Models of Agency in AI Ethics Education

Marco Peris[1], Teresa Scantamburlo[1,2,*]

[1]University of Trieste, via Economo 12/3, 34123 Trieste
[2]European Centre for Living Technology, Ca' Bottacin, Dorsoduro 3911, Calle Crosera, 30123 Venice, Italy

Abstract

This paper presents two models of agency that offer useful frameworks for understanding both machine and human action. We explore these models by examining their philosophical foundations, relevant literature, and implications for AI ethics. We argue that engaging with these concepts in AI ethics education can help clarify the respective strengths and limitations of human and artificial agents, thereby supporting the development of more effective oversight strategies. The paper concludes with preliminary recommendations for expanding the topics, competencies, and overall pedagogical approach in AI ethics education.

Keywords

AI ethics education, Human agency, Human oversight, Volitional agency, Mechanistic agency

1. Introduction

Rapid and pervasive advances of Artificial Intelligence (AI) innovation across various fields and human activities have sparked fresh discussions about the aims and methods of AI education. A central problem concerns the basic knowledge and competencies that all AI users should develop to ensure effective and beneficial use of AI, especially by those who may lack technical expertise [1, 2]. This led many organizations design and promote AI literacy initiatives for various targets, from children, teens to adults, e.g. see [3, 4, 5]. Similar efforts are also encouraged by the recent European AI Act which recommends AI providers and deployers to ensure that staff and users operating AI systems on their behalf have an adequate level of AI literacy, appropriate to their roles, background, and the context of use. (see art. 4 [6]). Building on this, the European AI Office has begun collecting examples of AI literacy initiatives in a publicly accessible living repository [7] to support the sharing of best practices and promote widespread adoption.

Another set of issues regards the education and training of AI scientists and practitioners. Past cases of AI-related harms, as well as studies in AI ethics and Science, Technology, and Society (STS) scholarship, have made it clear that the AI community should reconsider AI education programs with a view to integrating traditional technical knowledge and skills with ethical and social understanding, responsibility, and interdisciplinary perspectives. An example of such an effort is the Embedded EthiCS initiative at Harvard, which integrates ethical reasoning directly into computer science courses through interdisciplinary collaboration between philosophy and computer science faculty and students [8]. However, despite such promising approaches, a scoping review reveals that significant disparities remain in the coverage of AI ethics requirements across educational programs, with topics like Privacy and Data Governance more commonly addressed, while areas such as Societal and Environmental Well-being and Accountability are much less frequently included [9].

Integrating ethics into standard AI programs is not straightforward. A fundamental challenge lies in the fact that some ethical principles cannot be fully operationalized or translated into technical requirements. A key example is the notion of human agency and oversight, which often involves context-sensitive judgment and cannot be reduced to fixed rules or automated processes.

2nd Workshop on Education for Artificial Intelligence (edu4AI 2025 https://edu4ai.di.unito.it/), co-located with the 28th European Conference on Artificial Intelligence (ECAI 2025). October 26, 2025 in Bologna, Italy
*Corresponding author.
marco.peris@ds.units.it (M. Peris); teresa.scantamburlo@units.it (T. Scantamburlo)
0000-0002-3769-8874 (T. Scantamburlo)

In this paper, we argue that training future generations of AI developers requires expanding how human agency is conceptualized in standard AI literature [10]. Specifically, we suggest adopting a broader account of agency, one that enables exploration of the strengths and limitations of artificial agency while emphasizing the role of the human agent. To this aim, we compare two models of agency, show their philosophical roots (utilitarianism and virtue ethics), and discuss their different impact on ethical decisions and behavior. Our discussion will suggest implications for AI ethics education, in particular, recommendations for AI ethics courses dealing with human agency and oversight.

The paper is structured as follows. Section 2 outlines the importance of human agency and oversight in current AI ethics discourse and education. Section 3 presents the rational-agent model found in standard AI literature, examining its philosophical roots and ethical implications. Section 4 contrasts this with a volitional account of agency, associated with full ethical agency and grounded in classical philosophical traditions, which emphasizes internal motivations and moral responsibility. Section 5 explores the implications of these contrasting models for AI ethics education, offering preliminary recommendations regarding relevant topics, essential competencies, and pedagogical approaches. The paper concludes with a summary of our argument, highlighting the importance of incorporating a richer understanding of human agency into the training of future AI professionals.

