

THREE KINDS OF AI ETHICS

Emanuele Ratti¹

Department of Philosophy, University of Bristol

Abstract. There is an overwhelming abundance of works in AI Ethics. This growth is chaotic because of how sudden it is, its volume, and its multidisciplinary nature. This makes difficult to keep track of debates, and to systematically characterize goals, research questions, methods, and expertise required by AI ethicists. In this article, I show that the relation between ‘AI’ and ‘ethics’ can be characterized in at least three ways, which correspond to three well-represented kinds of AI ethics: ethics and AI; ethics in AI; ethics of AI. I elucidate the features of these three kinds of AI Ethics, characterize their research questions, and identify the kind of expertise that each kind needs. I also show how certain criticisms to AI ethics are misplaced, as being done from the point of view of one kind of AI ethics, to another kind with different goals. All in all, this work sheds light on the nature of AI ethics, and sets the groundwork for more informed discussions about the scope, methods, and training of AI ethicists.

1. INTRODUCTION

Literature in AI Ethics has recently exploded. What characterizes this growth is not just how quick it was, but also how varied it is from a disciplinary perspective. Contributions to AI Ethics come from a number of directions, including philosophy, human-computer interaction, political theory, computer science, social sciences, and law, just to name a few. It is difficult to keep track of AI ethics trends, because key concepts (e.g., algorithmic bias; surveillance; explainability, trustworthy AI; AI safety) come and go pretty quickly. Because of this sudden and chaotic growth, AI Ethics has not yet had an opportunity to stabilize as a unique discipline with its own specific questions, methodologies, exemplars, and good practices.

This is unlike much older applied ethics disciplines like clinical ethics or research ethics, which count their own standardized trainings, professional figures, and even textbooks (Beauchamp and Childress 2009; Shamoo and Resnik 2015). AI Ethics is moving its first steps towards these achievements. For instance, there is the *AAIE*², which is an association created to promote the professional interests and development of AI ethicists around the world. There are occasionally summer schools aimed at providing crash courses on AI Ethics and its many facets. What is missing though is an attempt to systematize AI Ethics as a discipline, by identifying common themes, and grouping them on the basis of the questions asked, the expertise needed, and the context of implementation.

The goal of this article is exactly to fill this gap. AI Ethics lies at the conjunction of the terms ‘ethics’ and ‘AI’, and here I show how relationships between these two words can be conceived in rather different ways. In particular, I claim that there are at least three kinds of AI Ethics that are currently well-represented in the literature: Ethics *and* AI, Ethics *in* AI, and Ethics *of* AI. The benefits of this analysis are not merely taxonomical. Instead, what I will show is that these three kinds of AI Ethics presuppose different relations between the two terms ‘AI’ and ‘ethics’, and they ask different questions about the relevance of ethics to AI and *vice versa*. As a consequence, the goals of AI Ethicists will be rather different across these three kinds of AI Ethics, and their methodologies and scope will be as

¹ mnl.ratti@gmail.com

² [Home - Association of AI Ethicists](#)

well. Different trainings will be required as a result of the kind of AI ethics an institution or a company is interested in. Another notable consequence is that classic criticisms raised against the status of AI Ethics can be misplaced: they might be raised from the point of view of the goals of one kind of AI Ethics, to another kind of AI Ethics that might not have the same goal.

In order to show the different relations between ‘AI’ and ‘ethics’ underpinnings the three types of AI Ethics, I will rely on the Capability Approach (Sen 1999; Nussbaum 2011) as an illustrative example through which elucidating the general characteristics of each kind of AI Ethics. This is especially useful to show how one and the same normative framework can serve three different goals (requiring three different kinds of skillsets and expertise) within AI Ethics. This will be complemented with concrete examples from recent scholarship in AI Ethics. While I have not followed any specific methodology of systematic review for identifying the relevant literature, Supplementary Table 1³ provides a list of all the articles I have read, and how I have decided to categorize them on the basis of my tripartite account of AI Ethics.

The structure of the article is as follows. First, there are preliminary considerations to lay out. In Section 1.1, I will give a succinct description of the Capability Approach, and in 1.2 I will define important terms that will be used in the article. In Section 2, I will illustrate themes and problems from Ethics *and* AI. In Section 3 and 4, I will do the same for Ethics *in* AI, and Ethics *of* AI respectively. In Section 5, I will show how this taxonomy can provide a better focus on the limitations of AI ethics, and will discuss two possible limitations of my analysis.

1.1 The Capability Approach

The capability approach (CA) is an approach to compare quality-of-life assessments. It is conceived as an alternative to more consequentialist approaches to measure quality of life. CA is used to assess policies and social arrangements in a variety of contexts, across high-, middle-, and low-income countries. There are a number of formulations of CA (Robeyns 2005), emphasizing different disciplinary angles (such as economics, politics, social sciences, or philosophy) and it is known under many names (e.g., ‘human development approach’). I especially rely on Nussbaum’s formulation (2011).

The central point of CA is that individuals not only should have access to concrete positive resources for improving their quality of life, but they also should be able to choose which resources to use, how, and to what purpose. This consideration has a number of ramification, the most notable being that our assessment of policies or institutions should be concerned with what people are (freely) able to do and be. From these basic considerations, there is a fundamental distinction, which is the one between functionings and capabilities.

Functionings are things that one might value doing or being. Functionings include rather different things, from basic states such as being nourished, to complex activities such as participating to political demonstrations. Most consequentialist theories of well-being tend to focus only on functionings when measuring quality of life, though there is significant variation as to what counts as important functionings for measuring quality of life. CA proposes something different: we should consider not only functionings, but also the freedom that an individual has in deciding which functionings to pursue, how, and why.

³ [Three kinds of AI Ethics - Supplementary Table 1.xlsx - Google Sheets](#) . This table is regularly updated to add new papers or categorize papers I read in the past

These ‘freedoms’ are called ‘capabilities’, defined as a range of potential functionings that are concretely feasible for an individual to achieve, and that such individual can freely choose to pursue. The emphasis is on choosing freely, and as such capabilities are seen as *substantial freedoms*, where individuals “may or may not exercise [the freedom] in action; the choice is theirs” (Nussbaum 2011, p 18). These substantial freedoms have an ethical dimension, because (at least in Nussbaum’s view) they provide a foundation for human dignity.

