Education, Ethical Dilemmas and AI: From Ethical Design to Artificial Morality

Joan Casas-Roma(✉), Jordi Conesa, and Santi Caballé

SmartLearn Research Group, Universitat Oberta de Catalunya, Barcelona, Spain
{jcasasrom,jconesac,scaballe}@uoc.edu
http://smartlearn.uoc.edu

Abstract. Ethical dilemmas are complex scenarios involving a decision between conflicting choices related to ethical principles. While considering a case of an ethical dilemma in education presented in [17], it can be seen how, in these situations, it might be needed to take into consideration the student's needs, preferences, and potentially conflicting goals, as well as their personal and social contexts. Due to this, planning and foreseeing ethically challenging situations in advance, which would be how ethical design is normally used in technological artifacts, is not enough. As AI systems become more autonomous, the amount of possible situations, choices and effects their actions can have grow exponentially. In this paper, we bring together the analysis of ethical dilemmas in education and the need to incorporate moral reasoning into the AI systems' decision procedures. We argue how ethical design, although necessary, is not sufficient for that task and that artificial morality, or equivalent tools, are needed in order to integrate some sort of "ethical sensor" into autonomous systems taking a deeper role in an educational settings in order to enable them to, if not resolve, at least identify new ethically-relevant scenarios they are faced with.

Keywords: AI Ethics · Online learning · Artificial morality · Ethical sensors

1 Introduction and Motivations

The new disciplinary approach of learning engineering as the merge of breakthrough educational methodologies and technologies based on internet, data science and artificial intelligence (AI) have completely changed the landscape of online education over the last years by creating accessible, reliable and affordable data-rich powerful learning environments [12]. Particularly, AI-driven technologies have managed to automate pedagogical behaviours that we would deem as "intelligent" within an online education setting.

This work has been supported by the project colMOOC "Integrating Conversational Agents and Learning Analytics in MOOCs", co-funded by the European Commission (ref. 588438-EPP-1-2017-1-EL-EPPKA2-KA), and by a UOC postdoctoral stay.

R. A. Sottilare and J. Schwarz (Eds.): HCII 2021, LNCS 12792, pp. 167–182, 2021.
https://doi.org/10.1007/978-3-030-77857-6_11

However, as reported in more mature sectors where AI-driven technologies have already been developed and deployed, automatic decision-making processes many times bear unexpected outcomes. For instance, machine learning (ML) based systems have been reported to discriminate certain social communities in the context of law courts, job applications or bank loans due to the use of biased datasets to feed the ML models [4,13,25]. Different studies conclude that, in order to avoid unforeseen outcomes in their integration, the ethical dimension of deploying AI in different settings must be taken into account. This becomes particularly important when thinking about the effects that applying AI systems to education could have to current and future generations of students. Due to this, special care needs to be taken when considering how AI systems could deal with ethical dilemmas that can appear in an educational setting.

In order to provide a starting point and guide our discussion throughout this paper, let us consider the following case of an ethical dilemma in the context of education, as it appears in [17]. An eight-grade student's marks are not enough to pass to ninth-grade, and her teachers agree that she is unprepared for the next grade. Should the student be allowed to pass? Given the standard norms, the automatic answer might be "no". However, we have some more information available about the student; we know she is likely to drop out entirely if she is not allowed to pass, and her teachers also note that she has put a lot of effort that resulted into improving her grades, until she recently grew discouraged. Given these new bits of information, should the student be allowed to pass? We still have some more details about this case, though: she has lived in three foster homes for the past years, and her brother died from a gunshot. Furthermore, a potential alternative school for struggling students is a well-known "school-to-prison" pipeline. Again, should the student be allowed to pass?

Although this is an example of a quite extreme case, it shows how, in order to evaluate and make a decision about a situation with clear ethical effects in their outcomes, one needs to consider a broad picture of the scenario. In this particular case, and even though the dilemma takes place in an educational setting, the elements that need to be considered step "beyond the classroom"; namely, the situation starts being shaped as a dilemma as soon as we start considering not just the student's information that we would normally find represented within the educational system (marks, grade pass, etc.), but also the student's personal and contextual situation. In this sense, what makes this situation particularly challenging goes beyond the usual norms that one would apply in the educational system and step right into the student's own case. As it is pointed out in [9], ethical dilemmas are often about the exception, rather than the norm, and they usually involve solutions with potentially conflicting goals that cannot all be fully satisfied at the same time. As such, ethical dilemmas do not usually have a clearly "good" outcome, as one solution favoring one dimension will often disregard another one.

After having introduced the guiding case study, we introduce the notion of the layered approach to ethical dilemmas in Sect. 1.1. We explore distinct considerations related to the integration of ethical behaviors in technological

tools in Sect. 2. With this considerations at hand, we discuss the challenges that each layer of our guiding case study would pose in Sect. 3. Having identified the complexities behind this kind of dilemmas, we introduce the notion of *ethical sensors* in Sect. 4. Finally, we provide some conclusions and directions of future work in Sect. 5.