Our contributions include a 1) a critical analysis of alternative models of moral agency expanding current discussions in the AI ethics literature; 2) suggestions for broadening the scope of human agency and oversight in AI ethics education.

2. Human agency and oversight in AI ethics education

Human agency and oversight is a central theme in AI ethics scholarship. Emphasized across various AI ethics policies [11], this requirement aims to protect human autonomy and foster fundamental rights [12]. It underscores the importance of designing AI systems that support users in making informed decisions and achieving their goals. This implies that such systems should be transparent and understandable, rather than functioning as inscrutable "black boxes." AI should be developed in a way that allows for meaningful human supervision throughout its operation - e.g. see strategies such as Human-in-the-Loop (HITL), Human-on-the-Loop (HOTL), or Human-in-Command (HIC) [12]. In high-risk contexts especially, this demands tighter control over AI, ensuring that humans retain the authority to choose whether to deploy a system and the ability to intervene in its decision-making processes.

There is ongoing debate about whether and how the requirement for human agency and oversight can be put into practice effectively [13, 14, 15]. Common ways to meet this requirement include monitoring system performance, setting up protocols for human intervention, and creating mechanisms to detect and respond to problems. The European AI Act also points out that carrying out human oversight requires proper training and the development of relevant competences [6]. However, what such training should entail remains open to discussion, and current AI ethics education offerings vary widely in both content and approach. Some take the form of standalone AI ethics courses, while others aim to integrate ethical considerations into technical AI curricula.

A review of AI ethics syllabi in the U.S. highlights that standalone courses often address broader social consequences, with some also examining the role and responsibility of humans as both developers and users. In contrast, technical AI courses tend to focus narrowly on topics such as bias, fairness, or privacy—typically framed as mathematical concepts and covered near the end of the course [16]. Another paper [17], which examined German data science programs, finds that in the field of AI education, ethics can be classified mostly from two perspectives: (1) "ethics as content" or (2) "ethics as a tool." The former perspective emphasizes moral and philosophical foundations, including moral theories, ethical principles and the discussion of moral dilemmas. The latter focuses more on practical aspects such as bias mitigation, transparency, explainability, data protection and legal compliance. This approach rarely engages with theoretical moral frameworks and is typically confined to introductory section within data science courses.

While human agency and oversight are closely linked to the theme of responsibility, there is little indication of how these concepts are addressed in AI ethics education, or what specific aspects and underlying assumptions are emphasized in teaching them. This reflects broader challenges in integrating ethical and technical content in computing courses - that is, how to link "ethics as tool" and "ethics as content" -, particularly the risk that students may struggle to connect ethical issues with technical topics. The disconnect between computing disciplines and ethics could be due to various factors, such as the perceived lack of relevance of ethics to technical subject matters and instructors' limited training in ethics [18]. This situation may also stem from the difficulty of engaging with alternative epistemologies or critical perspectives, which reflects deeper disciplinary assumptions that privilege technical rationality over humanistic or social ways of knowing [19].

Within this context, it remains unclear to what extent the concept of human agency, central to many AI ethics frameworks, is meaningfully addressed in AI ethics education. Few studies explicitly examine whether students are encouraged to reflect on their own role and responsibility in the development and use of AI systems. In addition, AI curricula often adopt a functional view of agency, such as rational agency [20], which tends to go unexplored in relation to other philosophical or existential perspectives. While this approach is well-suited for modeling agent's behavior in computational terms, it provides a limited account of the human dimensions involved in moral reasoning and ethical decision-making, both of which are essential for meaningful human oversight and supervision in AI systems. To address this gap, we compare two conceptualizations of human agency: one commonly found in AI, and another grounded in classical philosophical thought.