There are other important aspects of capabilities, especially concerned to their structure. However, this succinct introduction is enough for the moment, and other aspects of CA will be specified when they become relevant for the discussion on the three kinds of AI ethics.

1.2 Terminological Caveats

Let me now turn to a few terminological caveats. In the rest of this article, AI systems will overlap significantly (but not completely) with Machine Learning (ML) systems.

I take AI systems to be constituted by two kinds of characteristics: functional and structural characteristics. This distinction comes mainly from philosophy of technology, where a debate stemmed from Kroes’ characterization of technical artifacts as having a ‘dual nature’ (2002). With this, he means that there are two ways of looking at technical artifacts, one *qua* physical systems with certain *structural characteristics*, and one from the point of view of the *goals* they contribute to or task they fulfil. In line with this crude characterization of the rather complex thesis of Kroes, I take ‘functional characteristics’ of AI systems as a specification of their goals and tasks, and the structural characteristics as the ones referring to the way ML systems are constituted, both in terms of components, and the procedures followed to coordinate those components⁴. This means that AI systems are ML systems plus something else that goes beyond the boundaries of ML systems (Figure 1).

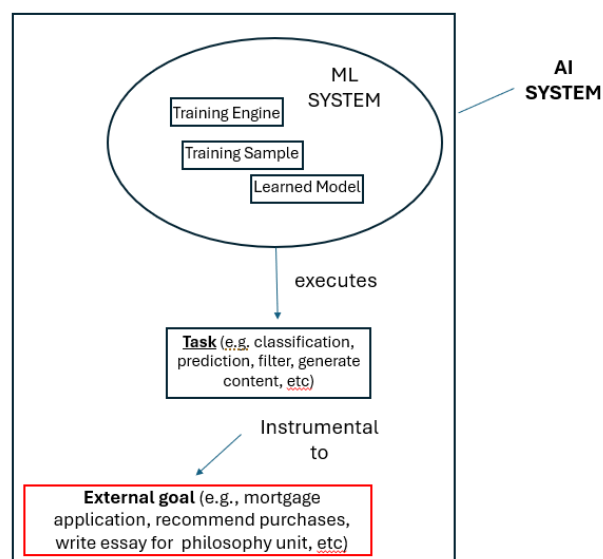


Figure 1. Functional and structural characteristics of AI systems

⁴ Kroes’ account stresses the importance of the ‘material’ nature of structural components of technical artifacts. In the case of AI systems, ‘structural components’ are either virtual (e.g. data sets, algorithms, technical requirements), material (e.g. the hardware that implements the software on which AI is implemented in turn) or hybrids. In this context, I focus especially on ‘virtual components’, for reasons that will become clear later on

Structurally, I see AI systems as constituted at least by three components which make up ML systems (Termine et al 2024). First, there is the training sample, which is the repository of data points that the system use to learn and adapt. Second, there is training engine, which is the computational/optimization ‘machine’ that the system uses to learn and adapt on the basis of the training sample. Finally, there is the learned model, which is what the ML system ‘learns’ and it is applied to execute a number of tasks. I include in the category ‘structural characteristics’ also the procedures and standards followed by AI practitioners to coordinate these components.

There can be a number of functional characteristics or attributes, which falls into two categories (Figure 1). First, there are ‘task-attributes’. AI systems can execute tasks such as classification, prediction, content filtering, content generation, etc – task attributes are the ‘bare’ outputs of an AI system. But these tasks are often instrumental to a goal, which is generally structurally external to ML systems, such as approving mortgage application, recommend purchases, etc. Goals are the ‘effects’ in the real-world to which bare outputs contribute to. I include also goal-attributes (or ‘contribution to goals’) as a functional characteristic⁵. Sometimes, functional characteristics (or their failures) are connected to macroproperties of AI systems that might refer to structural components of AI systems, but the technical dimension is overlooked or ignored, as I will show.

Let me now turn to the three kinds of AI ethics.

2. ETHICS AND AI

The first kind of AI Ethics is what I call Ethics *and* AI. I define it as follows:

Ethics and AI (EaA) = the study of the relation between the functional characteristics of AI systems and pre-existing normative commitments external to AI systems

Central to the definition of EaA are the functional characteristics of AI systems. In the context of EaA, AI systems are looked especially from this point of view. For instance, AI systems can be used to automate mortgage applications, but in doing so they might exclude individuals on unlawful or ethically problematic basis. In this case the problem is with the ‘task attribute’, namely the ‘bare output’ of the AI system. But in other cases it is the goal associated to the AI system that is problematic. The EU AI Act (Madiaga 2024) prohibits AI systems whose goals are associated with unacceptable risks (such as AI systems used for face recognition tasks).

The notion of ‘unacceptable task or goal’ has to be understood from the standpoint of specific normative frameworks (i.e., fundamental human rights). To use an expression from Kaplan’s work, EaA is concerned with questions of ethics that are *external* to AI systems, and analogously to what he calls ‘philosophy and technology’, the goal of EaA is to “analyze technology in terms of ready-made philosophical concepts, usually moral and political concepts such as ‘freedom’, ‘general welfare’, and ‘human nature’” (2009, p xiv).

From the perspective of EaA, AI systems are significantly blackboxed from a structural point of view. With this I mean that the attention to the structural characteristics of AI systems is *minimal*. While many examples of EaA stem from considering the opaque nature of ML systems, or the fact that

⁵ One could specify more precisely the distinction between tasks and goals by resorting to the distinction between effect role functions and purpose role functions (van Eck 2015). However, adapting this distinction to this context will require more work than it seems, and given the limitation of space I plan to do this in another work

data can be biased, the specific technical facets of these problems are not of particular interests, nor are technical solutions. Rather, what is of central interest to EaA is whether certain outputs or goals of the systems can possibly collide with specific normative commitments in our society. EaA ethicists' goal is to point out that there is a tension, suggest that the gap between the AI systems and the normative frameworks ought to be filled, provide in some cases a general and high-level strategy on how to do this (e.g. design publicity of Loi et al 2021), but then how to concretely do this is a task require another kind of skillset and a different set of considerations. This means that EaA ethicists need not considerable technical expertise about AI systems; rather, what is important is that outputs and goals of AI systems may raise some issues of, e.g., privacy or fairness. This is certainly not an easy task, especially when the effort is on showing why certain tasks (e.g. predicting behavior; filtering and ranking information, etc) coupled with macroproperties of AI systems (e.g. opacity, biased data, automated processes) pose specific normative problems, as in the case of showing why trust is indeed an important issue when AI is concerned (von Eschenbach 2021), that automated decision-making systems violate duties of consideration (Grant et al 2023), or that we have indeed a right to an explanation that opacity violates (Vredenburg 2022). But none of these analyses require any specific technical expertise concerning opacity or algorithmic bias.

