



The social dilemma in artificial intelligence development and why we have to solve it

Inga Strümke^{1,2} · Marija Slavkovic³ · Vince Istvan Madai^{4,5,6}

Received: 21 July 2021 / Accepted: 12 November 2021
© The Author(s) 2021

Abstract

While the demand for ethical artificial intelligence (AI) systems increases, the number of unethical uses of AI accelerates, even though there is no shortage of ethical guidelines. We argue that a possible underlying cause for this is that AI developers face a social dilemma in AI development ethics, preventing the widespread adaptation of ethical best practices. We define the social dilemma for AI development and describe why the current crisis in AI development ethics cannot be solved without relieving AI developers of their social dilemma. We argue that AI development must be professionalised to overcome the social dilemma, and discuss how medicine can be used as a template in this process.

Keywords Artificial intelligence · Social dilemma · Machine learning · Professional ethics

1 Introduction

A professional should not have to choose between their job and doing the right thing. Still, AI developers can be and are put in such a position. Take the example of a company that develops an AI tool to be used to guide hiring decisions: after the product has reached a certain stage, developers may identify ethical challenges, e.g. recognising that the tool is discriminatory against minorities. Avoiding this discrimination may require decreasing the performance of the product. What should the developers do to rectify the situation? They necessarily need to inform management about their concern, but their complaint can be met with indifference, and even a threat to replace them.¹

Situations such as these fall into the category of *social dilemmas*, and our goal in this paper is to highlight the impediment to ethical AI development due to the social

dilemma faced by AI developers. We argue that the current approaches to ethical practices in AI development fail to account for the existence of the challenge for developers to choose between doing the right thing and keeping their jobs.

A social dilemma exists when the best outcome for society would be achieved if everyone behaved in a certain way, but actually implementing this behaviour would lead to such drawbacks for an individual that they refrain from it. The problem we identify is that the current structures often put the burden to refuse unethical development on the shoulders of the developers who cannot possibly do it, due to their social dilemma. Furthermore, this challenge will become increasingly relevant and prevalent since AI is becoming one of the most impactful current technologies, with a huge demand for development [19, 68].

Advances in the field of AI have led to unprecedented progress in data analysis and pattern recognition, with subsequent advances in the industry. This progress is predominantly due to machine learning, which is a data-driven method. The utilized data are in the majority of cases historical, and can thus represent discriminatory practices and inequalities. Therefore, many machine learning models currently in use cement or even augment existing discriminatory practices and inequalities. Furthermore, AI technology does not have to be discriminatory for its development to be unethical. Mass surveillance, based on e.g. facial recognition, smart policing and safe city systems, are already used by several countries [29], news feed models used by social media create echo chambers and foster extremism [24], and autonomous weapon systems are in production [38].

¹ This example is a generalization of numerous experiences the authors have of being approached at relevant conferences by developers who perceive their work as unethical, e.g. as discriminatory against minorities. A common question within this context is whether they should risk losing their jobs for prioritising ethical considerations.

✉ Vince Istvan Madai
vince.madai@gmail.com

There has been rapid development in the field of AI ethics, and sub-fields like machine learning fairness, see e.g. [11, 35, 51, 69]. However, it is not clear that much progress is made in implementing ethical practices in AI development, nor that developers are being empowered to refuse engaging in unethical AI development. Reports such as The AI Index 2021 Annual Report [73] stress the lack of coordination in AI development ethics. Specifically, one of the nine highlights in this report states that “AI ethics lacks benchmarks and consensus”.

Major corporations, also referred to as “Big Tech”, are the ones developing the overwhelming majority of AI systems in use. These corporations have reacted to academic and public pressure by publishing guidelines and principles for AI development ethics. There has been what can be characterised as an inflation of such documents over the past years [37, 44, 64]. Although researchers and the society view AI development ethics as important [27], the proliferation of ethical guidelines and principles has been met with criticism from ethics researchers and human rights practitioners who, e.g., oppose the imprecise usage of ethics-related terms [30, 61]. The critics also point out that the aforementioned principles are non-binding in most cases and due to their vague and abstract nature also fail to be specific regarding their implementation. Finally, they do not give developers the power to refuse unethical AI development. The late firings of accomplished AI ethics researchers [39, 40, 46] for voicing topics inconvenient for the business model of the employer, demonstrate that top-down institutional guidelines are subject to executive decisions and can be overruled. While we acknowledge that we must be cautious when generalizing from single cases, we are not alone with our concern that ethical principles might be merely ethics washing [13, 54, 70], i.e., that corporations only give the impression of ethical practices in order to avoid regulation. Thus, the need for implementing ethical principles in AI development remains, and a crucial factor for this to succeed is removing the social dilemma for AI developers.

Social dilemmas exist in most areas where individuals, employers, and society are in a relational triangle around decisions that affect the society at large. AI is not an exception; there are many fields that encounter social dilemmas and some have successfully implemented mitigating measures. A very prominent example of this is medicine. In this paper, we argue that medicine’s strong focus on professionalization and the development of binding professional ethical codes is a powerful way to protect medical professionals from social dilemmas, and we discuss how structures like those in medicine can serve as a blueprint for AI development, thus leading to a lasting impact on ethical AI development.

Clearly, the issue of how to ensure ethical conduct of for profit companies cannot be reduced to resolving the issue

of the social dilemma for the employees. This is a complex question that relies on constructing legal frameworks to address “big tech” regulation and it is one of broad public interest [63]. In this paper, we focus only on the bottom-up contribution to the resolution of this complex problem, but recognise that the top-down legal regulation is a necessary component as well.

Before proceeding, we recognise that our analysis touches upon topics from other ethics sub-fields, namely business ethics, corporate ethics and research ethics. We do not adopt the viewpoint of any of these since we believe that our analysis can inform them and would be hampered by a too narrow focus.

Our aim is to raise awareness towards the existence of the social dilemma for AI developers, and by doing so outline the need for a systematic solution that removes that social dilemma. This process is a societal task, and will require interdisciplinary expertise from other fields outside of AI development.