3. Agency in machines

In the context of AI research the concept of agency is closely tied to the problem of AI definition. The standard view springs from a widely used framework dividing AI definitions in four main categories: i) systems that think humanly, ii) systems that act humanly, iii) systems that think rationally, and iv) systems that act rationally [10]. According to Russell and Norvig, the last category—systems that act rationally—best captures the essence of the AI field. Under this view, any AI system can, in principle, be considered a rational agent: an entity capable of pursuing goals and interacting with its environment (whether physical or digital) under conditions of uncertainty.

Consider common applications that are part of everyday life: anti-spam software, security cameras, advanced safety systems in cars, and smart assistants on smartphones and computers. These systems are all designed to achieve specific goals while minimizing errors. The concept of AI as rational agents has become increasingly established with the rise of Large Language Models (LLMs), which are now integrated into both task-specific applications, such as those in chemistry [21] and medicine [22], and more complex workflows [23].

Interestingly, many intuitive definitions of AI encapsulate the rational-agent perspective. A popular example includes the EU AI ethics guidelines which refers to AI as those "systems designed by humans that, given a complex goal, act in the physical or digital world by perceiving their environment, interpreting the collected structured or unstructured data, reasoning on the knowledge derived from this data and deciding the best action(s) to take (according to predefined parameters) to achieve the given goal." [24].

The rational-agent approach offers several advantages compared to other definitions. It is more flexible than other definitions (e.g. *systems that think rationally*) and, most importantly, its mathematical framing offers a precise and widely applicable standard of rationality beyond mimicking human thought or behavior. Despite its apparent simplicity, the rational agent model can be used to describe a wide range of behaviors, from simple devices like thermostats to complex systems such as crowd-based and social machines [25]. Before discussing the implications of this model, let us first consider what is meant by an "agent" and what kind of "rationality" this paradigm embodies.

3.1. The definition of a "rational agent"

Following Russell and Norvig, we define an "agent" as any entity — physical or virtual — that can perceive its environment through sensors and act upon it through actuators. It is important to note that the environment is not the same for all agents; it depends on the type of input and output sensors the entity possesses. For instance, an agent designed to send spam operates within a completely different environment than a security camera: one is purely virtual, while the other has both physical and virtual dimensions.

In the literature, agents are also classified based on the types of programs that govern their interaction with the environment. These include simple reactive agents, model-based agents, goal-based agents, and learning agents [10]. Most modern AI systems fall into the latter two categories: they are goal-oriented and capable of learning. Designing effective software for an artificial agent involves considering the components summarized by the acronym PEAS: Performance measure, Environment, Actuators, and Sensors. A rational agent, when attempting to complete a task, should select the action that is expected to maximize its performance measure, based on its knowledge of the environment as acquired through its perceptual history (also called "experience") up to that point [10].Therefore, an agent's rationality lies in its ability to determine the most appropriate action or course of actions to take.

Historically, this notion of rationality is rooted in the seminal work of J. von Neumann and O. Morgenstern, who modeled human behavior to predict decisions in game-like settings [26]. According to their framework, a rational agent chooses between options based on expected utility. Two observations are important here. First, the concept of *utility*: in economics, utility refers to a numerical representation of a user's satisfaction (or reward) derived from selecting one option among several (e.g., choosing between different routes to minimize travel time). This notion naturally aligns with the idea of a performance measure—typically, higher utility (or reward) corresponds to better performance. The second observation concerns the uncertainty under which the agent operates. The agent chooses among possible outcomes, each associated with a certain probability and utility. As a result, we typically refer to *expected utility* (rather than utility alone), since the agent makes probabilistic predictions about the consequences of its actions and selects the one with the highest expected reward.[1]

This discussion leads us to restate the earlier definition of an AI system in more precise terms: a goal-oriented entity that, given the current state of its environment, selects the action that maximizes its expected utility.

3.2. The moral implications of a rational-agent view

The model of rational agency reflects the moral theory of Utilitarianism, and more broadly, the perspective of Consequentialism [2]. To determine whether one action is better than another, we focus primarily on its consequences and aim for the best possible outcome, one that maximizes the reward or the expected utility.