1.1 Ethical Dilemmas in Education: A Layered Approach

The previous dilemma allows us to distinguish three different layers that should be taken into account when considering the ethical dimension of a conflicting situation like the one depicted in the case study: the *Educational* layer, the *Personal* layer, and the *Social* layer. In a nutshell, those layers (see Fig. 1) distinguish three contexts that, although being all potentially important in an ethically-relevant scenario, belong to different spheres of the student's learning context.

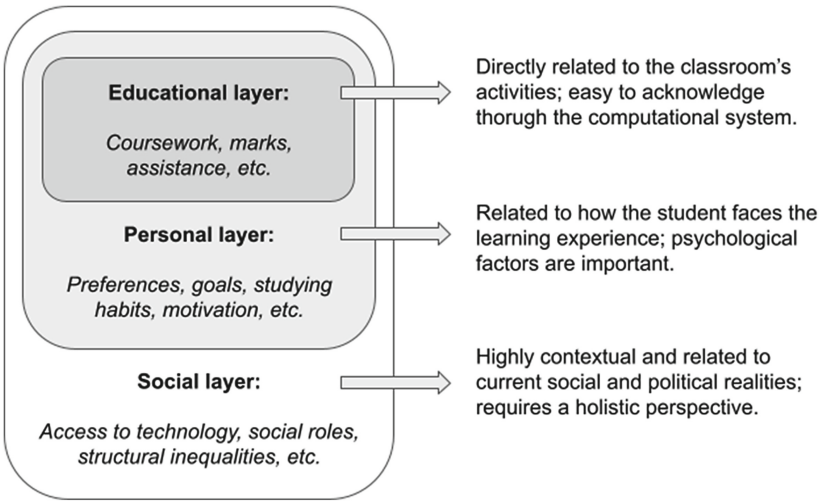

Fig. 1. The different ethically-relevant layers of a student's learning experience.

The Educational layer refers to those elements that belong to, and are explicitly accounted for, the educational context –namely, anything that would normally take place within the classroom. Course contents, classroom activities, evaluations, homework, etc. These elements are already part of the student's *persona* within the educational environment, and they aim to measure their knowledge, progress and skills within the learning process. They are the most readily-available elements for an educational institution to look at, as they naturally fall within the scope of what the students do in their learning process and within the standard course of events of their learning. In the case study presented in the previous section, those elements would correspond to:

- The student's final marks.
- The norm requiring students to achieve a certain mark in order to pass to next grade.
- The student's marks record.

The Personal layer refers to those elements related to the student's way of being, their goals, preferences and motivations, the way the student faces learning challenges, etc. –namely, they are part of what makes each and every individual person be the way they are. Even though these elements are not explicitly taken into account within the educational system, they have a direct effect in the way the student approaches their learning process. Even though not being explicitly represented in the educational environment, they clearly bear a direct relationship with the student's learning journey, and are often known and taken into account by human actors involved in the learning process. In the previous case study, those elements would correspond to:

- The student's intention to drop out from the educational system if she is not allowed to pass to ninth grade.
- The student's effort (and success) in raising her marks in the past through more dedication.
- The student's discouragement after having improved her marks, which resulted in her results worsening again.

The Social Layer refers to those elements belonging to the student's context, but which are external to their way of being. These include, but are not limited to, the people with whom they share their life (family, friends, etc.), the place where they live (home, geographical area), relationships and responsibilities they may have towards other people, past and current events that might be affecting the student's life significantly, as well as socio-political and historical particularities of the student's social context (which might be related to ethnicity, gender roles, etc.). These contextual elements can have a big effect on the student's life and, consequently, on the student's learning process. Aside from potentially affecting the student's access to educational resources, they can have an effect on the way the student behaves, the way the student devotes their time to learning, and can even frame the student in specific roles related to different social communities and contexts. In the previous case study, those elements would be (among others, but focusing on the ones that are being explicitly mentioned):

- The student having lived in three different foster homes for the past years, which indicates an unusual and potentially troublesome family structure for the student.
- The student's brother having died from a gunshot. This not only highlights an important personal loss for the student that can have profound emotional consequences, but might also suggest troublesome living conditions for the student and her family.

Even though this classification is not meant to be exhaustive, it is enough to show how these three different layers play a quite important role when considering an ethical dilemma such as the one presented in the case study. Furthermore,

this classification allows us to see how each further layer is harder to explicitly account for by using the tools of the educational system itself, but, at the same time, each further layer might point to deeper factors related to the student's situation that need to be considered in the dilemma. How, if possible at all, can all this be acknowledged in order to be used as part of a semi-autonomous decision-making system within a learning environment?

