomes. And to make things harder, one cannot explicate or enumerate at the design-time all the unwanted outcomes an artificial agent might produce in run-time; in other words, because of the open-ended nature of AI, some unwanted outcomes we will only recognize after they have happened. And thus, an almost paradoxical tension is created between our needs for control. We need the AI to autonomously exercise control over the situation it is placed into while expecting ourselves to also remain in control in the sense that we need the AI to avoid unwanted consequences – which we cannot enumerate fully in advance as many of them are unforeseeable.

Recently it is often the case that in very complex R&D projects, like in the fields of bioethics or medical research, regulation has been delegated to the practitioner in the form of self-regulation, since a more generic regulation was not possible. Now, it seems there is yet another level of immediation introduced: in an AIGU, we ask AI practitioners to self-regulate with regards to what decisions they further delegate to the AI agents, and how. This will

require the AIGUs to be rather abstract – a property that we investigate later in this paper.

We may conclude that the possibility of complete and profound behaviour design is what makes the ethics of AI uniquely challenging and distinct from the regulation of other novel and powerful technologies, like GMOs or blockchains. Since human behaviour and ethics (what should a human do) are subjects of perennial debates, we may have to prepare for never-ending debates about machine ethics (what should a machine do) as well.

This thought highlights, however, the need for a distinction between the kinds of AI. Some applications of AI allow for less autonomous, less agent-like AIs, even without machine learning, like a traditional chess algorithm. Naturally, in this case, the above argument about relegation control is less relevant.

In the last 25 years, no human has been able to defeat AI in chess and in many other applications, yet there was no boom of AIGUs in the nineties. This suggests that there is something in the latest wave of AI applications that has provoked the proliferation of AIGU projects.

Perhaps the current tidal wave of AIGUs has to do with the development of a more autonomous kind of AI, with a ubiquitous presence in our everyday lives. Also, perhaps some psychological–perceptual barriers were broken with the proliferation of mobile AI platforms (autonomous car, vacuum robot) as to when we perceive a machine as an actor worthy of governing like a person instead of just smart software. The next section examines the motivations of particular AIGUs to shed some light on this question.

The motivation of AI ethics guidelines

> *"As the use and impact of autonomous and intelligent systems (A/IS) become pervasive, we need to establish societal and policy guidelines in order for such systems to remain human-centric, serving humanity's values and ethical principles."* (IEEE 2019: p2)

As we examine our seven sources, looking for sections that describe their motivation, it becomes clear that the authors of these guidelines are united in anticipating that AI will become a pervasive, transformative, unavoidable force on society in the very near future. This technology can bring enormous positive change but comes with profound risks at the same time, especially that it will not remain "human-centric", a term used in several of these AIGUs.

Therefore, yet again a new technology presents itself as a set of hard trade-offs, as described in the science and technology studies (STS) literature (i.e. Feenberg 2003). Technology may bring good, but it can also bring harm, and the differentiating factor may be proper regulation or guidelines. A perception is that if left unregulated, AI could disrupt our economies and erode our values:

> *"(...) AI also raises challenges for our societies and economies, notably regarding economic shifts and inequalities, competition, transitions in the labour market, and implications for democracy and human rights."* (OECD 2019)

This characterization is clearly present in six out of seven of our sources (The Microsoft Principles does not contain a rationale section), and is very much like the characterization of most of the past technology regulation debates, from child labour (Feenberg 2003) to the original debates around the boiler code (Ferguson 1987) or any other reconstruction of debates by STS of risky technologies (for several examples, see Johnson and Covello 2012). These studies invariably show that the concerns and perceived trade-offs at the time of the debates turned out not to reflect the problems and opportunities that later materialized. In other words, at the time of the back-and-forth debates around the risks and regulation of the new technologies, the participants of the debates simply failed to anticipate the future properly, and while it is true that some risks, benefits, and transformations were later realized, these did not resemble the fears and visions imagined beforehand. Of course, this divergence could also be a result of the implemented regulation itself, i.e. a risk that generated the need for caution was indeed prevented from being realized by the regulation. However, the studies above show that the divergence between the anticipated future and what came to be is usually too profound to simply ascribe it to the negating effect of the intervention on prediction. Rather, it appears that prediction of what new technology may bring is inherently hard, and regulators are mostly in the dark.