The paper is structured as follows. In Sect. 2, we carefully define the social dilemma in AI development and highlight how it differs from known instances of social dilemmas. In Sect. 3, we elaborate on why professional codes of behaviour supplement legislation in tackling the serious social problem that is the regulation of technology impact. In Sect. 4, we take a lesson from the field of medicine on how social dilemmas faced by medical professionals are removed by establishing a professional code of conduct. In Sect. 5, we discuss issues with establishing ethical codes of conduct for AI developers, as we see them today, and in Sect. 6 we discuss related work. Finally, we outline limitations in Sect. 7 and our conclusions in Sect. 8.

2 The social dilemma in AI development

A social dilemma, also referred to as a ‘collective action problem’, is a decision-making problem faced when the interests of the collective conflict with the interests of the individual making a decision. It was established in the early analysis of the problems of public good cost by [57, 59], who stated that “rational self-interested individuals will not act to achieve their common or group interests”. Well known problems that can be considered instances of social dilemmas are the prisoner’s dilemma [50], the tragedy of the commons [49], the bystander-problem [26], fishing rights, et alia. The best known of these is perhaps the tragedy of the commons, which is a situation in which individuals with open access to a shared resource selfishly deplete the resource, thus acting against the common good and hurting their own individual interests as a result. All collective action problems concern situations in which individuals fail to behave according to the interests of the collective, although this

would ultimately benefit all individuals, or, as stated by [47]: “situations in which individual rationality leads to collective irrationality”. At the same time, all these examples are metaphors that stand as evidence for the difficulty of formulating an exact definition of social dilemmas [7].

In the context of AI, the social dilemma has been little discussed. The exception is in relation with autonomous vehicles [17]. Bonnefon et al [17] observe in their experiments that “people praise utilitarian, self-sacrificing AVs and welcome them on the road, without actually wanting to buy one for themselves.”, and state that this has “...the classic signature of a social dilemma, in which everyone has a temptation to free-ride instead of adopting the behavior that would lead to the best global outcome.” This is, in fact, the tragedy of the commons [41].

The social dilemma in AI development described in the introduction, however, does not fit the metaphor of the tragedy of the commons, or any of the other commonly used social dilemma metaphors. Consequently, we need to define the social dilemma in the context of AI development, and put forward the following definition: *a social dilemma exists when the best outcome for society would be achieved if everyone behaved in a certain way, but actually implementing this behaviour would lead to such drawbacks for individuals that they refrain from the behaviour.* In the social dilemma in AI development, we encounter three agents, each with their, possibly conflicting, interests: society, a business corporation, and an AI developer who is a member of society and an employee of the business corporation. The interest of society is ethical AI development; the interest of the business corporation is profit and surviving in the market; the interest of the developer is primarily maintaining their employment, but secondly ethical AI development, because developers are also a part of society. The developer is thus put in a situation where they have to weigh their interest as a member of society and their interests as an employee of the corporation. This is the social dilemma we want the AI developer **not** to face.

An analysis by PricewaterhouseCoopers [60] stated that AI has the potential to contribute 15.7 trillion dollars to the global economy by 2030. This puts business corporations in a competitive situation, especially regarding developing and deploying AI solutions fast. Fast development is potentially the opposite of what is needed for ethical development, which can require decreasing the development speed to implement necessary ethical analyses, or even deciding against deploying a system based on ethical considerations. This can create a direct conflict between the corporations’ motivation and the interest of society, which manifests in the work and considerations of the developers. These then find themselves in a situation where they might be replaced if they voiced concerns or refuse to contribute to the development.

The intention is not for AI developers to be the centre of decision making processes about what is ethical. There is a consensus [8, 23, 71] that the decision of what is ethical should be one taken as an agreement among all identified stakeholders in a society in which the AI system will be developed and deployed. However, what is frequently neglected is the specification of how that agreement on what is ethical should be reached. As Baum [12] elaborates, collectively deciding what is ethical is not a simple process. Since no mechanisms to reach a stakeholder agreement are put in place, developers are being put in a position to be the judge and jury on what is ethical, without having been trained at working with wider communities to achieve a collective understanding of the moral and societal impact of the system they are building. But even if they are trained for the task, their position in the company does not necessarily empower them to act upon it. This situation is reminiscent of the so-called principal-agent problem: The AI developers are agents of the principals in the form of their employers, and the social dilemma situation constitutes a conflict between the AI developers and their employers [5, 43, 62].

Expecting that AI developers will overcome this social dilemma without support is unrealistic. This stance is strengthened by the observation of other areas where social dilemmas are evident, e.g. climate change, environmental destruction, and reduction of meat consumption, where billions of people behave contrary to the agreed-upon common goal of sustainability, because of their social dilemmas.

AI development ethics, however, are much more complex than for example the ethics of meat consumption. The ethical challenges in AI are often both novel and complicated, with unforeseeable effects. While different approaches to ethics may not provide the same answer to the question “What is ethical development?”, the process of analysing the ethical aspects of a development process or system yields important information regarding the risks that can be mitigated by the developer. Yet, analysing a system and its potential impact from an ethical standpoint requires ethical training and a methodology. For the AI developer untrained in ethics and facing a social dilemma, it is unrealistic to perform this task, especially at scale [53].

We can also observe the potential for a social dilemma to occur on another level, this time for corporations: no single corporation or small group of corporations can take on the responsibility of solving AI development ethics, as this might put them at a disadvantage compared to other agents in the same market.² It is an interesting phenomenon that the social dilemma spirals upwards, in the sense that it

² We acknowledge that there might be cases where ethical development of a product can be considered a competitive business advantage. The existence of such cases does not preclude, however, that also cases exist where ethical development is a clear disadvantage.

can only be removed by solving it at the lowest level. If no corporation finds developers willing to engage in unethical development, they cannot end up in the corporation-level social dilemma. Furthermore, imposing corporation-level regulations for ethical conduct would likely lead to a search for loopholes, especially since there would always be gray zones, context-dependence and need for interpretation. From this perspective, solving the social dilemma for developers is also the approach that would lead to the most stable solution.