2 Beyond Tools: AI and Ethical Behavior

Before trying to answer that question, we first need to examine what the relation between technology and ethical[1] challenges has been. The use of AI in decision-making was seen, years ago, as the most reliable way of eliminating human bias and unfair decisions [8]; it was thought that data was objective and that computational systems were neutral with regards to interests and prejudices, and thus it was believed that those systems would be able to make neutral and fair decisions much more easily that any human would. Nevertheless, researchers soon realized that this was not the case. The way data was gathered, represented, selected and used, the way algorithms were encoded, the rules governing automated decision systems, all those pieces of the mechanism could easily encapsulate personal, social and historical biases in a wide variety of ways [8,13,18]. The question, then, arose: how could AI systems be made in such a way to prevent unintentional harm from being done?

Even though computers are clearly technological tools, the way computational artifacts have evolved in the recent decades sets them apart from other technological creations [14]. There currently is a strong distinction between a computer program and the traditional notion of a tool, such as a screwdriver, a jackhammer, or even a hand calculator. Perhaps the most evident distinction is that, while a traditional tool waits for someone to use it, AI programs can act somewhat autonomously react to, and affect their environment. Due to this, the ethical considerations traditionally applied to the design and use of technology (safety mechanisms, emergency buttons, etc.) no longer fill the needs behind AI systems. As a tool gets more autonomous, the responsibility for its ethical use gets farther away from its intended user and needs to consider a broader set of scenarios.

The study of the ethical dimension of artificial agents has led to some different classifications of both what constitutes and ethical agents, and what kinds of ethical agents there might exist. Moor distinguishes in [19] between four kinds of (non-exclusive) ethical agents:

– *Ethical impact agents*: Those agents whose actions have ethical consequences, regardless of whether these are intended, or not.

[1] Although the terms "ethics" and "morality" have slightly different definitions (one being a more reflective discipline, while the other one being more about prescription of behavior), we use them interchangeably in this work to refer to behaviors that are both in accordance to certain ethical principles, as well as considered to bear "good", or "right" outcomes.

- *Implicit ethical agents*: Those agents that have ethical considerations (normally, safety, or security considerations) built into their design.
- *Explicit ethical agents*: Those agents that can identify and process ethical information, as well as use it to make sensitive decisions on what should be done.
- *Full ethical agents*: Those agents who, aside from being able to identify and process ethical information, have those metaphysical features that are usually attributed to human agents; namely, consciousness, intentionality and free will.

Similarly, Wallach and Allen [23] define three layers of moral agency based on the two properties of *autonomy* (the degree in which an agent can act independently) and *sensitivity* (the degree in which the agent can identify and factor ethical information into their decision system):

- *Operational morality*: Agents with both low autonomy and low sensitivity, but which have some ethical considerations engineered in their design.
- *Functional morality*: Agents that either have high autonomy and low sensitivity, or the other way around (i.e.: low autonomy and high sensitivity).
- *Full moral agency*: Agents with high degrees of both autonomy and sensitivity, capable of acting as "trustworthy moral agents" [23, p. 26].

Wallach and Allen explicitly refer to Moor's categorization and, although they agree with Moor's aim and approach towards explicit ethical agents, they also point out how Moor does not provide instructions regarding how this direction should be pursued. In this sense, the authors defend how their account of the development of technology based on an interaction between autonomy and sensitivity provides good directions. As increased autonomy is an already ongoing trend in technological advancement, the question behind artificial moral agency requires an increase in ethical sensitivity.

The challenge behind the design of artificial ethical agents has usually been tackled through *ethical design*. Nevertheless, as soon as we recognize that the next steps behind that challenge lie in increasing ethical sensitivity, ethical design may be faced with certain limitations requiring us to divert our attention to a more explicit approach to ethical reasoning: the creation of *artificial morality*.

2.1 Ethics by Design: Forewarned is Forearmed

Ethical design faces the ethical challenges behind technology through anticipation [18]. When designing a new technological artifact, considering what this new artifact can do, who might use it, how it may be used, and what outcomes their uses can bring about helps understand situations in which the artifact can have an ethically-relevant effect. Once this combination of internal (the artifact's allowances) and external factors (its users, potential contexts, etc.) is considered, the designers can anticipate risks and dangers and integrate those into the design of the artifact itself.