A classic example of normative commitments characterizing EaA is the principlist approach characterising a significant portion of the AI Ethics literature (Jobin et al 2019). The strategy is to take general principles that are seen as constituting the (ethical, political, and civic) backbone of our society, and see to what extent the goals and tasks of AI systems put them in jeopardy. Literature in EaA has taken inspiration from principlist biomedical ethics, often explicitly (Mittelstadt 2019). This move has been misunderstood in various ways - I will come back to the misunderstandings in Section 5. For the time being, consider that the moves licensed by principlism in AI Ethics are not based on dubious similarities between AI and biomedicine as a profession (as correctly criticized by Mittelstadt 2019). Rather, they are based (or they can be based) on the idea that the principles of biomedical ethics are universal principles, at least in Beauchamp and Childress' perspective (2009). In particular, the source of the four principles of biomedical ethics is what they call 'common morality', which "is applicable to all persons in all places, and all human conduct is rightly judged by its standards" (Beauchamp 2007, p 7). Developing AI systems is just one instance of 'human conduct', and hence it has to be judged by the standards of those principles. This means that the 'principles' of EaA are conceptualisations of normative concerns and requirements that, at least from this perspective, stand for general desiderata of our society, and that everyone will likely accept (at least from the principlist perspective). The only genuine AI-based principle seems to be explainability, but it is not difficult to show that explainability can be reduced to a combination of other more general principles, such as justice and respect for autonomy (REF), or the importance of trust (von Eschenbach 2021), or in general that 'explaining' can be reconceptualized as 'justifying' AI systems in the face of the normatively-laden goals (Loi et al 2021). In principlist EaA, the goal of AI ethicists is to gauge the feasibility and appropriateness of AI systems and the goals they serve within a given context. AI ethicists will orient discussions about those normative commitments in the context of AI systems, to make sure that high-level principles are fruitfully interpreted and/or developed.

Let me illustrate how EaA would work with CA. In 2019, a roadmap for supporting a more equitable development of AI was published by the United Nations (2019), in line with a plan for achieving the so-called Sustainable Development Goals. This was part of a larger effort by United Nations to address the emerging political and ethical issues raised by frontier technologies. AI here is not considered a unique technology: the main issue is what risks cutting edge technologies like AI raise for Sustainable Development Goals in general. What emerged from this document is that AI systems should be evaluated on the basis of how well they align with already pre-existing developmental goals

and human rights, and that whatever AI system one wishes to develop, it should “balance economic, social, and environmental goals” (p 3), where those goals are understood on the basis of the categories provided by the United Nations’ frameworks. In line with the Human Development Approach (namely, CA) endorsed by United Nations, the emphasis and the focus is especially on low- and middle-income countries, which are typically at the mercy of high-income countries when emerging technologies (and an equitable share of burdens and benefits) are concerned. More specifically, four distinct layers of ‘capacity development’ are identified for AI, from infrastructure, to data, human capabilities, and human rights-based laws and policies. All those typical issues raised by a CA-based approach to policy-making are present here: issues of digital divide, conversion factors related to capability-expansion, and focus on human rights. Therefore, the goal of CA-based EaA, as envisioned by United Nations, is to direct all AI-related policies towards those concerns typically raised by CA. EaA will then shape the AI community by directing the attention of their practitioners towards those CA-based concerns

There are several examples of EaA, and most of them are admittedly linked to principlism. Char et al (2020) is a textbook example: it looks at the classic ethical issues raised in biomedicine, and then adds ‘AI’ to the healthcare context to see how this complicates the situation – which it does. A rising trend in EaA is to pay attention to ‘safety’, which can be redefined in terms of ‘non-maleficence’ (Gyevnar and Kasirzadeh 2025). But there are also examples of ‘principlism-free’ EaA. For instance, Waelen (2022) redefines ethical issues typically framed in the principlist way in terms of *power*. The central concern is about emancipation and empowerment, and because there is evidence that goals associated to AI systems might jeopardize them, then AI Ethics should be framed in terms related to critical theory.

3. ETHICS IN AI

A second kind of AI Ethics that is well-represented in the literature is what I call Ethics *in* AI, which is defined as follows:

Ethics in AI (EiA)= the study of how AI systems, given their design flexibility, can be constructed such that their structural characteristics reflect given ethical and/or political commitments

In line with what I said in Section 2, by ‘structural characteristics’ I mean the internal characteristics and components of ML systems, in particular the training engine, the training sample, and the model which is constructed as a result of the training of the engine on a given data set (Termine et al 2024). Each of these components is constructed and coordinated by data scientists/AI practitioners by following certain procedures on the basis of benchmarks and standards. I include those procedures and their standards in my definition of ‘structural characteristics’.

EiA is based on the idea that ethical, societal, and political issues raised by AI systems are often failures of design, in the sense of failures to pick up and coordinate the right structural characteristics. AI systems generate outputs that are unfair, violate privacy, or jeopardize safety, because they have been designed to generate those outputs. If this is true, then it is possible to design AI systems such that they will deliver the right ethical outcomes, and this can be done by modifying the structural characteristics of AI systems themselves.

These ideas have emerged in an engineering context that sees ethics as a matter of ‘techno-fixes’. Ethics is not about a community effort to orient functional characteristics of AI systems towards the right normative desiderata. Rather, ethical problems are just one set of problems that could be addressed with an engineering mindset. As noticed by Wiggins and Jones (2023), most of ‘techno-fixes’

have focused especially on privacy and fairness, leading to a proliferation of attempts to ‘code’ fixes into AI systems – e.g., k-anonymity, differential privacy, independence, separation, sufficiency, etc. Especially when it comes to fairness, there has been also an explosion of workshops and meetings dedicated explicitly to the construction of novel tech fixes (e.g. FAcCT).