Yet, a sense of urgency prevails in all of our AIGUs. There is no exact reason given for this, but we can infer that the authors are worried that technologies may get locked in and may turn unmodifiable as they become ubiquitous. For instance, in the EU Guidelines:

> *"(...) While offering great opportunities, AI systems also give rise to certain risks that must be handled appropriately and proportionately. We now have an important window of opportunity to shape their development."* (AI HLEG 2019)

In other words, the situation appears to be set up like a Collingridge dilemma, in that arguably one of the most complex and unpredictable technology is being attempted to be controlled.

To sum up, there is a shared perception, one could even say a sense of hype, that the AI industry is just about to take off and hence some form of regulation or social control is immediately necessary. There is no consideration of the possibility that the development of AI may disappoint, especially in light of these elevated expectations (Floridi 2020). This does not mean that the progress of AI will stop, but it could mean that the overarching, society-transforming change

that these AIGUs are tuned for is generations away. In contrast, more mundane practices, some even bordering on the criminal (Hagendorff 2020), do not get enough attention, while in reality, these could be more effectively regulated against.

The specificity of the concerns in AIGUs

> *"The principle of prevention of harm:*
> *AI systems should neither cause nor exacerbate harm or otherwise adversely affect human beings.(...)"* (AI HLEG 2019)

Another research question about AIGUs is their specificity. What we investigate here is what are the challenges unique to AI and therefore that require unique guidelines and what would be relevant to any technology? This investigation also serves as input to our first question about the justification of the existence of AIGUs. That is, should we come to a conclusion on the extreme end – that the recommendations in AIGUs are not AI-specific after all – the logical consequence should be that these recommendations should be called simply engineering ethics considerations, without any need for AI guidance in particular. And yet, we find that some AIGUs are very ambitious, and rather than attempting to build on existing guidance, they seek to cover all concerns.

We can see if this is really the case with a simple method we may call the "water boiler" test. Let's replace "AI" with "water boiler" in the guidelines and see if the sentence still makes sense and remains valid. If yes, we may conclude that the specific piece of guidance is technology-agnostic and not AI-specific.

> "~~*AI systems*~~ ***Water boilers*** *should neither cause nor exacerbate harm or otherwise adversely affect human beings.(...)"* (AI HLEG 2019);
> *"~~AI systems~~* ***Water boilers*** *should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk."* (OECD 2019)
> *"Creators and operators shall provide evidence of the effectiveness and fitness for purpose of ~~A/IS~~* ***water boilers****."* (IEEE 2019)

As we can see all of these fail the test by remaining just as relevant with water boilers as with AIs. This, of course, does not mean that they are wrong, and should not be followed. It just shows that AIGUs include generic engineering guidelines, and this raises the question of whether these could be referenced from prior work instead of being re-invented. The examples above are far from the most generic; those would be the elements of guidance that remind developers to adhere to the law; explain the risks to customers; emphasize operator training.

Yet, we may be charitable in our evaluation and say that the inclusion of these "principles" or "guidelines" that pass the water boiler test serve a purpose: in this way, the AIGU is a one-stop-shop of guidance. Our only complaint would be then that these generic recommendations appear to be rather narrow, not mentioning such mundane requirements as the proper documentation of source codes and so on.