One can find ethical design in technology way before complex AI-driven systems, and it can easily be found in almost any kind of technological tool [14,23]; emergency buttons on tools that, if they were to get out of control, could cause severe damage (like jackhammers, kitchen blenders, motorbikes), manual safety blocks and latches in firearms to prevent unintended firing, etc. However, the more autonomous technological artifacts become, the more complex their "safety mechanisms" must become. For instance, internet search engines are equipped with automated filtering tools to prevent showing inappropriate content to unintended audiences; a search engine *could* show these results, but its potential searches are limited beforehand due to ethical reasons. Similarly, a plane's autopilot system has a constrained range of manoeuvres it can perform, with limited speeds, turning, and ascend/descend angles; beyond what is mechanically and physically possible, these limitations are imposed in order to avoid discomfort to the passengers. More complex systems, such as an ML-based algorithm programmed to decide whether an applicant can get a bank loan, have been known to show biases and unfair behaviors [13]; among other options, ethical design can be applied in order to pre-process the data to filter those fields that should not play a role in the decision-making.

When considering the categories introduced in the Sect. 2, ethical design would likely lead to what Moor classifies as ethical impact agents and implicit ethical agents, and to what Wallach and Allen call operational morality. Even in the case of fairly complex systems (such as ML-based automated decision systems), ethical concerns are explored *beforehand*, planned and dealt with in advance. Although this does not mean that the system cannot be checked, revised and improved over time, aside from the ethically-relevant situations that have been foreseen in its design, the system does not adapt. Furthermore, there is no explicit representation of the ethical weight of the system's actions. As such, ethical design leads to systems that, regardless of their degree of autonomy, lack ethical sensitivity (following Wallach and Allen's terminology).

In order to leap this gap and reach some sort of explicit ethical sensitivity, which would be necessary for artifacts exhibiting functional morality, we need to define and embed morality as part of the system's decision procedures. In order to do this, ethical design is not enough: we must take a step forward and venture into the realm of artificial morality.

2.2 Artificial Morality: Towards Encoding Moral Value

Even though certain behaviors can be encouraged or limited through rules, norms and patterns, ethical behavior usually requires some sort of awareness of what is at stake in a situation. Take, for instance, a famous case in the fiction literature of ethical autonomous systems: Isaac Asimov's *I, robot* [5]. Although being a fictional work, Asimov's rules of robotics have been thoroughly considered and discussed as a potential starting point for ethical machines [3,11]; needless to say, this set of rules has been shown to lead to paradoxes that would make them insufficient to guide artificial ethical behavior. Nevertheless, even if we hypothetically accept that those rules are good enough to guide ethical robots,

the robots would still need to be *aware* of what constitutes an ethically-relevant fact. Take, for instance, the first rule governing the ethical behavior of robots:

> A robot may not injure a human being or, through inaction, allow a human being to come to harm.

In order for a robot to act according to this rule, it must be able to understand what "harm" means to a human, and what situations could possibly lead to a human coming to harm. In fact, a different understanding of what counts as "harm" could lead to many different interpretations to guide the robot's behavior[2]. Even being just a fictional example, this helps to highlight how, in order to exhibit ethical behavior and adapt to potentially unforeseen situations involving multiple agents, interests and contexts, an explicit awareness of what counts as "moral" is needed.

This requirement for explicitness, which would be needed in order to achieve explicit moral agents and functional morality (as well as beyond that), makes artificial morality a more promising avenue than ethical design, which is based on *a priori* anticipation to ensure that behavior is constrained *according* to certain ethical principles. Instead, artificial morality is rooted in the notion of "agent" and "agency"; automated decision-making systems here are not considered mere tools, but they are implicitly considered to be autonomous over certain decisions. As such, this approach is based on integrating moral reasoning into the decision system itself. The agent is given agency to identify, evaluate and potentially make autonomous decisions over potentially new ethically-relevant situations –just as we humans do.

The overall idea behind the engineering of these systems is simple: the "morality" of a decision should be identified, weighted and brought into the picture, just as it is already done with other notions (such as "utility", "performance", "benefit", etc.) that are factored into the decision procedure. Intuitively speaking, this sounds quite natural to what we humans do when we reason about a situation; sometimes, our decision is entirely based on the benefit we would receive from acting in a certain way; other times, we become aware of the moral weight involved in such decision and choose to act in a different way, even if it not as beneficial to us as it could be. Beyond this intuition, however, identifying, capturing and weighing morality in a computational way suddenly becomes a huge conceptual challenge where, for every answer, we are faced with a plethora of both theoretical and technical questions. As identified by [24], this challenge can be looked at from three main design perspectives:

1. *Top-down approaches* are based on understanding and defining beforehand all those situations that could be relevant in order to distill a set of rules to guide the behavior of the artificial moral agent (some examples are [6,7,21]).
2. *Bottom-up approaches* are inspired by trial-and-error learning which, in fact, we humans use while developing our moral character. Machine learning and

[2] In the story "Liar!" [5, ch. 6], precisely, a robot continuously lies to the characters in order to avoid hurting their feelings, which is an unintended understanding of the term "harm" that was not planned in the design of that robot.

evolutionary algorithms are some of the underlying mechanisms that could be used under this approach (see [1,15]).