3 Professional codes versus legislation

Ethical perspectives in AI development are important since unethical development of AI can have a profound, negative impact both on individuals and society at large. Motivated by recent efforts to propose a regulatory framework for AI [28], one might be tempted to think that the challenge of ethical AI development could be solved solely by legislation. However, there are several reasons why legislation cannot fill this role: legislation develops at a much slower pace than current technology, implying that legislation is likely to arrive after harm has already been done, or even worse, after customary practice has been established. Furthermore, legislating against anything that could potentially be unethical or misused would disproportionately hinder progress, which is both undesirable and would in practice affect small businesses more than large ones, reinforcing the already problematic power imbalance between users and providers.

Note that we do not argue against legislative regulation. We are convinced that regulation is an important part of the overall approach towards ethical AI development. So are alternative approaches to AI governance such as human rights-centered design [52, 65, 72], AI for social good [31], algorithmic impact assessment [21], or Ubuntu [55]. We argue, however, against the notion that regulation and these alternative approaches suffice to create a stable solution. They have a blind spot and we thus face the challenging situation where—for this blind spot—we have to entrust corporations developing AI to take the ethical responsibility, despite not being motivated purely by the benefit of society. If the corporations do not shoulder this ethical responsibility, individual developers will be hindered in pursuing ethical development due to their social dilemma. We now describe a possible solution to this problem, recognising that the described phenomenon is not novel from a societal point of view.

Historically, societies have understood early that certain professions, while having the potential to be valuable for society, require stronger oversight than others due to their equally substantial potential for harm. While the necessity of a certain autonomy and freedom for professionals is acknowledged, it is important to simultaneously expect

professionals to work for the benefit of society. As [32] puts it: “Society’s granting of power and privilege to the professions is premised on their willingness and ability to contribute to social benefit and to conduct their affairs in a manner consistent with broader social values”. Camenisch [20] even argues that the autonomy of a profession is a privilege granted by society, which in turn leads to a moral duty of a profession to work for societal benefit. Professional codes that are not in line with societal good will be rejected by society [42].

Professional codes have been used to promote the agreed upon professional values in areas where legislative solutions are inadequate. Members of a profession are tied together in a “major normative reference group whose norms, values, and definitions of appropriate [professional] conduct serve as guides by which the individual practitioner organizes and performs his own work” [58]. Most importantly, in the context of this work, professional codes are a natural remedy against social dilemmas encountered in professional settings. The individual is relieved from the potential consequences of criticising conduct or refusing to perform behaviour in violation of their professional codes, and it would be highly unlikely that another member of the same profession would be willing to perform the same acts in violation of the professional code. Furthermore, the public would have insight into what is the standard ethical conduct for the entire profession.

Naturally, professional codes do not develop in a void. They draw from ethical theories, the expectations of society and from the self-image of the professionals. Consequently, professional codes are never set in stone but are constantly revised in light of technical advancement, development in societal norms and values, and regulatory restrictions. However, although they are dynamical, there is still at any given time a single version that is valid, protecting the individual professional from the social dilemma and maximizing the benefit for society.

We acknowledge that the development of professional codes is not an easy task. We thus argue that it is best to draw from a field that has succeeded at the task, as described in the next section.

4 Professional codes in medicine

As stated in Sect. 1, we suggest using medicine as a template for a professional code for AI development ethics. Although other fields have also developed professional codes, we argue based on societal impact that medicine is the most suitable example to follow. Medicine—primarily responsible for individual and public health—has a tremendous impact on society, at a level which few, if any, other professions share. AI has the potential of a similarly or even more substantial impact, depending on future development.

Medicine is an ancient profession, with the first written records dating back to Sumeria 2000 BC [14]. The Hippocratic oath, the first recorded document of medical professional codes, was introduced by the ancient Greeks, and its impact was so large that many laypeople, incorrectly, believe that it is still taken today [74]. The British and American medical associations drafted their first codes for ethical conduct in the 19th century [9]. Modern medicine has evolved considerably over the past 150 years, and milestones in the development of professional codes have been the declarations of the World Medical Association [34], which states promoting ethical professional codes as one of its main goals. The two most prominent declarations are the declarations of Geneva as a response to the cruelties performed by medical professionals in Germany and Japan during World War II [1]; and the declaration of Helsinki for ethical conduct of medical research [2]. These documents are continuously updated and received further refinement especially after disastrously unethical events, e.g. the revelation of the Tuskegee Syphilis study [22]. Based on these documents, professional medical associations around the world have drafted professional ethical codes. Importantly, these codes are specifically designed to not be dependent on legislation which can highly differ between countries [34].

Medical professionals are guided in their work by these ethical codes, and are protected from the social dilemma as the publicly known ethos enables them to refuse unethical behaviour without the fear of repercussions.³ Due to the similar level of expected impact on society, we view medicine as a suitable template for a professional code for AI development ethics. In the following section we outline how the field of AI needs to adapt in order to develop robust, impactful, and unified professional codes in analogy to medicine.

5 Towards professional codes in AI

In this section, we discuss the present issues with establishing ethical codes of conduct for AI developers and outline some possible paths towards establishing them.