There is, however, a set of recommendations that would not pass the water boiler test but would work with "information system". These typically have to do with data collection and privacy:

> *"We will incorporate our privacy principles in the development and use of our AI technologies information systems. We will give opportunity for notice and consent, encourage architectures with privacy safeguards, and provide appropriate transparency and control over the use of data."* (Google 2018)

Again, we don't see the AI-specificity in such claims; however, it is clear that they do no harm for the purpose of the AIGU as a whole, despite being already mandated by privacy protection legislation for several years in most jurisdictions, therefore being redundant.

There is a class of non-AI-specific recommendations, however, that raises more important questions than the issue of redundancy. These are the guidelines that advise on the acceptable overall motivation of AI projects, like:

> *"A/IS Water boiler creators shall adopt increased human well-being as a primary success criterion for development."* (IEEE 2019, principle 2) *"A/IS should prioritize human well-being as an outcome in all system designs, using the best available and widely accepted well-being metrics as their reference point."* (IEEE 2019, recommendation for principle 2)

The section containing the quote above goes on to explain that GDP or consumption levels are possibly not the right way to assess the state of society and recommends instead well-being metrics form the OECD to guide the development of AI. We can find a similar recommendation in (EU HLEG 2019) and, an explicit reference, although with less emphasis in (Holdren et al. 2016). There are similar thoughts in the Microsoft, Google and Beijing recommendations as well.

The reason this is interesting is that while the quote semantically works with a water boiler, there is no tendency of calling out water boiler manufacturers to do more for social well-being in particular. Of course, they still need to be compliant with environmental, safety, financial, and consumer protection, etc. regulations and those could very well serve well-being, but they don't need to engage with the concept on an explicit level, and they seem to be allowed to pursue profit as a primary motive, as long as they remain within the legal boundaries.

This shows a profound shift in society's expectations towards innovators, but is maybe in conflict with the incentives of innovation, which in almost all economic theories has to do with a profit motive and competition. For some thinkers, this shift may be a welcome one, signifying the long-needed next step of technological enlightenment (Ropohl 1998), in that technology will be finally made to face the expectations its power warrants, or some form of successful democratization (Feenberg 2003) of technology, or a serious attempt of social control. Although, it should not be taken for granted that the pursuit of well-being has such an exact framework that could be implemented on a company level. It is, though, out of the scope of this paper to evaluate this techno-political shift and the possible social and economic consequences of it.

However, the more intense involvement of broad society in technological governance brings some methodological challenges. That is, this approach cries for an empirical investigation of what the public wants, in quantitative terms. Currently, as far as the documents reveal the methods they used to compile them, it appears that predominantly theoretical work has been conducted so far, reinforced by the experience of professionals in the field – quite a number of them in some cases, but ultimately a small subculture in comparison to all the members of society affected by AI.

Positive exceptions in this case are the EU and IEEE regulations, in which case a debate was induced, and a request for comments and questions made. The other AIGUs seem to be declarations not calling for public approval. This means that the current generation of AIGUs were created with the armchair method and refined in debates and by invited inputs.

This does not mean that there is no empirical research that could facilitate AIGU creation. For instance, the famous Moral Machine Experiment (MME) (Awad et al. 2018) set out to empirically measure the moral preferences of different social groups and cultures with the stated intent of facilitating the debate around AI ethics and hence to provide input to AIGUs. Yet, in order for such data to be useful, it needs to be established that the measurements are relevant to the actual design decisions. This is yet to be done (Dewitt et al. 2019; Jaques 2020; Héder 2020). Even if they were, the separate ethical question is whether it is the right thing to implement the most popular expectations. Is it not the case that some protection of minority opinions and value systems against majority expectations would result in fairer and more liveable societies?

Finally, one practical concern is worth pointing out: in a possible future where well-being is made to be the first priority, combined with some level of enforcement potential (soft or hard law), developers will be incentivized to define their work as *not* AI; and here, the rather numerous and often inexact definitions of AI provide the room for interpretation to do just that. This incentive will be there as long as AI projects need to live up to stricter requirements than other technological ventures.