3. *Hybrid approaches* combine both previous approaches in order to dynamically learn from relevant cases, while sticking to a certain set of rules that might constrain or guide the way those cases are processed. Hybrid approaches have the advantage of being more flexible than pure top-down approaches, while being less unpredictable than purely bottom-up ones (see [2,22]).

Despite the clear challenge behind the computational representation of something as contextual as "morality", several prototypes have been designed and implemented in order to explore this uncharted territory [10]; this sheds some light into this and provide some first steps that can be followed to enhance this kind of explicit ethical systems[3].

3 Exploring the Challenges Behind the Case Study

Although it is true that ethical dilemmas in the context of education need not be as deeply nuanced as the one we present in this paper, that case is useful to understand the multiple layers that may be involved in those scenarios. Needless to say, a case like that, where not even human teachers can agree on (different professionals propose very different approaches to it in [17]), would be extremely challenging to solve computationally. However, we can tentatively venture into exploring some of the many challenges that each layer of that case would need to be faced by autonomous ethical agents. Far from trying to provide a solution to that problem, this exploration can help us understand the challenges that a computational approach to it will face, thus guiding future steps in this line.

3.1 The Educational Layer

As it has been explained in Sect. 1.1, the Educational layer represents the most direct representation of the student's activities as part of the expected learning process. The information contained in this layer encapsulates the standard norm and conditions required to pass grade, as well as the student's actual results; furthermore, and through the student's record, one can get a picture of the student's performance in the past, and might allow to spot performance trends and unusual variances that could be used to support making a decision.

If we are to focus, for now, only on the Educational layer, the ethical dilemma depicted in the case study somewhat vanishes –or, at the very least, gets strongly diluted. Because this layer is driven by a clear rule (the required mark to pass

[3] It is worth mentioning that these two approaches to ethical systems, ethical design and artificial morality, are not mutually exclusive. In fact, Moor points out in his work how the categories he defines in [19] are not exclusive either –an explicit ethical agent can easily be an ethical impact agent and an implicit ethical agent as well. Following this, furnishing an agent with some artificial morality mechanisms does not imply having to ditch ethical design approaches beforehand.

grade) and a fact (the mark scored by the student), there is not much to consider at this point... unless we bring the student's record into the picture. A student whose marks are good enough, but which have shown improvement in the past (even if with ups and downs), might be able to keep the pace with the next grade; conversely, a student whose marks have been consistently low would not probably be able to cope with the next grade any better than with the current one.

This task at hand could be approached through ethical design by a set of rules, without the need to furnish the agent with any sort of explicit moral reasoning. Because the rule and the data are clear, the only thing that could be taken into account in this decision is the tendency depicted in the student's records. One can easily imagine an automated system that, provided the students' records follow a certain specified tendency, is more prone to either rounding up a slightly low mark[4] to allow a grade pass, or, at least, bring a human-in-the-loop to make a decision on a "fringe case".

3.2 The Personal Layer

The Personal layer poses some computational challenges with respect to the Educational layer. Namely, as the Personal layer is directly related to the student's beliefs, goals, intentions, etc., which are way harder to capture from "within" the computational setting that an artificial agent, such as a personal tutoring system, would have access to. This layer would normally be grasped and understood, in a traditional learning context, by the student's teachers[5].

This layer presents severe conceptual representation challenges. Because we are focusing on the personal attitude of the student towards the situation, we are faced with notions such as the student's *intention*, *effort*, *dedication* and *discouragement*, among others. While it can be argued that, in an online learning environment, dedication could be somewhat acknowledged (through maybe hours connected on campus, activities and exercises solved, etc.), the other three belong to the internal mental state of the student and are hard to account for through the usual environment in online learning. An even trickier part of this layer lies in the fact that this decision is all about potential future outcomes. This decision is no longer about what the normal rule about grade passing and

[4] This would then open up the Sorites question about "how low is low enough" for the system to make this decision, but this question falls outside the scope of this paper.

[5] It is worth recalling a recent case that occurred during 2020 in the UK in which, due to students being unable to attend an A-level exam due to the Covid-19 pandemic, an automated system was implemented in order to predict the student's grades [16]. It turned out that the predictions made by the students' teachers and the ones made by the automated system were quite different (being way lower in the automated prediction), which resulted in several protests that led to the UK government disregarding the automated predictions and following the human teachers' predicted grades. This ties up directly with the fact that human teachers had access to this Personal layer of their students that the automated system, which was fed only on data of what we call the Educational layer, lacked.

the student's grades are, but rather it is about how the future scenario could be, after a decision has been made, and how this may affect the student's possibilities.

In this sense, the student's success in raising her grades (which can be acknowledged by the system through her records) is key in considering whether the student should be given this chance; a student whose grades have been invariably low would not probably be able to cope with the next grade –or, at least, there is no evidence supporting that. However, and because this not guarantee, in any way, that the student will be able to keep the pace, this might require some sort of vote of confidence where a teacher might agree to allow the student to pass grade, but only if the student promises to make an effort to keep the pace with it; this agreement might work, or it might not, but it is something that only time will tell.