5.1 Current issues

The topic of professional codes for AI has raised considerable interest during recent years, and several works have pointed out the fluid nature of this field and its complexity, see e.g. [15, 48]. Yet, we observe that there is little tangible practical impact, in the sense that there are no broadly

³ We do of course not claim that this system is foolproof and can prevent the social dilemma fully. The professional codes in medicine are, however, arguably the ones that protect their professionals the best in an area with highest societal impact.

accepted professional codes today. We believe that this can be attributed to two major reasons:

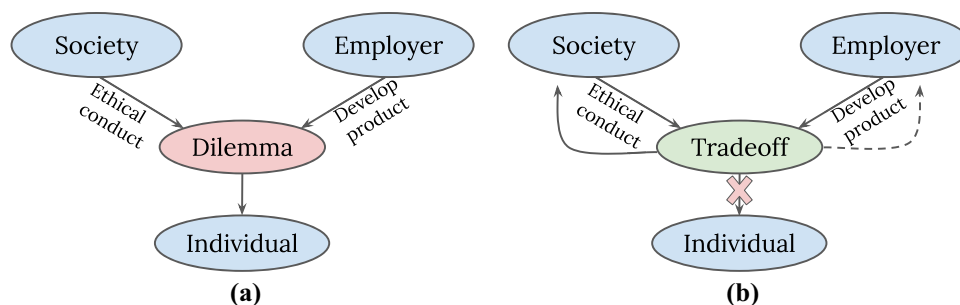
Primarily, current analyses accept many boundary conditions as given, instead of suggesting how to change these conditions. We exemplify this as follows. It is true that even if large organizations, such as the IEEE, adopted professional codes of AI development, a plethora of challenges would remain. Who is an AI developer? What is the incentive for someone to join these organizations and abide by the codes? What keeps the organisation from diverging from the code if the potential gain or cost avoided is substantial? Even worse, several competing organizations might publish different codes, fragmenting the field and making it impossible for the developers and the public to know the norms of the profession. Lastly, such efforts might even be hijacked by Big Tech: they could “support” certain organizations in publishing ethical codes, with significant influence on content, essentially leading to a new form of ethics washing. We agree that under these boundary conditions, an implementation of professional codes for AI seems difficult or even impossible. However, we argue that we must distinguish between an analysis of a situation and practical suggestions regarding how conditions can and should be changed. In brief, we argue that analyzing a situation will not change it; only changing relevant determining factors will.

Secondly, as long as current initiatives do not free developers from the social dilemma outlined earlier, an implementation of professional codes will inevitably fail. For example, recent work has addressed how embedding ethicists in development teams might support ethical AI development [53]. However, if these ethicists are themselves just other employees of the same corporation, the social dilemma applies to them as well.

5.2 Possible ways forward

We argue that in order to solve the crisis in AI development ethics, a process that addresses the two points in Sect. 5.1 must be initiated. Our primary proposition is that AI development must become a unified profession, taking medicine as an example. And, as in medicine, it must become licensed. The licence must be mandatory for all developers of medium to high risk AI systems, following, e.g., the Proposal for a Regulation laying down harmonised rules on artificial intelligence by the European Union [28]. This would protect the individual developer in an unprecedented manner. The chances of being replaced by another professional would be very small, since employers would know that all AI developers abide by the same code. Thus, AI developers could refuse to perform unethical development without fear of the social dilemma consequences. The difference before and after introducing a professional ethos is depicted in Fig. 1.

Fig. 1 a Now: Society's need for ethical conduct and the employer's need to develop products together put the developer into a dilemma, and **b** after introducing the ethos: what was previously a dilemma for the developer is now a trade-off that society, together with the employer, has to handle using established methods



Secondly, national AI developer organisations maintaining registers of employed AI developers must be established, analogously to national medical societies. These, including all their members, would serve as nuclei for the development of professional codes, and be responsible for maintaining, updating and refining them. With such a system in place, understanding and following the codes—the professional ethos—would replace the need for individual formal training in the methodology of ethics, as is the case in medicine. Lastly, unethical behaviour could lead to the loss of one's license, which is a strong incentive not to take part in unethical development, even if required by an employer. Note how legislation does not influence the content of the professional codes but facilitates it by creating the right boundary conditions.

In his 1983 work discussing professional obligation to society [4], Abbott stated that one of five basic properties of professional ethics is that “nearly all professions have some kind of formal ethical code”. While we argue that this should be the case for AI professionals, it is not the case today. A concrete suggestion of forming an ethos along the lines of the Hippocratic oath for developers of technology was recently put forward by Abbas et al. [3], who suggested a Hippocratic oath for technologists, consisting of three main parts: understanding the ethical implications of technology, telling the truth and acting responsibly. The authors suggest that technologists sign the oath publicly and digitally, receiving a digital badge to be included in online profiles. The main objective of the oath being to “raise awareness of the ethical responsibility of the technologist as a user and creator of technology”, the authors do not further define what exactly a ‘technologist’ is, other than a creator of technology.

We do not claim that unifying AI development into one profession is a simple task. On the contrary, we acknowledge all the challenges other authors, e.g. [56], have pointed out regarding defining who is an AI professional, and the complex interactions between all stakeholders in AI governance. The difference is that we do not focus on what hinders the process, but argue that establishing ethical AI development

will otherwise fail: As long as professionals can be uncertain regarding whether they are an AI developer, as long as corporations can claim that their employees are not AI developers, as long as we leave developers alone with their social dilemma, as long as there are no single international institutions serving as contact points for governments and corporations, and as long as there is no accountability for unethical AI development, no stable solution securing future AI development to be ethical will be found.

Although overcoming all obstacles to a unified AI developer profession will be a tedious endeavour, it will remove the social dilemma for developers. We argue that this is the only realistic way to ensure that AI development follows goals in alignment with societal benefit. Once a unified profession with professional codes exists, it will serve as a safeguard against unethical corporate and governmental interests. This is important as the role of corporations can be manifold, and a unified profession will help to steer their decisions in a direction aligned with society's ethical expectations. Removing the need for internal guidelines would also remove the possibility of using AI ethics as merely a marketing narrative. On the contrary, proven and audited adherence to professional codes could provide an economic benefit to AI companies.

5.3 An aside on AI as a profession

An exact definition of who is an AI professional is useful to identify who should—or even must—receive training in ethics for AI. The discussion about the professionalisation of AI is outside the scope of the current discussion on the social dilemma for AI developers. We do however, need to recognise the challenges with this point.