Genuinely AI-specific concerns in AIGUs

The truly AI-specific concerns in AIGUs seem to be in correspondence with the unique features of AI: autonomous decision-making, learning capability and the high level of potential opacity.

A genuinely AI-specific requirement, as recommended by several AIGUs, is human oversight:

"Human oversight. Human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects. Oversight may be achieved through governance mechanisms such as a human-in-the-loop (...), human-on-the-loop (...), or human-in-command (...) approach." (EU AI HLEG 2019)

While it is true that the possibility of human oversight should be maintained in any technological solutions, even with water boilers, the elevated level of autonomy that is quite specific about AI warrants special attention. This is why discovering, defining and differentiating the various methods and levels of maintaining human oversight ("in-the-loop", "on-the-loop", etc.) seems to be a proper concern of AIGUs, and cannot be imported from the guidelines of other engineering fields.

Another AI-specific example is the requirement that *"the basis of a particular A/IS decision should always be discoverable."* (IEEE 2019), provided that we don't trivialize our decision definition to include the "decision" of the thermostat, etc. , this indeed does not make sense in any other context than AI. This requirement, that in other AIGUs is often called "explainability", is in connection with the high complexity of AI and machine learning as the method for tuning the system. Opacity is an issue not only in AI: any sufficiently complex system, even a fully mechanical one quite distant from AI, may become highly opaque to the users and even for the operators. Yet, while in those cases opacity can be ascribed to poor documentation and a gradual degradation of knowledge of the artefact over time, AI seems to take it to a whole new level by generating models with genetic algorithms and reinforcement learning that are opaque from the very beginning.

We can conclude that the issues of *transparency* and explainability are truly specific to the field of AI. Consequently, if we are to establish guidelines for these concerns, we will find little related work in other fields. However, this still does not mean that the pursuit of AI transparency is beyond criticism.

It seems that there are serious theoretical (Grünke 2019) and practical (Héder 2020) limitations to achieving a high level of transparency, and therefore if the requirement of transparency is not defined in an appropriate level of abstraction, it could become a serious hindrance to AI development. Moreover, if transparency is not combined with some measure of intelligibility, AI developers may get away with pretend transparency – in this context, this would mean the release of an unmanageable amount of code and configura-

tion, where the developers could claim that they released "everything", yet it would be impossible to for an outsider to gain any useful insight from this.

Finally, it is claimed with quite convincing arguments, that compared to human decision-making, a strong transparency requirement towards AI is a double standard (Zirelli et al. 2019). That is, the level of transparency we enjoy about our fellow humans' decision-making processes is very low. Obviously, we cannot investigate the brain in any useful level while the decision is made, and what is more, the decision-maker cannot give us a full account even with the best intentions, as the relevant processes are partially opaque, even for the person making the decision itself.

Still, we accept the opacity of humans – we have no choice. While it is plain that a strong transparency requirement against AI is a double standard when compared with humans, this is arguably a positive development. First of all, it fits into the narrative explained above, about society's ever-increasing expectations towards technology. Also, just because for practical reasons human decision-making is very opaque, we may contend that this would be very beneficial in certain situations – only if we could achieve it. And since we have much fewer limitations when it comes to AI, we may mandate it to a larger extent.

Discussion

This paper investigated the current wave of Artificial Intelligence Ethics Guidelines (AIGUs). Two research questions were pursued. First, what is the justification for the proliferation of AIGUs, and what are the reasonable goals and limitations of such projects? Second, what are the specific concerns of AI that are so unique that general technology regulation cannot cover them?

Our first question was answered by putting AIGUs into historical context with other kinds of regulation and guidelines. The result of this revealed that there is an ever-increasing trend of elevated expectations of social control, from since at least the mid-20th century. AIGUs are expressions of a yet higher level of expectations, and some elements of AIGUs are not specific to AI, rather they seem too new and to be general requirements from society's part towards any technology; and this development just coincides with the current wave of AI technologies and their successes.