Automatizing this to some degree becomes challenging, as it risks turning the exception into a norm. In the case that occupies us, the fact that the student has shown that she can keep up with the workload, given enough effort and dedication, should be key in deciding whether the exception should be considered. An automated system, therefore, should be able to look up for signs in the student's record showing that they could, indeed, face the challenge a new grade would present[6].

3.3 The Social Layer

The Social layer requires an understanding that goes beyond the students and that deals with the context they are living in. This context has an influence over the student and their behavior, as well as over potential opportunities and limitations the student might have. Among other factors, family, social roles, historical and geographical inequalities can have a strong impact on the student and their learning process. Understanding how those factors can influence the student is a challenge for both human teachers and automated systems. Although demographic data of the students might be available, inferring information automatically from that data in order to understand the social context of the students can potentially lead to biases, profiling and discriminatory segmentation of the students, as it has been seen in other areas where ML technologies have been used on demographic data [13]. This layer, therefore, presents a particularly deep challenge: whereas the social context is indeed relevant and can provide important information regarding the students' living conditions, it is also mainly based on generalizations that can lead to unfair decisions based on social profiling.

Whereas the Personal layer had to do with the student's particular case and, therefore, was hard to generalize into an "exceptional norm", the Social layer deals with generalizations and, as such, is hard to apply to the particular case. Even though social circumstances can lead to systematic profiling and

[6] Learning analytics could help understand the student's performance and dedication and provide some grounds for a more informed decision.

discrimination, in our case study the Social layer is meant to highlight how, because of the inherent problems the student faces in her social environment, further care could be placed in her case in order to make an exception to the usual norm. In this case, therefore, the Social layer should be meant to play a "positive discriminatory" effect, rather than a negative one. Understanding how the student's environment might constrain her performance, her dedication and her emotional well-being can be a reason to provide an additional margin of tolerance to that student's case in order to make up for that.

However, and just as it happened in the aforementioned sectors where AI has systematically been discriminatory because of social considerations, the same thing could happen if the trend was inverted. Because social differences reflect *general* patterns, those patterns might not apply to the particular cases of people belonging to those social groups. Belonging to a particular social circle might indicate a propensity, or a major probability, towards being affected by some of the characteristics (positive, or negative) that are considered "common" of that social group. Nevertheless, transforming a general tendency, depicted by statistical patterns, into particular cases, always bear the risk to unfairly misrepresent those particulars. The characteristics that are more prone to occur in that social group could not, however, be systematically taken for granted, but they could rather be open for consideration.

The fact that the social environment is highly contextual and usually requires a holistic understanding of the social reality of the moment could suggest that a human-in-the-loop should be needed in order to understand each particular case. The role of a pedagogical agent in this case would be not to make a decision on its own, but rather to identify the case as potentially relevant and bring a human-in-the-loop in order to make a supported decision. This collaborative role that technology can have in ethical decision-making, where an automated system may be able to identify cases that require more careful human-in-the-loop supervision, leads us to the identification of the need for ethical sensors in this kind of systems, which we explain in the next section.

4 Awaking Awareness: From Physical to Ethical Sensors

Even with just a brief analysis of the guiding case study, it is enough to show the complexity that ethical dilemmas carry with them. Our reflections around this case lead us to realize that, before even attempting to imagine a computational system able to provide a satisfactory decision to a situation as complex as that, the system would need to raise an *awareness* of the ethical dimension behind that decision. As we have previously mentioned, the Educational layer in itself had almost no hint of an ethical dilemma at all. Nevertheless, by directing our attention to the Personal and Social layers of that particular student, the ethical dimension behind such decision arose. Only once we become aware of such dimension we start wondering whether following the standard norm would be the right thing to do, or whether we need to carefully consider other factors bearing an ethical weight before making a decision. Following this line, we recognize that our autonomous ethical agents need something else, before they can

even consider how to make an acceptable decision as part of ethical dilemmas: they need to be able to *recognize* them.