Gasser et al. [33] argue that fundamental conceptual issues such as the notion of what constitutes various “AI professions” remain not only open, but constitute a main challenge when discussing a single professional norm for developers of AI. The lack of clear definitions of the term ‘AI’ itself, what exactly a profession is and what constitutes

professional ethics create is, as formulated by [33], “a perfect definitional storm”. A central challenge in defining who is an AI professional, is that AI system design is multi-disciplinary and often involve individuals that already belong to another profession, with its own professional association. AI development can also be performed by “those working entirely outside the framework of any professional accreditation”, as remarked in [16]. These aspects are precisely those we state must be overcome, acknowledging that that AI development happens in a variety of contexts where other norms may already be relevant.

While [33] speculate that a sufficiently strong driver for an AI profession to evolve might emerge from a crisis—as seen historically when modern medical ethics formed from a realisation that there was more at stake than merely individuals and their professional work [10]—we urge that the social dilemma for AI developers must be resolved *before* a crisis is caused by unethical AI development.

6 Related work

The idea of a professional ethos for AI professionals and its role in achieving AI ethics has been argued for in the literature. We give a summary of some of these arguments.

Stahl [66] reviews different proposals put forward to address the challenges in the ethics of AI on three levels: the policy-level, the organisational level and the individual level. The latter consists mainly of guidance principles and documents for individuals, designed to handle AI systems that are already under development or in place. Stahl also points to the observation made by [30], that the large number of guidelines for individuals can cause confusion and ambiguity, a challenge we have also addressed in this paper, and argue is best solved by developing a professional ethos for AI professionals.

Stahl [66] focuses on the dilemma of control - he stresses the importance of identifying and considering ethical issues early on during the development process, pointing out the relevance of the Collingridge dilemma [25]. This dilemma, also known as the dilemma of control, is the observation that “it is relatively easy to intervene and change the characteristics of a technology early in its life cycle. However, at this point it is difficult to predict its consequences. Later, when the consequences become more visible, it is more difficult to intervene”. This dilemma is particularly relevant for those in the position to address ethical issues during the development process. As the developers of a system are in a position to make changes to a system during the early stages, these should also be made most responsible for—and capable of—performing such changes. They should not be hindered by fear of repercussions, but rather encouraged by their professional responsibility.

In another recent paper [67], Stahl analyses similarities between the computer ethics discourse of the 1980s and the AI ethics debate of today. He argues that focus should not be on the relevant “technical artefact”, i.e. the computer or AI system, but rather that ethical issues arise in the context of socio-technical systems [67]. He points out that “One proposal that figured heavily in the computer ethics discourse that is less visible in the ethics of AI is that of professionalism.”, highlighting exactly the part of the discourse we argue is missing.

Referring to literature on ethics for computing professionals, i.a. [6, 45], Stahl observes that the development of professional bodies for computing has been driven exactly by the idea of “institutionalising professionalism as a way to deal with ethical issues” [67].

Analysing the normative features of ethical codes in software engineering, Gogoll et al. [36] argue that codes of conduct “are barely able to provide normative orientation in software development”, and that their value-based approach potentially prevents them from being useful from a normative perspective. Such codes being underdetermined, the authors argue that they cannot replace ethical deliberation, and rather damage the process and decrease the ethical value of the outcome. The authors instead propose to implement ethical deliberation within software development teams. This is in line with the arguments of Borenstein et al. [18], who in their recent work on the need for AI ethics education, discuss the fostering of a professional mindset among AI developers.

Discussing how AI developers view their professional responsibilities, Gogoll et al. [36] observe that “Oftentimes, developers believe that ethics is someone else’s problem.”, conveying that AI developers sometimes view themselves as dealing with the technology, and ethics being the responsibility of somebody else. This attitude can be seen as a symptom of social dilemmas the developers face, but do not necessarily recognise as such. Namely, it would be reasonable that the developers would be keen to avoid being put in an ethically challenging situation for which they might be held liable but are not given the tools or power to address adequately.

The discussion on how to educate AI professionals, including whether they should be trained in ethics, is highly relevant in the context of professionalising AI, but somewhat outside of scope of our social dilemma for AI developers discussion which is why we refrain from a detailed related work overview on this topic. We will mention Borenstein et al. [18] who, stressing the need for AI ethics education, state that while many remedies to the ethical challenges resulting from AI have been proposed, a key piece of the solution is “enabling developers to understand that the technology they are building is intertwined with ethical dimensions, and that, as developers, they have a vital role and responsibility to

engage with ethical considerations". The authors conclude that an important part of future ethical AI is "making sure that ethics has a central place in AI educational efforts."

We do need to remark that professional education is a core offer of professional societies and thus professionalization of AI development would in turn allow broad education about ethical issues. However, the role of continuous accreditation cannot be discounted by education alone. A parallel can be drawn to existing credential maintenance programs, such as for example that of the US Green Building Council's Leadership in Energy and Environmental Design (LEED) professional credentialing service.⁴

7 Limitations

In the absence of the possibility to test policy interventions on societies in a randomized and controlled fashion, the decisions about the best way to achieve a certain goal, e.g. ethical AI development, naturally remain uncertain. We acknowledge that the impact of professional codes might be less prominent than we believe. We also acknowledge the uncertainty in whether the professionalization of AI development will lead to the desired effect that we outline. We are convinced, however, that the current debate will profit from the inclusion of the social dilemma aspect and a discussion of the potential solution that we suggest.

8 Conclusion

AI technology has the potential for substantial advancements but also for negative impacts on society, and thus requires assurance of ethical development. However, despite massive interest and efforts, the implementation of ethical practice into AI development remains an unsolved challenge, which in our view renders it obvious that the current approach to AI development ethics fails to provide such assurance. Our position is that the current, guideline-based, approach to AI development ethics fails to have an impact where it matters. We argue that the key to ethical AI development at this stage is solving the social dilemma for AI developers, and that this must be done by unifying AI development into a single profession. Furthermore, we argue that, based on observations from the mature field of medicine, a unified professional ethos is necessary to ensure a stable situation of ethical conduct that is beneficial to society. While we certainly do not claim that removing the social dilemma for AI developers is sufficient for solving all issues of AI development ethics, we argue that it is necessary.