While this elevated level of concern expressed by society may justify a rather comprehensive set of regulations, the unwanted side-effects should also be considered: the cost of regulation in non-realized benefits of a timely application of AI, and the potential market capture that a misconstructed regulation may enable. The AIGUs investigated contain no obvious reference to any of these issues.

Our second question was: What are the ethical concerns truly specific to AI? The answer to this question can be derived by considering the unique features of AI systems, that no other technology exhibit. We found that the most genu-

inely AI-specific issues are transparency and *human oversight* (there are some other different names to these concepts that are also included regardless). Other concerns and requirements have been shown to be quite non-specific, raising the question of whether they should be really considered in general engineering ethics.

The issue of transparency is a quite complex one. One problem with this demand against AI applications is that one may offer remote/pretend transparency, that is, release full source code and data and still remain opaque as the released information is not intelligible. However, and an intelligibility test raises its own problems, i.e. how do we arrive at a characterization of the person that need to understand a system and then how do we measure it? Another issue is that this can be seen as a double standard: we have no transparency requirement against humans, furthermore, we are arguably opaque to our own selves.

The core of the issue around human oversight is that AI is a transfer of control from humans to machines. A paradoxical tension is created by this situation, in which we wish to delegate as much control as possible, since control is hard intellectual work, and yet still wish to keep some control over AI in the sense that we want to avoid negative outcomes and maintain our capacity for intervention. This means that we want to both delegate oversight in one sense and retain it in another, leaving the AI practitioner in a predicament of how to make this decision.

Where all these leads is a high level of abstraction: AIGUs cannot prescribe the exact details of wanted and unwanted AI systems. Therefore, they are forced to operate on the level of principles and general recommendations. Therefore, guided self-regulation is expected from the developers. To make matters even more complicated, one area of AI development is to establish what decisions to delegate to the autonomous systems and how in order to get to the best results. Therefore, it seems that we would require the *artefacts* produced by AI developers (the autonomous systems) to be "ethically aligned", and to have the developers figure out how, while mandating that they themselves are "ethically aligned". This double indirection means that all the AIGUs can do is guide the developers on "guiding" the AI systems. Perhaps this remoteness of the actual artificial agent from the committee that is formulating the guidelines is why the AIGUs have a tendency of being abstract to the point of ineffectiveness.

Possibly, the current wave of development of AI Ethical Guidelines (especially the more comprehensive ones) represent the most ambitious and demanding regulatory efforts towards technology this far in the history of humanity. The reason for this appears to be an ever-increasing need for well-being and other societal goals, paired with the willingness for social control off technology. A recognition of the complexity and potential of AI and the ambition to regulate it nevertheless, perhaps shows that we may really call this shift in technology reception as "technological enlightenment".

References

AI HLEG. “Ethics Guidelines for Trustworthy AI.” Accessed December 30, 2019. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Awad, E., Dsouza, S., Kim, R. et al. “The Moral Machine experiment.” *Nature* 563 (2018), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Beijing Academy of Artificial Intelligence. “Beijing AI principles.” Accessed December 30, 2019. https://www.baai.ac.cn/blog/beijjng-ai-principles

Crevier, D. *AI: The Tumultuous Search for Artificial Intelligence*, New York, NY: BasicBooks, 1993. ISBN 0-465-02997-3

Cohen, I. G., Evgeniou, T., Gerke, S., & Minssen, T. “The European artificial intelligence strategy: implications and challenges for digital health.” *The Lancet Digital Health* 2, no 7 (2020): 376–379.

Collingridge, D. *The Social Control of Technology.* Open University Press, 1981.

Dewitt B., Fischhoff B. & Sahlin N. “<Moral machine> experiment is no basis for policymaking.” Nature 567 (2019): 59–64. https://doi.org/10.1038/d41586-019-00766-x

European Commission. “*On Artificial Intelligence – A European approach to excellence and trust. COM 65.*” Accessed November 1, 2020. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

Floridi, L. “AI and Its New Winter: from Myths to Realities.” Philosophy and Technology 33 (2020): 1–3. https://doi.org/10.1007/s13347-020-00396-6

Feenberg, A. *Transforming Technology: A Critical Theory Revisited*. Oxford University Press: 2002.