Unlike EaA where AI systems are black-boxed, in EiA AI systems are indeed opened up, and their structural characteristics modified to make sure that they deliver the right results. Special attention is especially devoted to output metrics and data preparation, which is where concerns of privacy and fairness are more likely to emerge. This emphasis is structural rather than functional: it is not necessarily about the outputs of algorithms; rather, it is about what kind of metrics we use for measuring the ethical relevance of outputs, and whether one metric rather than another reflects our moral and political commitments. This has been explicitly raised in the debate between ProPublica and Northpointe/Equivant on the alleged discriminatory nature of COMPAS, where one side accused the other of using the wrong notion of fairness to inform choices regarding how to measure ‘discrimination’ (Ratti and Russo, 2024).

The contribution of ethics as a discipline to EiA is to provide the right conceptual resources related to those moral and political commitments that the structural characteristics of AI systems ought to reflect. In the literature on techno-fixes, this has been done especially by data scientists or computer scientists: an EiA ethicist is often a well-rounded computer scientist or data scientist who takes normative concerns as something internal to, or part and parcel of designing AI systems. Representative of these efforts is the rich literature on *value alignment*. A classic of this approach is Gabriel’s work (2020), which considers the different philosophically-informed standpoints through which aligning AI to human beings’ normative commitments or, briefly put, to human values. Moreover, Gabriel reviews the main challenges for this endeavour, from formally encoding values or principles in AI systems, to choosing the right principles on the basis of resources provided by philosophy, social sciences, and political theory. But computer scientists or data scientists need not be AI ethicists. In fact, philosophers, social scientists, and political theorists have been collaborating with computer and data scientists exactly to contribute to the process of choosing the right ethical and political conceptual tools, as well as implementing them. For instance, in a seminal article Binns (2018) shows how works in moral and political philosophy can inform emerging (at the time) debates about fairness in ML.

It is useful to illustrate an example of EiA that uses CA as a resource to design AI systems that deliver the correct outputs. London and Heidari (2024) are concerned that most approaches to value alignment are too centrally focused on the values stemming from the cultural background of creators (that is, high-income countries), and that they are not emphasizing larger impacts of AI. The importance of interacting with AI systems in such a way that AI systems are meaningfully beneficial to individuals, they say, cannot be adequately prioritized in current AI alignment strategies. Doing this requires reconceptualizing AI systems as *assistive technologies* – which means redefining their functional characteristics (that is, the goals and functions that AI systems serve). But in order to do this, it is necessary to lay out precise structural characteristics of AI systems as well, which they do in terms of a formal characterization of Nussbaum and Sen’s CA. This is because, in their opinion, redefining the structural characteristics on the basis of CA is the only way to build AI systems that are indeed assistive technologies. This is a convincing case of EiA: AI systems are conceptualized and structurally characterized through a theoretical framework provided by economics and philosophy.

4. ETHICS OF AI

The third kind of AI Ethics is what I call *Ethics of AI*. This is defined as follows:

Ethics of AI (EoA) = the study of how AI systems and the communal practices of the contexts in which they are implemented, shape each other

Central to EoA is the idea of ‘communal practice’. This term has its origin in the context of the contemporary revival of virtue ethics, especially the one inspired by MacIntyre’s work (2011). What is exactly a ‘practice’, in this particular context? MacIntyre defines a ‘practice’ as

“any coherent and complex form of socially established cooperative human activity through which goods internal to that form of activity are realized in the course of trying to achieve those standards of excellences which are appropriate to (...) that form of activity, with the result that (...) human conceptions of the ends and goods involved, are systematically extended” (p 218)

There is a lot to unpack in this definition. ‘Complex’ refers to the idea that communal practices can be described at a number of different levels, while the collaborative nature of communal practices refers to the fact that they require individuals to negotiate their goods, standards, and procedures. Most important, communal practices are sustained activities that are goal-oriented, and as such requires certain normative conceptions of what count as a good for that practice, why, and which are the legitimate means to get to those ends.

How is ‘communal practice’ connected to ethics? Common examples of communal practices – science, medicine, theatre, education, chess, football, etc – show the importance of a collective effort in which individuals might compete with each other, but also work together “to advance shared goods seen as having intrinsic values” (Hicks and Stapleford 2016, p 454). As such, it is possible to see communal practices as those spaces where individuals, in collaboration with other individuals, act and make choices that are instrumental to pursue goals that are considered intrinsically valuable and, in some cases, are relevant to live the kind of life that individuals have reasons to value. This is especially relevant in communal practices with an essential aspirational character, such as science (Ratti and Stapleford 2021), or education. Being these spaces where individuals pursue their life aspirations, communal practices are tightly connected to implicit or explicit conceptions of the Good Life. But ethics is, by definition, about choices and actions as they unfold with respect to conceptions of how we ought to live. Therefore, the structure and the dynamics of communal practices are ethically salient, because they will have an effect on questions about agency and choices which are relevant for the Good Life.

But what is the relation to AI ethics? Given the connection between communal practices and ethics, AI systems can be said to ‘mediate’ the Good Life, because they ‘mediate’ our experiences in the environments and the contexts in which communal practices unfold. Here I take ‘mediation’ in the technical way this term has been understood in postphenomenological philosophy of technology (Ihde 1990; Verbeek 2004). At a very basic level, AI systems ‘mediate’ as any other technical artifacts do, namely by shaping our perceptual and interpretative abilities⁶. The way AI systems do this is typically by shaping, constraining, and transforming the same environment in which we act and make choices (Danaher 2016). While this applies to other technical artifacts, the scale (Creel and Hellman, 2022) and the invisibility (Moor 1985) of AI systems seem unprecedented.