We have discussed ethical considerations from the perspective of added cost, but would like to also point out that ethical development has itself proper value. Awareness of ethical responsibilities both inwards (towards the corporation and peers) and outwards (towards clients and society) leads directly to the protection of assets and reputation. Professional objectives in line with ethical values leads to increased dedication and sense of ownership, resulting in higher quality deliverables. Practice in ethical consideration and evaluation processes improves professionals' decision making and implementation abilities, making them more willing to adapt to changes required for sustainability. Focus on ethical considerations fosters a culture for openness, trust and integrity, which again decreases the risk of issues being downplayed. Outstanding professionals with the privilege to choose among several employers are likely to consider not only the opportunity for professional growth, but also whether they can expect their future employer to treat them and their peers justly and ethically.

By focusing on the social dilemma we have added additional pressure to motivate the development of professional codes for AI ethics. Much remains to be done to operationalise this desired professional certification framework.

We can observe that the medical professional ethical code is built on a long-standing tradition of professional codes. In the field of AI, we do not have the benefit of such a historical and globally recognised entity. Thus, the first step will be to agree on the core values and principles that apply to any AI developer in any context. The next step will be to operationalise those values and principles on a national level by establishing a certification framework for AI developers. Governments do not need to be left on their own when developing this certification frameworks, as it can be based on the experience with many national medical certification frameworks.

Acknowledgements We thank Dr. Daniel Strech, Dr. Michelle Livne and Dr. Nora A. Tahy for the thorough review of our manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

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

⁴ <https://www.usgbc.org/resources/cmp-guide>.




References

1. WMA—the world medical association-WMA declaration of Geneva. <https://www.wma.net/policies-post/wma-declaration-of-geneva/>
2. WMA—the world medical association-WMA Declaration of Helsinki—ethical principles for medical research involving human subjects. <https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/>
3. Abbas, A.E., Senegés, M., Howard, R.A.: A hippocratic oath for technologists. *Next-Generation Ethics: Engineering a Better Society* 71 (2019)
4. Abbott, A.: Professional ethics. *Am. J. Sociol.* **88**(5), 855–885 (1983)
5. Akerlof, G.A., Kranton, R.: Identity and the economics of organizations. *J. Econ. Perspect.* **19**, 9–32 (2005)
6. Albrecht, B., Christensen, K., Dasigi, V., Huggins, J., Paul, J.: The pledge of the computing professional: recognizing and promoting ethics in the computing professions. *SIGCAS Comput. Soc.* **42**(1), 6–8 (2012). <https://doi.org/10.1145/2422512.2422513>
7. Allison, S.T., Beggan, J.K., Midgley, E.H.: The quest for “similar instances” and “simultaneous possibilities”: metaphors in social dilemma research. *J. Pers. Soc. Psychol.* **71**(3), 479–497 (1996). <https://doi.org/10.1037/0022-3514.71.3.479>
8. Association, I.S.: Ieee 7000-2021—ieee standard model process for addressing ethical concerns during system design (2021). <https://standards.ieee.org/standard/7000-2021.html>
9. Backof, J.F., Martin, C.L.: Historical perspectives: development of the codes of ethics in the legal, medical and accounting professions. *J. Bus. Ethics* **10**(2), 99–110 (1991). <https://doi.org/10.1007/BF00383613>
10. Baker, R.: Codes of ethics: some history. *Perspect. Prof.* **19**(1), 3–4 (1999)
11. Barocas, S., Hardt, M., Narayanan, A.: *Fairness and Machine Learning*. fairmlbook.org (2019). <http://www.fairmlbook.org>
12. Baum, S.D.: Social choice ethics in artificial intelligence. *AI Soc.* **35**(1), 165–176 (2020). <https://doi.org/10.1007/s00146-017-0760-1>
13. Bietti, E.: From ethics washing to ethics bashing: A view on tech ethics from within moral philosophy. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT 20*, pp. 210–219. Association for Computing Machinery, New York, NY, USA (2020). <https://doi.org/10.1145/3351095.3372860>