The notion of a "sensor" is quite common in computational technology. Physical sensors allow computers to take parts of the physical worlds as input and react to them accordingly: light, temperature, movement, writing, faces, etc. Physical sensors allow to leap the existing gap between certain non-digital parts of the world and digitalize them in order to be used in a computational environment. The sensory organs in the human body, in fact, do pretty much the same: our senses capture the "raw" information we perceive in the world (light, sound, touch, etc.) and send it to our brains so that we can make a model and interpret them. We can follow this analogy further on and go beyond the physical sensors in our body. Aside from being able to gather and interpret physical information about the world, we are also able to extract and infer other kinds of information from a perceived situation, such as psychological, emotional and ethical information. Just as we can perceive and "mindread" someone else's thoughts, worries and feelings, we can also intuitively perceive when a decision carries out an ethical weight with it. The ethical relevance of the situation is not clearly signaled by any physical, or mental signal; instead, when we put together our mental model of a situation we are faced with, we can (often) recognize whether it carries out an ethical component in it. What makes a particular situation qualify as ethically-relevant, however, greatly varies depending on the context. Regardless of the particularities behind each situation, the important thing is that, just as we have sensory inputs directed towards sensing the physical world, we are also able to "sense" the ethical dimension of a scenario. Following this analogy, then, it is only natural to ask ourselves the question: could artificial agents be equipped with a similar sort of *ethical sensors*, just like how they already are equipped with physical sensors?

We believe that, in order to design artificial moral agents able to recognize and deal with the ethical dimension of situations they may face, they should be equipped with some sort of ethical sensors. Although the parallelism with physical sensors is useful, ethical sensors would need to be, in fact, an internal procedure within the artificial agent, rather than an input hardware component. In other words, ethical sensors would need to interact not with the "outer world", but rather with the "inner model" of the outer world already created within the agent's representation system. These sensors should allow the agent to raise awareness about decisions that could have a profound ethical impact. Unlike physical sensors, which gather "actual" data about the world and construct a picture of the current scenario, ethical sensors would need to consider "potential" outcomes that could follow a decision and understand how these would shape future scenarios.

In this sense, we claim that artificial morality is key. Whereas ethical design allows to encapsulate and foresee specific situations that may arise in certain domains, it is not grounded on the compositional evaluation of the different "pieces" that form the situation, but rather on the consideration of the situation as a whole. Conversely, artificial morality can take a more granular approach

to ethically-relevant scenarios that might *arise*, or *emerge*. For instance, [1,2] provide hybrid and bottom-up prototypes of artificial moral agents that learn, via examples, how to deal with new morally-relevant scenarios. Even though these prototypes focus on solving moral dilemmas by following other examples and, therefore, by capturing their patterns, one can conceive these agents as being able to "distill" the ethical relevance of the different elements forming the distinct cases that are used to train them; in other words, the agents could grow an awareness about which elements can potentially bear ethical weight within them. This would be a first step towards being able to recognize when a certain scenario, even if brand new in the set of examples used to train the agent, might bear ethical relevance.

It is worth noting that the path between recognizing ethical dilemmas and solving them is far from direct. Nevertheless, if artificial agents are to be more and more integrated in our educational environments, then it we need to ensure that they can, at the very least, *recognize* when a decision can potentially have an ethical impact on students, or teachers. Furthermore, ethical dilemmas in education can be quite complex to evaluate and solve, and it might be challenging, or even impossible, the find the "right" choice –even among human teachers. In those cases, artificial agents deployed in an online learning environment (personal tutoring systems, pedagogical conversational agents, etc.) could, if equipped with ethical sensors, recognize when a decision could qualify as an ethical dilemma and bring a human-in-the-loop to supervise and consider how to deal with that particular case.

5 Conclusions and Future Work

Ethical dilemmas usually bear a huge degree of complexity with them and it is often not even clear whether a "right choice" exists. Making a decision usually involves adhering to certain principles while disregarding others, and this may bring severe detrimental consequences to some of the parties involved in the situation. In this sense, ethical dilemmas in education are no different. As artificial agents gain more autonomy and become more integrated in the educational system, one must consider how these artificial systems could be equipped to handle ethically-relevant situations that arise in online learning environments. In this work we have briefly compared two main approaches to ethical artificial agents: ethical design and artificial morality. Each one have their strengths and weaknesses and can, therefore, tackle different parts of the problem at hand.

Far from aiming to provide a general solution to the automated solving of ethical dilemmas, in this work we direct our attention to a challenge that would come prior to solving an ethical dilemma: recognizing that there is one. While ethical design preemptively constrains certain behaviors and functionalities that fit a pre-specified pattern, existing prototypes of systems equipped with some sort of artificial morality behave as "ethical problem solvers": their inputs are decisions *known* to be ethically-relevant, and so they are designed to munch those scenarios and spit out an ethically-acceptable solution. Although this is something that ethical autonomous agents will need to do eventually, these agents

would firstly need to be able to recognize ethically-relevant scenarios, prior to using their moral reasoning mechanisms to understand what the best choice would be –and, in case there is no clear answer, bring a human professional into the decision process. In this line, we introduce the notion of *ethical sensors* as mechanisms aimed not at solving ethical problems, but rather at identifying what can make a decision qualify as "ethically-relevant". Being better suited to the design of explicit ethical agents, we believe that artificial morality provides a better starting point towards the design of ethical sensors. The need to find a balance between ethical sensitivity and privacy remains open and needs further discussion, however, as in order to achieve more ethical sensitivity the system would probably need to have more data available.