The way AI systems shape our environment – and hence possibly our communal practice - come from how structural and functional characteristics ‘react’ to the characteristics of the context of implementation, and this can happen in rather unpredictable ways. For instance, recommender

⁶ There is a lot to say about this, but for reasons of space, consider Ihde’s (1990) and Verbeek’s (2004) works

systems, by filtering content, might shape the online environment in unpredictable ways, such as insulating users from exposure to different viewpoints (Milano and Prunkl 2024), thereby depriving them of alternative conceptions of goods, ends, and life aspirations, with noteworthy consequences for a number of communal practices. But it is not just the fact that an AI system filters content that makes echo-chambers inevitable. This happens because an AI system filters content within a particular context in which users interact with virtual environments in specific ways (e.g., little exposure to the news or educational content outside the virtual environment itself) such that it then results in echo-chambers impacting also communal practices. The point is that it is difficult to anticipate effects on communal practices on the basis of functional and structural characteristics of AI systems alone – we need also to consider the context in which AI systems are implemented. EoA is interested in discerning the relation between functional and structural characteristics of AI systems on the one hand, and the context in which AI systems are implemented on the other, in order to uncover possible effects on communal practices. Because of this interest, EoA is a genuine sociotechnical approach (Fazelpour and Danks 2021): it focuses on structural and functional characteristics of AI systems (i.e., the ‘technical’) and the characteristics of the context in which such systems are employed (i.e., the ‘socio’).

The goal of AI ethicists in the context of EoA follows from the above considerations. Given an AI system x , with functional and structural characteristics $c1, c2, \dots, c3$ to be implemented in an environment y with features $f1, f2, \dots, fn$, the goal of an AI ethicist is to analyze how cs might negatively shape communal practices in y , where this analysis is based on a thorough investigation of cs and fs . This analysis is typically carried out from the standpoint of a framework coming from a number of disciplines, be they moral philosophy, political theory, social sciences, etc.

Let me now turn this picture into something more concrete by illustrating EoA through CA. In the case of EoA, CA is used as a method to make sense of the relation between AI systems and communal practices (Ratti and Graves 2025). As I have explained in 1.1, capabilities are functionings that are concretely achievable by individuals, and that individuals can freely choose to pursue. Nussbaum (2011, pp 33-34) identifies ten central capabilities, which are deemed relevant for human dignity, and the task of a government is to ensure that all citizens possess at least a threshold of these capabilities. Because capabilities are about choices, actions, and agency, if AI systems mediate actions by restructuring the environment where communal practices unfold, then they shape capabilities. But this does not say much: it is just a different formulation of the basic idea described above, namely that AI systems shapes our choices (i.e., capabilities) by structuring the context where communal practices happen.

CA becomes helpful only when we pay more attention to the structure of capabilities, which Nussbaum calls *combined capabilities*. In addition to capabilities proper, combined capabilities include what are called *conversion factors*. Whether a functioning is effectively achievable – whether it is indeed a capability – will depend on factors that allow a person to turn (to convert) a possibility into something actual. Therefore, whether one can make a choice that matters to them, will depend especially on these conversion factors. There are several types of conversion factors, including personal (e.g. reading skills, physical condition, metabolism, etc), social (e.g., public policies, social norms, societal hierarchies, etc), environmental (e.g. climate, geographical location), digital (e.g. access to computers, phone network, broadband). In order to understand the connection between capabilities and conversion factors, take a classic example of the CA literature. Bicycles are technical artifacts that can potentially expand a number of central capabilities. For instance, the capability of affiliation can be greatly expanded as a result of, let’s say, young kids using bicycles to join a sport club in a nearby town. Senses, imagination, and thought can be also expanded as a result of individuals biking beyond large urban areas and engage in sightseeing. However, the functional and structural characteristics of

bicycles *alone* cannot guarantee that capabilities will be expanded, nor we can predict that they will do only on their basis. In order for bicycles to do so, several conditions must apply: one needs to have the right physical factors enabling the use of bicycles; a certain infrastructure facilitating the movement of bicycles must exist, etc. For instance, a bicycle in the Netherlands (with its extensive bike path) will expand capabilities better than in the Amazon Forest. In other words, bicycles can convert possibilities into actual functionings, only if some personal, social, and environmental factors are already in place (that is, conversion factors).

AI systems are no exceptions. Consider a famous case in medical AI discussed a few years ago. Obermeyer et al (2019) investigate the performance of a widely used health-risk algorithm. The goal of this AI system can be conceptualized as expanding health as a capability (namely, pursuing health-related goals that one can freely choose), where health is a necessary condition for engaging in communal practices in general⁷. Obermeyer et al found that this system falsely concluded that Black patients were healthier than equally sick White patients, though the algorithm appeared to be well calibrated across races. What emerged from their analysis was that the AI system used 'health expenditure' as a proxy for health risk. It is not unreasonable to think about 'health expenditure' as a proxy for risk: the more one spends on health, the more this person might have health-related problems. However, 'health expenditure' is a proxy only for those individuals who already have access to healthcare, where 'health care access' is a conversion factor (Prah Ruger 2010) that depends in turn on other conversion factors, such as having a full-time job; health insurance; a stable income; living in an area where healthcare is accessible; etc. In other words, the AI system was performing well only for those individuals whose 'communal practices' were characterized by a number of factors of personal, social, and environmental nature. AI practitioners have assumed that all end-users had a homogeneous level of conversion factors; however, with this controversial assumption, they automatically exclude and make invisible all those who lack that particular level of conversion factors. This means that structural characteristics of AI systems (i.e. the ones determining 'health care access' functioning as a proxy for health risk) have interacted with the characteristics of the context of communal practices (i.e. the lack of certain personal, social, and environmental factors characterizing the context of certain AI stakeholders), in such a way that the AI system would negatively affect the participation of individuals to certain communal practices by excluding them from the important radar of healthcare assessment risk. But it is difficult to anticipate this unfortunate outcome on the basis of structural and functional characteristics of AI systems *alone*. What the CA suggests instead is that one ought to look at the AI system *and* the context of implementation, by identifying the relations between functional and structural characteristics on the one hand, and conversion factors on the other.