Moral aggregation. One issue with this approach is that the moral value of the consequences is not considered individually, but only in terms of their total sum. For example, when deciding which action is better, we might include some morally bad consequences in the overall calculation, yet still reach a positive outcome. This raises concerns about the reduction of individual moral worth, as some scholars have argued that utilitarian reasoning tends to treat people as a collective whole, overlooking the fact that each person lives—and is responsible for—their own life [28].

A classic illustration of this issue is the trolley problem [29], where a runaway tram is heading down a track, and a person must decide whether to pull a lever to divert the tram—saving five people but

[1] Specifically, it selects the option that maximizes utility by weighing the value of each possible outcome against the probability of its occurrence.

[2] It is important to observe that the moral theories of Utilitarianism and Consequentialism are not to be intended as interchangeable, but the latter is a branch of the former. The term "consequentialism" was used for the first time by an English philosopher, G.E.M Anscombe, in her article "Modern Moral Philosophy"[27] in 1958.

causing the death of one—or do nothing and allow the tram to kill the five[3]. This thought experiment has informed the design of the Moral Machine project, an online platform hosted by MIT that collected over 40 million decisions from participants across 233 countries to capture how people prioritize lives in autonomous driving scenarios [30].

Utility maximization. Moreover, the consequentialist perspective does not take into account what precedes the consequences — the process that leads to choosing one action over another — but focuses solely on achieving the most desirable outcome (i.e. the highest utility). From this standpoint, it becomes possible to construct computational models that behave in ways similar to humans, primarily by replicating the kinds of outputs typically produced by human agents.

A recent study describes similar consequentialist assumptions about AI in terms of a mechanistic view of agency [31]. According to this view, humans are regarded as moral agents whose ethical decision-making can be understood by analyzing how they typically respond to hypothetical moral dilemmas (similar to the trolley problem and the moral machine experiment) — that is, by weighing potential outcomes against specific inputs. Within this framework, it is assumed that as AI systems gain access to more data, their moral capabilities increase accordingly, enabling them to make better decisions. Note that this perspective underpins several AI research efforts aimed at evaluating LLMs in ethical decision-making contexts (e.g. see [32, 33]).

The consequentialist view of (moral) agency—and, accordingly, its mechanistic interpretation—offers a theoretical foundation that aligns with the needs of AI designers to translate actions into quantitative terms. However, the dominance of this perspective in AI research and practice risks reducing human agency to a mechanistic model and overlooking the qualitative dimensions of human action.

4. Moral Agency: From machines to humans

So far, we have examined how agency is conceptualized within the field of AI. We now turn to a broader framework of agency that can help us connect and reflect on the relationship between machine agency and human action. In particular, we draw on Moor's framework, developed to explore the theoretical foundations of machine ethics [34]. While we do not engage directly with his central research question ("Could we ever teach robots right from wrong?"), Moor's classification of artificial agents offers a useful structure for developing our discussion and has become a classic reference in the literature. According to Moor, ethical agents might be of four types:

1. *The ethical impact agent.* In the first category, the ethical agent is understood in a "weaker sense," meaning that any robot could potentially be considered an ethical agent if its actions have ethical consequences. For example, even a thermostat might qualify if it contributes to a sense of well-being in the home. Moor also identifies unethical agents—those whose actions result in negative consequences—which he refers to as "bad" agents.
2. *The implicit ethical agent.* In the second category, ethical considerations are embedded in the agent's software by its developers. A typical example is a safety or security system, such as an aircraft warning system (e.g., for altitude or collision alerts). Moor describes these agents as having a kind of built-in virtue, not derived from habit, as in humans, but from programming. An example of an unethical agent in this category is a spam zombie agent, which operates with harmful intent coded into its behavior.
3. *The explicit ethical agent.* Explicit ethical agents are "agents that can identify and process ethical information [...] and make sensitive determinations about what should be done" [34]. These agents act not only "according to" ethical principles, but also "from them." Unlike the previous types, which are passive with respect to moral values, explicit ethical agents can reason about ethical considerations and resolve situations where principles may conflict. Their actions have

[3]This is, of course, a difficult moral dilemma, but it helps highlight what factors should be considered when choosing one action over another.

ethical significance due to this explicit relationship with moral reasoning. An example of such an agent might be a large language model–based system used in healthcare, which analyzes patient data and ethical frameworks to reason through dilemmas and offer justified recommendations.