As part of our future work, we want to study how artificial morality prototypes shaped in either a hybrid or a bottom-up fashion, such as [1,2,20], could be used as a starting point to the design of ethical sensors. Intuitively, the patterns learned by those systems could be used to identify what makes a morally-relevant situation be considered so. Once those elements have been identified, we could start prototyping ethical sensors that, given situations that are not preemptively identified as being morally-relevant, could identify whether those situations could carry moral relevance by following a compositional, or emergent analysis of the elements that form such situations. Although designing artificial pedagogical agents that are able to solve ethical dilemmas is, indeed, our ultimate goal, we first need to ensure that those agents will be able to recognize them.

References

1. Anderson, M., Anderson, S.: Geneth: a general ethical dilemma analyzer. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 28 (2014)
2. Anderson, M., Anderson, S.L., Armen, C.: Medethex: a prototype medical ethics advisor. In: AAAI, pp. 1759–1765 (2006)
3. Anderson, S.L.: Asimov's "three laws of robotics" and machine metaethics. Ai Soc. **22**(4), 477–493 (2008)
4. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. ProPublica, 23 May 2016 (2016)
5. Asimov, I.: I, robot. HarperCollins Publishers (2013)
6. Blass, J.: Interactive learning and analogical chaining for moral and commonsense reasoning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30 (2016)
7. Blass, J., Forbus, K.: Moral decision-making by analogy: Generalizations versus exemplars. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 29 (2015)
8. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017)
9. Casas-Roma, J., Conesa, J.: Towards the design of ethically-aware pedagogical conversational agents. In: Barolli, L., Takizawa, M., Yoshihisa, T., Amato, F., Ikeda, M. (eds.) 3PGCIC 2020. LNNS, vol. 158, pp. 188–198. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-61105-7_19

10. Cervantes, J.A., López, S., Rodríguez, L.F., Cervantes, S., Cervantes, F., Ramos, F.: Artificial moral agents: a survey of the current status. Sci. Eng. Ethics **26**(2), 501–532 (2020)
11. Clarke, R.: Asimov's laws of robotics: implications for information technology. Mach. Ethics 254–284 (2011)
12. Dede, C., Richards, J., Saxberg, B.: Learning Engineering for Online Education: Theoretical Contexts and Design-Based Examples. Routledge (2018)
13. Favaretto, M., De Clercq, E., Elger, B.S.: Big data and discrimination: perils, promises and solutions. A systematic review. J. Big Data **6**(1), 1–27 (2019)
14. Gunkel, D.J.: The Machine Question: Critical Perspectives on AI, Robots, and Ethics. MIT Press, Cambridge (2012)
15. Honarvar, A.R., Ghasem-Aghaee, N.: An artificial neural network approach for creating an ethical artificial agent. In: 2009 IEEE International Symposium on Computational Intelligence in Robotics and Automation-(CIRA), pp. 290–295. IEEE (2009)
16. Kolkman, D.: "f**k the algorithm"?: What the world can learn from the UK's a-level grading fiasco, August 2020. https://blogs.lse.ac.uk/impactofsocialsciences/2020/08/26/fk-the-algorithm-what-the-world-can-learn-from-the-uks-a-level-grading-fiasco/. Accessed 10 Feb 2021
17. Levinson, M., Fay, J.: Dilemmas of educational ethics: Cases and commentaries. Harvard Education Press (2019)
18. Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., Floridi, L.: The ethics of algorithms: mapping the debate. Big Data Soc. **3**(2) (2016)
19. Moor, J.: Four kinds of ethical robots. Philosophy Now **72**, 12–14 (2009)
20. Muntean, I., Howard, D.: Artificial moral agents: creative, autonomous, social. An approach based on evolutionary computation. In: Seibt, J., Hakli, R., Norskov, M. (eds.) Sociable Robots and the Future of Social Relations: Proceedings of Robo-Philosophy 2014, pp. 217–230. IOS Press (2014)
21. Vanderelst, D., Winfield, A.: An architecture for ethical robots inspired by the simulation theory of cognition. Cogn. Syst. Res. **48**, 56–66 (2018)
22. Wallach, W., Franklin, S., Allen, C.: A conceptual and computational model of moral decision making in human and artificial agents. Top. Cognit. Sci. **2**(3), 454–485 (2010)
23. Wallach, W., Allen, C.: Moral Machines: Teaching Robots Right from Wrong. Oxford University Press, Oxford (2008)
24. Wallach, W., Allen, C., Smit, I.: Machine morality: bottom-up and top-down approaches for modelling human moral faculties. AI Soc. **22**(4), 565–582 (2008)
25. Yapo, A., Weiss, J.: Ethical implications of bias in machine learning. In: Proceedings of the 51st Hawaii International Conference on System Sciences (2018)