There are a number of interesting works in EoA in the literature. Some works are devoted to make broader points. For instance, Heuser et al (2025) argue in favour of a more comprehensive AI ethics, which also includes an analysis of how AI systems and the lifeworld interact sometimes with surprising outcomes, where the idea of lifeworld is connected to ethics in a way comparable to how I have done it above. Other contributions take a particular perspective on the relation between AI systems and communal practices from a more explicit normative standpoint. For instance, Longo (2025) describes how algorithmic systems embedded in social media mediate political judgement. Most important, he provides an analysis of how AI do not 'replace' or 'determine' judgement as others have argued; rather, AI systems mediate (in the classic postphenomenological sense) judgement. This

⁷ In fact, some think (Nussbaum 2011; Venkatapuram 2011) that 'health' is not a proper capability; rather, it is a functioning that can be characterized as a necessary condition for all other capabilities to be expanded. The same can be said for communal practices: without health, it is difficult to engage in any meaningful communal practice

analysis is done from the standpoint of Arendt’s normative views (in the sense that Arendt’s perspective is helpful for illuminating these dynamics).

5. DISCUSSION

To summarize, Figure 2 represents the relation between ‘ethics’ and ‘AI’ in the three kinds of AI Ethics. In the case of EaA, AI systems and ‘ethics’ are not overlapping, but there is a line connecting them, suggesting the importance of the community effort to bring AI and ‘ethics’ closer by making functional characteristics of AI systems compatible with ‘ethics’. In the case of EiA, ‘ethics’ is literally inside AI systems: it is conceptualized as part and parcel of the structural characteristics of AI systems. Finally, in EoA, because AI systems are seen from the standpoint of a specific normative framework, they are ‘contained’ in the ‘ethics’.

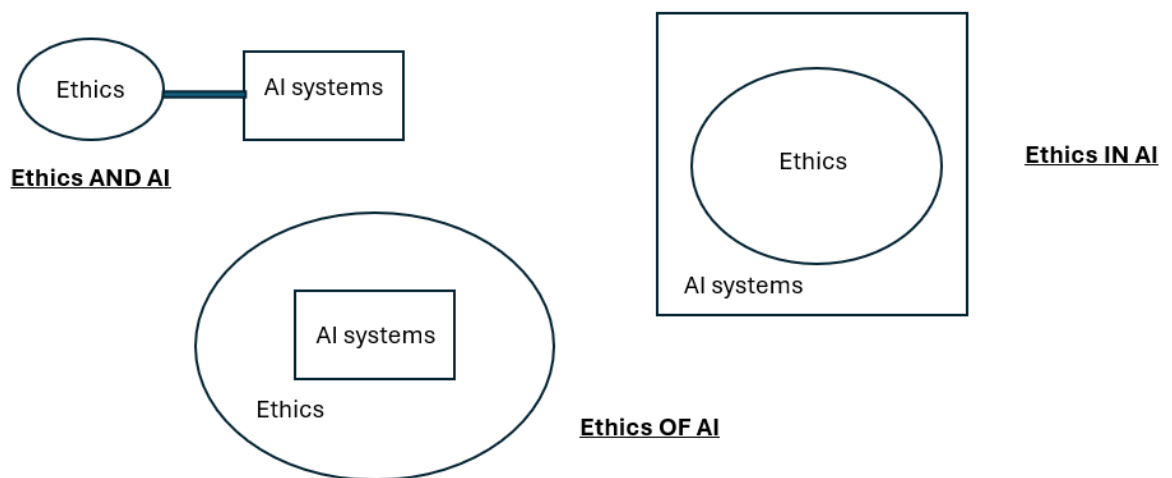


Figure 2. Relationships between ‘ethics’ and ‘AI systems’

This characterization of AI Ethics raises a number of questions concerning its benefits, limitations, and prospects.

An advantage of this conceptualizing is that it reveals some limitations of classic criticisms against AI ethics. Consider the harsh criticisms raised against the principlist approach in EaA. For instance, Munn (2023) provides a comprehensive account of flaws of principlist AI Ethics. Principles are deemed meaningless, in the sense of “highly abstract and ambiguous, becoming incoherent” (p 870), and a gap with practical implementation has been also considered as a fatal flaw for this kind of AI ethics (Morley et al 2020). Other problems include that lack of consequences resulting for not complying with principles. These are important concerns, but it is important to realize that, from a principlist and EaA perspective, some of them are simply misplaced. Consider the early days of the principlist approach in applied ethics. Arguably, its origin is the *Belmont Report* in 1978 (Wiggins and Jones 2023), which set up principles and guidelines for research on human subjects after years of discussion. The report conceptualized ethics as a negotiation of tensions concerning means and ends, on the basis of the acceptance of three principles that were seen as ‘epistemic backstop’, “the consensus on which all parties can agree, even when disagreeing about specific applications” (Wiggins and Jones 2023, p 239). Principles in the Report are explicitly ‘comprehensive’ and general enough to cover present and future normative concerns. But then the question is how can these general (and meaningless, toothless, remote from practice, to use Munn’s terminology) principles apply? First, we

need to be clear on what we should expect from principles: they are not rules, as rules trump further reasoning, while principles allow for flexibility (Ratti and Graves 2021). This means that, unlike rules, you do not follow principles: you *weigh* them one to the other. Therefore, if we say that we cannot apply principles because they do not provide enough information to be followed, then we are simply ascribing to principles a task that should be assigned to rules. Second, we better think about ‘principles’ in applied ethics as akin to what is prescribed by constitutions, where these provide orientation of a general nature, and it is then up to communities to formulate more specific standards for individual cases. We should not expect principles to do all the work, as rules do; rather, it is an individual community “which does the hard work to distill these principles into standards, rules, and therefore into practice” (Wiggins and Jones 2023, p 240). Therefore, some of the criticisms of the dominant form of EaA (i.e. principle-based) simply miss what principlism is up to. And most important, this idea that normative commitments like principles are just orienting design and implementation of AI systems within a specific context, can be used to anticipate similar objections of limited applicability that can be raised to other forms of EaA: the goal of AI ethicists, as shown above, is to facilitate discussion leading to the formation of the backbone of these communities, and it is up to these communities to distil the general normative commitments into standards and benchmarks. This can suggest that harsh criticisms against EaA have been made from the perspective of EiA: if a proposal in EiA had the same characteristics of a principlist-based EaA, then criticisms made by, e.g., Munn, would be on target. However, principlist-based AI Ethics is often EaA, so those criticisms are unfair. Similarly, it has been sometimes said that off-the-shelf tools provided by EiA tend to uncritically assume certain underlying normative concepts, where choice between them is value-laden and requires “philosophical arguments and considerations that fall outside of the narrow technical scope of the standard approach to fair ML” (Fazelpour and Danks 2021, p 10). But one can argue that this is more the role that EaA ethicists should play, rather than EiA ethicists.