4. *The full ethical agent.* This final category consists of agents that possess certain "metaphysical characteristics" [34, p. 12], such as consciousness, intentionality, and free will. Based on these criteria, humans represent the primary - and currently only - example of full ethical agents.

In the following subsection, we take a closer look at the third and fourth categories, with the third encompassing the most advanced AI applications and the fourth reserved for humans. We emphasize that our aim is to examine the relationship between artificial and human agency in order to support human oversight and supervision. Rather than speculating on whether modern AI systems could be considered full ethical agents - now or in the near future - we focus on arguments that help highlight the distinct nature of human moral agency, which may be overlooked in standard models of rational agency. These reflections contribute to expanding relevant topics in current AI ethics curricula and further developing ongoing research discussions on human agency and oversight.

4.1. Explicit and full ethical agents

Building on prior analyses of Moor's classification [35], we point out that a fundamental difference between explicit ethical agents and full ethical agents lies in the nature of their relationship with (moral) values. A explicit ethical agent is only capable of acting *in accordance with* moral values — that is, acting procedurally based on values that have been assigned to it and that we expect to be followed during the machine's operation. Even when machine learning techniques are employed, AI systems do not determine values of their own; rather, they reconstruct—through data—what holds moral significance for humans.

In contrast, full ethical agents (i.e., humans) are capable not only of acting *in accordance with* moral values but also of freely determining those values. They maintain a direct relationship with values, one that is not mediated, as it is in the case of explicit ethical agents. As Fossa notes, human moral experience is defined by "the effort to live by affirming what we care about and opposing what we find unacceptable," a process that is "inseparable from questioning what it is that we truly care about and what we deem unacceptable" [35, p. 99]. When we act, we (as humans) choose to follow a course of action aligned with a value we recognize and affirm as such. This dynamic does not apply to artificial moral agents, for whom the selection of moral values is carried out in advance by humans — typically by relevant stakeholders — prior to deployment.

What we want to emphasize is that human moral experience cannot be understood in purely mechanistic terms, as if it were simply a predictable system that produces specific decisions (outputs) in response to external stimuli (inputs). Human desires and motivations are dynamic components that shape moral action and personal identity: they do not merely aim at achieving a given outcome, but rather question how that outcome should be pursued, in accordance with the agent's character and moral disposition.

The distinct character of full moral agency can also be understood through the lens of "volitional agency," a concept recently discussed within AI research to highlight the distinction from the mechanistic view discussed above [31]. The notion of volitional agency emphasizes that what characterizes human agency is the orientation toward what the agent wants to achieve, in other words, intention or volition. This perspective can be traced back to Aristotle, who emphasized the agent's internal disposition over the external consequences of action, and was further elaborated by Thomas Aquinas[4].

Thomas Aquinas offers a systematic analysis of human action grounded in the dynamic interplay between intellect and will. Unlike the mechanistic model, Aquinas's framework emphasizes a volitional structure in which action arises not merely from external inputs but from the agent's internal orientation toward the good [5]. The focus on the "volitional" aspect is crucial to critically address issues in AI such

[4]See Thomas Aquinas, Summa Theologiae (I-IIae, qq. 6–17).

[5]In the Thomistic tradition, the human faculty directed toward the Good in this way is the will or volition.

as responsibility, agency and human oversight, because the attention is not only on the the results of the action itself, but on what pre-consitutes it, such as motives, reasons, intentions and the link between means and ends.

Aquinas identifies a sequence of interrelated acts - ranging from the apprehension and volition of the end, to deliberation, choice, and execution - demonstrating that moral agency involves both rational judgment and affective commitment [36]. Central to this view is the notion of the will as a rational appetite: not an automatic response mechanism, but a faculty capable of freely choosing contingent goods in light of a universal good. This rich account helps illuminate what is at stake when contrasting human volitional agency with the procedural logic of artificial systems.