Feenberg, A. “Democratic rationalization: Technology, power, and freedom.” In *Philosophy of technology*, edited by R. Sharffand V. Dusek, 652–665. Blackwell publishing, 2003.

Ferguson, E. S. “Risk and the American engineering profession: the ASME Boiler Code and American industrial safety standards.” In *The Social and Cultural Construction of Risk*, edited by B.B. Johnson and V.T Covello, 301–316. Dodrecht: Springer, 1987.

Gerrie, J. “Three species of technological dependency.” *Techné: Research in Philosophy and Technology* 12, no, 3 (2008): 184–194.

Google. “Artificial intelligence at Google: Our principles”. Retrieved December 30, 2018. https://ai.google/principles/.

Google. “Perspectives on issues in AI governance”. Retrieved February 11, 2019. https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf.

Green, A. M. *History of the ASME boiler code.* New York: American Society of Mechanical Engineers, 1953.

Grünke, P. “Chess, Artificial Intelligence, and Epistemic Opacity.” *Információs Társadalom* 19, no. 4 (2019): 1–7, http://dx.doi.org/10.22503/inftars.XIX.2019.4.1

Hagendorff, T. “The ethics of AI ethics: An evaluation of guidelines.” *Minds and Machines* 30 (2020a): 99–120.

Hagendorff, T. “Forbidden knowledge in machine learning reflections on the limits of research and publication.” *AI & Society* (2020b): 1–15. https://doi.org/10.1007/s00146-020-01045-4

Héder, M. “The epistemic opacity of autonomous systems and the ethical consequences.” AI & Society (2020): 1–9. https://doi.org/10.1007/s00146-020-01024-9

Hendler, J. "Avoiding Another AI Winter." *IEEE Intelligent Systems* 23, no. 2 (2008): 2–4. https://doi.org/10.1109/MIS.2008.20

Holdren, J. P., Bruce, A., Felten, E., Lyons, T. and Garris, M. *Preparing for the future of artificial intelligence*. Washington, DC: Springer, 2016.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. "*Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition.*" Accessed December 30, 2019. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/ autonomous-systems.html

Jaques, A. E. "Why the moral machine is a monster. Self-archived manuscript at University of Miami School of Law." Accessed August 24, 2020. https://robots.law.miami.edu/2019/wp-content/uploads/2019/03/MoralMachineMonster.pdf

Jobin, A., Ienca, M. and Vayena, E. "The global landscape of AI ethics guidelines." Nature Machine Intelligence 1 (2019): 389–399. https://doi.org/10.1038/s42256-019-0088-2

Johnson, B. B. and Covello, V. T. (Eds.). *The social and cultural construction of risk: Essays on risk selection and perception (Vol. 3).* Springer Science & Business Media: 2012.

Marcuse, H. "The New Forms of Control." In Marcuse H. *One-Dimensional Man,* 1-18. Boston: Beacon 1964.

Microsoft Corporation. "Microsoft AI principles." Accessed December 1, 2019. https://www.microsoft.com/en-us/ai/our-approach-to-ai

OECD. "Recommendation of the Council on Artificial Intelligence". Accessed August 20, 2019. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Posner, R. A. *Theories of economic regulation (No. w0041).* National Bureau of Economic Research, 1974.

Ropohl, G. "Technological Enlightenment as a Continuation of Modern Thinking". *Research in Philosophy and Technology* 17 (1998): 239–248.

Zerilli, J., Knott, A., Maclaurin, J. et al. "Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?." Philosophy and Technology 32 (2019): 661–683. https://doi.org/10.1007/s13347-018-0330-6