These last considerations raise the question about the relation between the three kinds of AI ethics. Are these approaches in competition one with the other? Are they complementary? Can one be an AI ethicist in all three senses? I do not have a definitive answer to these questions. At first glance, there is indeed a relation of continuity between EaA and EiA. Because AI ethicists belonging to the first kind orient discussions around AI systems towards relevant normative commitments that are seen as valuable by a given society, then it is reasonable to conclude that they are also moving the first steps towards the process of implementing specific moral and political concepts in AI systems (namely, EiA). In fact, one can see an EaA ethicist playing a role also as an EiA ethicist, by enumerating to computer and data scientists the impressive variety of conceptual tools that can be possibly formalized. But the expertise of a full-blown EiA ethicist goes beyond what is required to EaA ethicists: not only one has to be proficient in the relevant normative vocabulary, but also they have to know the context in which AI systems are constructed and negotiated. This point should be explicitly stressed. A significant portion of EiA takes place in private companies. We are all well aware of the highly controversial outcomes of implementing AI ethics in private companies, such as the elimination of the Google’s Ethical AI Team which led to the firings of key EiA ethicists such as Timnit Gebru⁸. As documented by Metclaf et al (2019), it is not just a matter of finding the ‘right ethics’ to implement in AI systems. To succeed, EiA ethicists need to persuade members of the organization of the importance of the nature of the constraints that EiA can put on AI systems themselves. As Wiggins and Jones point out, “it is unclear how to convince colleagues to value ethical principles [that] could serve as any constraint – particularly if those constraints would reduce profit” (p 245). This is to say that knowledge and experience of corporate and management dynamics, which is not required to EaA ethicists, might be

⁸ <https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/>

required to EiA ethicists. Similar considerations apply to EoA. Ethicists in this kind of AI Ethics might be indeed useful in the context of EiA, but the limitations provided by corporate structures can in turn limit the scope of EoA. Moreover, while the goal of EiA is indeed practical (i.e. design better AI systems), a successful EoA can just be limited to the analyses it provides.

Another important point to raise concerns what I have left out from my taxonomy. There is indeed one promising trend in the literature, which I am just not sure whether it is robust enough to have its own category. There are a number of articles (Grosz et al 2019; Bezuidenhout and Ratti 2021; McLennan et al 2022) doing exciting research on how to teach AI ethics to engineering students, researchers, or practitioners. Particularly famous is the so-called ‘embedded ethics’ approach (Grosz et al 2019), launched by Harvard University. This approach to teaching AI ethics is based on the idea that systematic exposure to ethical problems as they arise in technical contexts, will habituate students or practitioners to anticipate ethical pitfalls, or simply will create a ‘feeling’ for ethics, that is now missing. As shown in Grosz et al’s foundational article (2019), in embedded ethics students learn about the ethical implications of AI “while they are learning ways to develop and implement algorithms” (p 56). As a consequence, ethics modules are not about exposing students or practitioners to moral concepts or theories *in abstracto*, as it is often done in traditional units in moral or political philosophy taught in technical curricula; rather, students will be exposed to morally charged situations in the context of technical units that they are already attending. While the number of embedded ethics programs around the world is growing, there is not much writing about it. There are a few attempts at theorizing (Bezuidenhout and Ratti 2021; Ferdman and Ratti 2024), as well as more methodological articles on how to measure impact (Kopec et al 2023). However, much has to be theorized and conceptualized about embedded ethics, and this is why I decided not to create an *ad hoc* kind of AI Ethics for it.

Finally, the reader may have noticed that in Supplementary Table 1 there is a column for ‘Academic Ethics and AI’. I think about this as a variety of EaA, because in this literature ethical questions about AI systems are external to AI systems themselves, as in traditional EaA. However, the angle is slightly different. In Academic EaA, the focus is not on how the relationship between AI systems functional characteristics and external normative frameworks. Rather, Academic EaA takes AI as an interesting case study to develop an already existing discussion in an academic debate. In Academic EaA, it is shown that AI can shed light on the debate on a certain concept (or not, contrary to expectations or previous works). For instance, in Schuster and Lazar (2025), judicious attention allocation is the main topic, and AI is treated as an interesting illustration of the normative problems it raises. In other words, there is a pattern of moral problems related to attention allocation that philosophers and social scientists have been discussing for a while, and AI follows (and possibly adds to) this pattern. AI raises problems of attention allocation because of the goals that AI systems are typically built for (e.g. nudging users).

6. CONCLUSION

In this article, I have provided a structured analysis of AI ethics as a discipline. I have distinguished three different senses in which AI ethics has been understood so far, highlighting the research questions, role for AI ethicists, problems and prospects for these three kinds of AI ethics. I have also shown the possible relations between the three kinds, and I have highlighted a number of interesting trends in the literature that, in the future, might enrich my analysis. All in all, this article is important to understand that AI ethics is not one thing; rather, it is many things, and there is currently no AI ethicist that can reasonably claim to be able to cover all three sets of questions and methods. This

article will hopefully orient more informed discussions on what the role of AI ethicists is, their limits, the expertise needed, and their training.