4.2. The moral implications of full ethical agency

Drawing on the components of full ethical agency, as articulated in the volitional account, we aim to highlight key human dimensions that carry important implications for the (educational) formation of AI professionals.

Freedom and self-examination. As humans, we act with the intention of achieving something, but not in a predetermined or necessary way. We retain the ability to stop at any point and reassess our goals if circumstances change. We are thus free both in choosing the ends we pursue and in deciding how to pursue them. Our decision-making process remains open and revisable throughout. As Dai notes, citing Taylor, human agents are the source of original purposes, whereas AI agents operate on derived purposes—goals "imposed" from outside.

This difference can be illustrated with a metaphor: a human agent is like an archer, while an AI agent is like the arrow. The archer can aim the arrow (the AI system) toward a target (its goal), but cannot control every detail of its trajectory. This "trajectory" highlights the inherent opacity of AI decision-making processes and underscores the need for human oversight when evaluating AI outputs.

Of course, human decision-making is also opaque to observers: we cannot directly access another person's internal motivations—only the agent themselves has insight into their reasons (at least the conscious ones). However, a human can engage in a kind of moral self-auditing: a reflective process in which one re-evaluates the action they are about to take, for example, by considering alternative moral constraints or verifying whether the chosen action continues to meet the standards initially set [37].

This reflective capacity allows the person to account for their actions, both to themselves and to others. In this way, the decision becomes an act of moral responsibility, satisfying the ethical criteria of responsibility and accountability [37].

By contrast, an AI agent remains inert without an external input, such as a predefined goal. When a human chooses consciously to do nothing, however, this too is a morally significant act - whether good or bad - resulting from deliberation rather than passivity.

Strong vs. weak evaluation. In full ethical agency, the evaluation of both ends and means can occur on two distinct levels. We may engage in a weak evaluation [31], focused on outcomes, to assess the most efficient or effective way to achieve a goal. Alternatively, we can perform a strong evaluation, which involves judging the quality of our choices in light of our intentions and values. Of the many ways a result can be achieved, some paths may more faithfully express the agent's underlying motivations and moral commitments. As Dai emphasizes, quality in this context does not refer to utility or efficiency, but to unquantifiable attributes of the agent's motivation [31, p. 6].

This capacity for strong evaluation relates to how much good a particular choice embodies, reflecting both the agent's intention and the values they uphold. Crucially, this form of evaluation is distinctive to human beings, for the reasons outlined earlier—particularly our ability to determine moral values freely and to question them at any moment.

To illustrate this, consider a person who chooses to act in accordance with deeply held values, even when the likelihood of success is minimal. For instance, someone may jump into the sea to try to save a drowning child, fully aware that the chances of survival are very low for both. Although they could

reasonably choose to stay on shore to preserve their own life, they nevertheless act on what they believe to be right, despite the risks. This is not a weak evaluation, driven by efficiency or expected outcomes, but a strong evaluation, focused on the moral quality of the act and the integrity of the agent's values.

Affective dimension. A volitional account of agency shifts the focus from externally observable actions to the agent, whose morality is defined by internal judgments and states [31, p. 7]. An action performed by a person is considered ethical not only in terms of the end it aims to achieve, but also in terms of how that end is pursued, according to the agent's motivation and the moral quality of their choices, as illustrated in the previous example.

Whereas mechanistic agency operates primarily within a performance-functional paradigm, volitional agency incorporates an affective dimension that cannot be reduced to quantitative terms. We describe this principle as affective because moral agency concerns quality, which in philosophical terms is closely associated with the good. This concept of the good is not reducible to something merely measurable or empirical; rather, it refers to a qualitative dimension of moral life that transcends calculation and instrumental reasoning.