14. Biggs, R.D.: Medicine, surgery, and public health in ancient Mesopotamia. *Civil. Ancient Near East.* **3**, 1911 (1995). ((ISBN: 9780684197227))
15. Boddington, P.: *Towards a Code of Ethics for Artificial Intelligence Artificial Intelligence: Foundations, Theory, and Algorithms*. Springer, Cham (2017). <https://doi.org/10.1007/978-3-319-60648-4>
16. Boddington, P.: *Towards a Code of Ethics for Artificial Intelligence*. Springer, Berlin (2017)
17. Bonnefon, J.F., Shariff, A., Rahwan, I.: The social dilemma of autonomous vehicles. *Science* **352**(6293), 1573–1576 (2016). <https://doi.org/10.1126/science.aaf2654>
18. Borenstein, J., Howard, A.: Emerging challenges in AI and the need for AI ethics education. *AI Ethics* **1**(1), 61–65 (2021)
19. Bughin, J., Seong, J.: Assessing the economic impact of artificial intelligence (2018). https://www.itu.int/dms_pub/itu-s/opb/gen/S-GEN-ISSUEPAPER-2018-1-PDF-E.pdf
20. Camenisch, P.F.: *Grounding Professional Ethics in a Pluralistic Society*. Haven Publications, New York (1983)
21. Castets-Renard, C.: *Human Rights and Algorithmic Impact Assessment for Predictive Policing*. SSRN Scholarly Paper ID 3890283, Social Science Research Network, Rochester, NY (2021). <https://papers.ssrn.com/abstract=3890283>
22. Chadwick, G.L.: Historical perspective: Nuremberg, Tuskegee, and the radiation experiments. *J. Int. Assoc. Physicians AIDS Care* **3**(1), 27–28 (1997)
23. Charisi, V., Dennis, L.A., Fisher, M., Lieck, R., Matthias, A., Slavkovik, M., Sombetzki, J., Winfield, A.F.T., Yampolskiy, R.: Towards moral autonomous systems. *CoRR arXiv:1703.04741* (2017)
24. Cinelli, M., Morales, G.D.F., Galeazzi, A., Quattrociocchi, W., Starnini, M.: Echo chambers on social media: a comparative analysis (2020)
25. Collingridge, D.: *The social control of technology* (1982)
26. Darley, J.M., Latané, B.: Bystander intervention in emergencies: diffusion of responsibility. *J. Pers. Soc. Psychol.* **8**(4, Pt.1), 377–383 (1968). <https://doi.org/10.1037/h0025589>
27. Ebell, C., Baeza-Yates, R., Benjamins, R., Cai, H., Coeckelbergh, M., Duarte, T., Hickok, M., Jacquet, A., Kim, A., Krijger, J., MacIntyre, J., Madhamshettiwar, P., Maffeo, L., Matthews, J., Medsker, L., Smith, P., Thais, S.: Towards intellectual freedom in an ai ethics global community. *AI and Ethics* pp. 1–8 (2021). <https://europepmc.org/articles/PMC8043756>
28. European Commission: *Proposal for a regulation laying down harmonised rules on Artificial Intelligence* (2021). <https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence>
29. Feldstein, S.: *The global expansion of AI surveillance*. Tech. rep., Carnegie Endowment for International Peace (2019). <http://www.jstor.org/stable/resrep20995.4>
30. Floridi, L., Cowl, J.: A unified framework of five principles for AI in society. *Harvard Data Science Review* **1**(1) (2019). <https://doi.org/10.1162/99608f92.8cd550d1>. <https://hdsr.mitpress.mit.edu/pub/10jsh9d1>
31. Floridi, L., Cowl, J., King, T.C., Taddeo, M.: How to design AI for social good: seven essential factors. *Sci. Eng. Ethics* **26**(3), 1771–1796 (2020). <https://doi.org/10.1007/s11948-020-00213-5>
32. Frankel, M.S.: Professional codes: why, how, and with what impact? *J. Bus. Ethics* **8**(2–3), 109–115 (1989). <https://doi.org/10.1007/BF00382575>
33. Gasser, U., Schmitt, C.: The role of professional norms in the governance of artificial intelligence. In: *The Oxford Handbook of Ethics of AI*, p. 141. Oxford University Press, Oxford (2020)
34. Gillon, R.: Medical oaths, declarations, and codes. *Br. Med. J. (Clin. Res. Ed.)* **290**(6476), 1194–1195 (1985). <https://doi.org/10.1136/bmj.290.6476.1194>
35. Goelz, P., Kahng, A., Procaccia, A.D.: Paradoxes in fair machine learning. In: H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, R. Garnett (eds.) *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc. (2019). <https://proceedings.neurips.cc/paper/2019/file/bbc92a647199b832ec90d7cf57074e9e-Paper.pdf>
36. Gogoll, J., Zuber, N., Kacianka, S., Greger, T., Pretschner, A., Nida-Rümelin, J.: Ethics in the software development process: from codes of conduct to ethical deliberation. *Philos. Technol.* **20**, 1–24 (2021)
37. Hagendorff, T.: The ethics of AI ethics: an evaluation of guidelines. *Mind. Mach.* **30**(1), 99–120 (2020). <https://doi.org/10.1007/s11023-020-09517-8>
38. Haner, J., Garcia, D.: The artificial intelligence arms race: trends and world leaders in autonomous weapons development. *Glob. Pol.* **10**(3), 331–337 (2019). <https://doi.org/10.1111/1758-5899.12713>