REFERENCES

- Beauchamp, T., & Childress, J. (2009). *Principles of Biomedical Ethics* (Sixth). Oxford University Press.
- Bezuidenhout, L., & Ratti, E. (2021). What does it mean to embed ethics in data science? An integrative approach based on microethics and virtues. *AI and Society*, 36(3), 939–953. <https://doi.org/10.1007/s00146-020-01112-w>
- Char, D. S., Abràmoff, M. D., & Feudtner, C. (2020). Identifying Ethical Considerations for Machine Learning Healthcare Applications. *American Journal of Bioethics*, 20(11), 7–17. <https://doi.org/10.1080/15265161.2020.1819469>
- Creel, K., & Hellman, D. (2022). The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems. *Canadian Journal of Philosophy*. <https://doi.org/10.1017/can.2022.3>
- Danaher, J. (2016). The Threat of Algocracy: Reality, Resistance and Accommodation. *Philosophy and Technology*, 29(3), 245–268. <https://doi.org/10.1007/s13347-015-0211-1>
- Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, 16(8). <https://doi.org/10.1111/phc3.12760>
- Ferdman, A., & Ratti, E. (2024). What Do We Teach to Engineering Students: Embedded Ethics, Morality, and Politics. *Science and Engineering Ethics*, 30(1), 7. <https://doi.org/10.1007/s11948-024-00469-1>
- Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3), 411–437. <https://doi.org/10.1007/s11023-020-09539-2>
- Grant, D. G., Behrends, J., & Basl, J. (2023). What we owe to decision-subjects: beyond transparency and explanation in automated decision-making. *Philosophical Studies*. <https://doi.org/10.1007/s11098-023-02013-6>
- Grosz, B. J., Grant, D. G., Vredenburg, K., Behrends, J., Hu, L., Simmons, A., & Waldo, J. (2019). Embedded EthICS: Integrating Ethics Broadly Across Computer Science Education. *Communications of the ACM*, 62(8), 54–61. <http://arxiv.org/abs/1808.05686>
- Heuser, S., Steil, J., & Salloch, S. (2025). AI Ethics beyond Principles: Strengthening the Life-world Perspective. *Science and Engineering Ethics*, 31(1). <https://doi.org/10.1007/s11948-025-00530-7>
- Hicks, D. J., & Stapleford, T. A. (2016). The Virtues of Scientific Practice: MacIntyre, Virtue Ethics, and the Historiography of Science. *Isis*, 107(3), 449–472. <https://doi.org/10.1086/688346>
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- Kaplan, D. (2009). *Readings in Philosophy of Technology*. Rowman & Littlefield Publishers.

- Kopec, M., Magnani, M., Ricks, V., Torosyan, R., Basl, J., Miklaucic, N., Muzny, F., Sandler, R., Wilson, C., Wisniewski-Jensen, A., Lundgren, C., Baylon, R., Mills, K., & Wells, M. (2023). The effectiveness of embedded values analysis modules in Computer Science education: An empirical study. *Big Data and Society*, *10*(1). <https://doi.org/10.1177/20539517231176230>
- Kroes, P. (2002). Design methodology and the nature of technical artefacts. *Design Studies*, *23*.
- Loi, M., Ferrario, A., & Viganò, E. (2021). Transparency as design publicity: explaining and justifying inscrutable algorithms. *Ethics and Information Technology*. <https://doi.org/10.1007/s10676-020-09564-w>
- London, A. J., & Heidari, H. (2024). Beneficent Intelligence: A Capability Approach to Modeling Benefit, Assistance, and Associated Moral Failures Through AI Systems. *Minds and Machines*, *34*(4), 41. <https://doi.org/10.1007/s11023-024-09696-8>
- Longo, A. (2025). Algorithmically mediated judgment: an arendtian perspective on political subjectivity in social media. *AI and Society*. <https://doi.org/10.1007/s00146-025-02230-z>
- MacIntyre, A. (2011). *After Virtues*. Bloomsbury.
- Madiega, T. (2024). The EU AI Act.
- Metcalf, J., Moss, E., & boy d. (2019). Owing ethics: Corporate logics, Silicon Valley, and the institutionalization of ethics. *Social Research An International Quarterly*, *82*(2), 449–476.
- Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 501–507.
- Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From What to How. An Overview of AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics*, *0123456789*. <https://doi.org/10.1007/s11948-019-00165-5>
- Munn, L. (2023). The uselessness of AI ethics. *AI and Ethics*, *3*(3), 869–877. <https://doi.org/10.1007/s43681-022-00209-w>
- Nussbaum, M. (2011). *Creating Capabilities - The Human Development Approach*. Harvard University Press.
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. <https://www.science.org>
- Ratti, E. (2020). “Models of” and “models for”: On the relation between mechanistic models and experimental strategies in molecular biology. *British Journal for the Philosophy of Science*.
- Ratti, E., & Graves, M. (2021). Cultivating Moral Attention: a Virtue-Oriented Approach to Responsible Data Science in Healthcare. *Philosophy and Technology*, *34*(4), 1819–1846. <https://doi.org/10.1007/s13347-021-00490-3>
- Ratti, E., & Graves, M. (2025). A Capability Approach to AI Ethics. *American Philosophical Quarterly*.
- Robeyns, I. (2005). The Capability Approach: a theoretical survey. *Journal of Human Development*, *6*(1), 93–117. <https://doi.org/10.1080/146498805200034266>
- Ruger, J. P. (2010). Health capability: Conceptualization and operationalization. *American Journal of Public Health*, *100*(1), 41–49. <https://doi.org/10.2105/AJPH.2008.143651>

- Schuster, N., & Lazar, S. (2024). Attention, moral skill, and algorithmic recommendation. *Philosophical Studies*. <https://doi.org/10.1007/s11098-023-02083-6>
- Sen, A. (1999). *Development as Freedom*. Anchor Books.
- Termine, A., Ratti, E., & Facchini, A. (2024). *Machine Learning and Theory-Ladenness: a Phenomenological Account*, <https://arxiv.org/abs/2409.11277>
- United Nations. (2019). *Summary of deliberations Addendum A United Nations system-wide strategic approach and road map for supporting capacity development on artificial intelligence*.
- van Eck, D. (2015). Mechanistic explanation in engineering science. *European Journal for Philosophy of Science*, 5(3), 349–375. <https://doi.org/10.1007/s13194-015-0111-3>
- von Eschenbach, W. J. (2021). Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy and Technology*. <https://doi.org/10.1007/s13347-021-00477-0>
- Vredenburg, K. (2022). The Right to Explanation. *Journal of Political Philosophy*, 30(2), 209–229. <https://doi.org/10.1111/jopp.12262>
- Waelen, R. (2022). Why AI Ethics Is a Critical Theory. *Philosophy and Technology*, 35(1). <https://doi.org/10.1007/s13347-022-00507-5>
- Wiggins, C., & Jones, M. (2023). *How Data Happened: A History from the Age of Reason to the Age of Algorithms*. WW Norton.