Accepting this view of agency implies recognizing that different agents may reach the same moral end through different, equally valid paths, shaped by distinct moral reasoning, values, or principles. As Dai notes, this framework allows for, and respects the existence of genuine moral disagreement across populations, avoiding the reductive treatment of moral deviation as statistical noise. This stands in contrast to mechanistic models, which approach moral decision-making as something that can be captured through quantitative analysis and simulated algorithmically. While such systems may be able to infer patterns of average moral behavior from large datasets, this is not necessarily suitable, or morally appropriate, for every context or situation, especially in high-stakes domains such as healthcare.

5. Implications for AI ethics education

What can we take from this discussion of human agency, understood from both machine and human perspectives? In what follows, we offer a preliminary reflection with recommendations concerning the *topics*, *competencies*, and *overall approach* that should inform AI ethics education—particularly with regard to human agency and oversight.

Topics Our analysis suggests the need for a broader account of human agency—one that enables reasoning and comparison across different forms of action. As discussed above, a mechanistic view of agency cannot adequately account for the role of moral principles or the affective dimension, which are foundational to human moral experience. AI systems may function as co-agents in decision-making contexts, particularly in sensitive domains, by serving as moral evaluators—tools that support human judgment by providing relevant information to help individuals reflect more critically.

AI agency could be presented as a form of *delegated agency*, whose essential aspects are determined in advance, or at least can be predefined by human agents. This means that the goals, constraints, and decision-making parameters of AI systems are externally set, rather than internally generated. Recognizing AI as delegated agency reinforces the importance of human responsibility in the design and oversight of these systems, as the moral and functional boundaries are established prior to deployment.

Table 1 offers a summary of key topics for educators, providing a preliminary framework that connects philosophical distinctions about human agency with the technical accounts that underpin AI systems. By juxtaposing mechanistic and volitional models, it serves as a point of reference for designing learning activities that help students critically engage with both the computational logic of AI and the moral dimensions of human action.

Competences We believe that a key competence AI experts should develop is the ability to recognize the qualitative distinction between actions performed by humans and those performed by AI systems. Addressing the distinction—and the relationship—between human and artificial agency allows us to

	Agency	
	Artificial	**Human**
Relation with values	indirect / mediated	direct
Evaluation	performance / efficiency	quality of motives
Processed information	measurements / quantities	motives / desires / emotions
Processing	computing / optimization	ethical reflection / group discussion
AI literature	mechanistic / rational / explicit agency	volitional / full ethical agency
Philosophical literature	Utilitarianism	Virtue Ethics / Thomas Aquinas

Table 1
Summary of comparable topics on human agency for AI ethics education

acknowledge a fundamental difference between performing a morally right action (including the ability to appeal to acceptable ethical principles) and being held morally responsible for that action [38].

For instance, while AI systems may perform actions that align with ethical principles and justify those actions within programmed or learned frameworks, they lack the qualities—such as intentionality and free will—that are necessary for true moral responsibility. This distinction is reflected in key terms used in AI regulation, such as accountability and responsibility. The former is a foundational requirement for trustworthy AI [6], and although AI systems can be held accountable for their outcomes in a procedural or operational sense, they cannot be held fully responsible, since responsibility remains distributed among various human actors involved in the system's design, deployment, and oversight.

The difference between accountability and responsibility means recognizing the qualitative gap between actions carried out by artificial agents and those of human agents. It also points to the necessity of human oversight as a safeguard for moral reasoning and responsibility. As we will argue, the fundamental difference between human agency and the agency of AI systems lies in the internal processes that animate human decision-making, which are absent in artificial moral agents.

Approach Finally, we argue that AI ethics education should take into account the subject of ethics: that is, the human being. Before asking what an AI system should or should not do, we must first cultivate our own moral experience [39]: understand the principles that guide us, reflect on what we consider right and wrong, and sincerely engage with tensions and moral dilemmas. Ethics education should begin not with machines, but with ourselves.