39. Hao, K.: I started crying: Inside timnit gebru's last days at google—and what happens next (2020). <https://www.technologyreview.com/2020/12/16/1014634/google-ai-ethics-lead-timnit-gebru-tells-story/>
40. Hao, K.: We read the paper that forced timnit gebru out of google. Here's what it says (2020). <https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/>
41. Hardin, G.: The tragedy of the commons. *Science* **162**(3859), 1243–1248 (1968). <https://doi.org/10.1126/science.162.3859.1243>
42. Jamal, K., Bowie, N.E.: Theoretical considerations for a meaningful code of professional ethics. *J. Bus. Ethics* **14**(9), 703–714 (1995). <https://doi.org/10.1007/BF00872324>
43. Jensen, M.C., Meckling, W.H.: Theory of the firm: managerial behavior, agency costs and ownership structure. *J. Financ. Econ.* **3**(4), 305–360 (1976)
44. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* **1**, 20 (2019). <https://doi.org/10.1038/s42256-019-0088-2>
45. Johnson, D.G.: Computing ethics computer experts: guns-for-hire or professionals? *Commun. ACM* **51**(10), 24–26 (2008)
46. Johnson, K.: Google targets AI ethics lead margaret mitchell after firing timnit gebru (2021). <https://venturebeat.com/2021/01/20/google-targets-ai-ethics-lead-margaret-mitchell-after-firing-timnit-gebru/>
47. Kollock, P.: Social dilemmas: the anatomy of cooperation. *Ann. Rev. Sociol.* **24**(1), 183–214 (1998). <https://doi.org/10.1146/annurev.soc.24.1.183>
48. Larsson, S.: On the governance of artificial intelligence through ethics guidelines. *Asian J. Law Soc.* **7**(3), 437–451 (2020). <https://doi.org/10.1017/als.2020.19>
49. Lloyd, W.F.: W. F. Lloyd on the checks to population. *Popul. Dev. Rev.* **6**(3), 473–496 (1980). <http://www.jstor.org/stable/1972412>
50. Luce, R.D., Raiffa, H.: *Games and Decisions: Introduction and Critical Survey*. Wiley, Chicago (1957)
51. Mary, J., Calauzènes, C., Karoui, N.E.: Fairness-aware learning for continuous attributes and treatments. In: K. Chaudhuri, R. Salakhutdinov (eds.) *Proceedings of the 36th International Conference on Machine Learning, Proceedings of Machine Learning Research*, vol. 97, pp. 4382–4391. PMLR (2019). <http://proceedings.mlr.press/v97/mary19a.html>
52. McGregor, L., Murray, D., Ng, V.: International human rights law as a framework for algorithmic accountability. *Int. Comp. Law Q.* **68**(2), 309–343 (2019). <https://doi.org/10.1017/S0020589319000046>
53. McLennan, S., Fiske, A., Celi, L.A., Müller, R., Harder, J., Ritt, K., Haddadin, S., Buyx, A.: An embedded ethics approach for AI development. *Nat. Mach. Intell.* **2**(9), 488–490 (2020). <https://doi.org/10.1038/s42256-020-0214-1>
54. Metzinger, T.: Ethics washing made in Europe (2019). <https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html>. Editorial
55. Mhlambi, S.: *From Rationality to Relationality: Ubuntu as an Ethical and Human Rights Framework for Artificial Intelligence Governance*. Carr Center Discussion Paper Series (2020-009) (2020)
56. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. *Nat. Mach. Intell.* **1**(11), 501–507 (2019). <https://doi.org/10.1038/s42256-019-0114-4>
57. Olson, M.: *The Logic of Collective Action: Public Goods and the Theory of Groups*, Second Printing with a New Preface and Appendix. Harvard University Press (1971). <http://www.jstor.org/stable/j.ctvjf3ts>
58. Pavalko, R.M.: *Sociology of occupations and professions*. Itasca, Ill. : F.E. Peacock (1988). <http://archive.org/details/sociologyfoccup00pava>
59. Perrow, C.B., Olson, M.: Review: [untitled]. *Soc. Forces* **52**(1), 123–125 (1973). <http://www.jstor.org/stable/2576430>
60. Rao, A., Verweij, G.: Sizing the prize (2017). <https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf>
61. Rességuier, A., Rodrigues, R.: AI ethics should not remain toothless! a call to bring back the teeth of ethics. *Big Data Soc.* **7**(2), 2053951720942541 (2020). <https://doi.org/10.1177/2053951720942541>
62. Sabel, C.F.: Beyond principal-agent governance: experimentalist organizations, learning and accountability *De Staat van de Democratie. Democratie voorbij de Staat. WRR Verkenning* **3**, 173–195 (2004)
63. Scheck, J., Purnell, N., Horwitz, J.: Facebook employees flag drug cartels and human traffickers. The company's response is weak, documents show. *The Wall Street Journal*. <https://www.wsj.com/articles/facebook-drug-cartels-human-traffickers-response-is-weak-documents-11631812953?mod=e2tw>
64. Schiff, D., Biddle, J., Borenstein, J., Laas, K.: What's next for AI ethics, policy, and governance? a global overview. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES '20*, pp. 153–158. Association for Computing Machinery, New York, NY, USA (2020). <https://doi.org/10.1145/3375627.3375804>
65. Smuha, N.A.: Beyond a human rights-based approach to AI governance: promise, pitfalls AND Plea. *Philos. Technol.* (2020). <https://doi.org/10.1007/s13347-020-00403-w>
66. Stahl, B.C.: *Addressing Ethical Issues in AI*, pp. 55–79. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-69978-9_5
67. Stahl, B.C.: From computer ethics and the ethics of AI towards an ethics of digital ecosystems. *AI Ethics* **20**, 1–13 (2021)
68. Szczepański, M.: Economic impacts of artificial intelligence (AI) (2019). [https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_BRI\(2019\)637967](https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_BRI(2019)637967). Briefing
69. Verma, S., Rubin, J.: Fairness definitions explained. In: *Proceedings of the International Workshop on Software Fairness, FairWare '18*, pp. 1–7. Association for Computing Machinery, New York, NY, USA (2018). <https://doi.org/10.1145/3194770.3194776>
70. Wagner, B.: Ethics as an escape from regulations: From “ethics-washing” to ethics-shopping?. *Amsterdam University Press*, pp. 84–89 (2018). <http://www.jstor.org/stable/j.ctvhrd092.18>
71. Winfield, A.F.T., Booth, S., Dennis, L.A., Egawa, T., Hastie, H., Jacobs, N., Muttram, R.I., Olszewska, J.I., Rajabiyazdi, F., Theodorou, A., Underwood, M.A., Wortham, R.H., Watson, E.: Ieee p7001: a proposed standard on transparency. *Front. Robot. AI* **8**, 225 (2021). <https://doi.org/10.3389/frobt.2021.665729>
72. Yeung, K., Howes, A., Pogrebna, G.: AI Governance by Human Rights-Centered Design, Deliberation, and Oversight (2020). <https://doi.org/10.1093/oxfordhb/9780190067397.013.5>. <https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780190067397.001.0001/oxfordhb-9780190067397-e-5>
73. Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., Lyons, T., Manyika, J., Niebles, J.C., Sellitto, M., Shoham, Y., Clark, J., Perrault, R.: The AI index 2021 annual report (2021). <https://aiindex.stanford.edu/report/>
74. Zwitter, M.: Ethical Codes and Declarations. In: *Medical Ethics in Clinical Practice*, pp. 7–13. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-00719-5_2

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

Author

Inga Strümke^{1,2}  · Marija Slavkovik³  · Vince Istvan Madai^{4,5,6} 

Inga Strümke
inga.strumke@ntnu.no

Marija Slavkovik
marija.slavkovik@uib.no

¹ Department of Engineering Cybernetics, NTNU, Trondheim, Norway

² Department of Holistic Systems, SimulaMet, Oslo, Norway

³ Department of Information Science and Media Studies, University of Bergen, Bergen, Norway

⁴ QUEST Center for Responsible Research, Berlin Institute of Health (BIH), Charité Universitätsmedizin Berlin, Berlin, Germany

⁵ Charité Lab for AI in Medicine, Charité Universitätsmedizin Berlin, Berlin, Germany

⁶ School of Computing and Digital Technology, Faculty of Computing, Engineering and the Built Environment, Birmingham City University, Birmingham, UK