As discussed in section 2, it is possible to distinguish two perspective on how ethics can be addressed in teaching - respectively "as a tool" or "as content" -, and to improve and enrich our own moral experience it is necessary to concentrate also on the second perspective, i.e. ethics as content. Adopting this approach enables the education of individuals not only as student or AI experts, but primarily as person. This means that ethics should not be treated as an auxiliary component - a tool - that as student/expert we use to solve problems, but rather as a fundamental aspect of human existence, as one of our most authentic dimensions. The teaching of ethics "as a content" should be seen *not merely* as the transmission of abstract definitions, values and principles, but also but as an engagement with ethical reflection that connects theory to real-world experience.

All the teaching activities should not be purely frontal but they should integrate case-studies, ethical dilemmas via group work to use knowledge of ethics in a practical context. We recommend an active learning approach [40] where students could deal with a decision about an action (e.g. "Is it ethical to deploy a chatbot that gives human-like responses without disclosing that it is an AI?"), that could be analyzed form different perspectives (such as utilitarianism, deontology and virtue ethics), highlighting the strengths and weakness of each approach. When we address "ethics as a content," we are laying the foundations upon which work on "ethics as a tool" can be built. For example, when confronted with an ethical issue in AI, such as the deployment of autonomous weapons in military contexts, before asking what the software should or should not do, we should first ask ourselves what we would do. This additional layer of reflection enables us to approach the "applied" problem with greater awareness, as it

involves a twofold examination: one personal and one technical. Such self-reflection also engages the affective dimension of agency (section 4.2), since ethical understanding arises not only from reasoning about actions from an external or objective standpoint, but also from a subjective perspective that involves the agent's inner motivations, emotions, and moral sensibility.

Example of a teaching activity for an AI ethics course Where ethics and AI intersect, it is essential that learning occurs through practical activities such as case study analysis and group discussions of real-world issues. To highlight the difference between a mechanistic and a volitional approach—that is, between merely achieving a goal and reflecting on *how* that goal is achieved - we recommend to complement the question "How do we achieve the goal?" with a deeper inquiry: "What am I doing when I choose this particular means to an end?" The latter question is not concerned with performance or efficiency but instead invites consideration of moral experience, shifting attention to the individual steps that lead to a result. For instance, we might consider an AI application designed to assist doctors and nurses in allocating medical resources during a pandemic, when the number of patients exceeds the available equipment.

We envision a task where students are asked to design a program that determines how to allocate medical resources to save as many patients as possible. The activity can be structured in two sequential phases. In the first phase, students work collaboratively to maximize the number of lives saved under the explicit criterion of maximizing the probability of success. The instructor provides examples that students analyze and solve, organized along a scale of increasing complexity: initial scenarios may involve a small number of patients with comparable ages and medical conditions, whereas later ones may introduce more challenging situations, such as significant age disparities or severe shortages of medical equipment.

In the second phase, students revisit the same scenarios, but without the constraint of maximizing the probability of success. They are free to decide how resources should be allocated in each case, engaging in collective discussion and justification of their choices. They may then perform a "moral check" on the decisions made across the two sessions to assess whether the paradigm they adopted influenced their results. When divergent decisions arise for the same scenario, this moral check can prompt reflection on the values and ethical principles that guided the choices in the second phase, and how these differ from those informing the initial, success-oriented approach.

The first part of the activity represents the task from the perspective of the mechanistic model of agency, where the goal is to achieve the most efficient result. To do so, it is necessary to reduce the number of variables, assign numerical values to them (for example, the condition of a young patient or the survival probability of an elderly one), and calculate the likelihood of each outcome. Once these parameters are established, the next step is to minimize the output function to identify the optimal solution — namely, the most effective course of action. This model of reasoning, or mechanistic agency, reflects the way computational systems operate and should be placed in dialogue with the alternative, volitional model. The same problem can thus be reinterpreted from a volitional perspective, asking instead: "What are we doing when we choose this particular means?" This shift invites reflection on questions such as how we assign value to human life and under what assumptions (for instance, is a young life considered "worth" more than an older one?). Addressing such questions calls upon our moral awareness rather than merely our capacity to design efficient computational solutions.

The collaborative nature of this exercise is essential: engaging with others allows students to broaden their moral outlook and critically examine their principles—an enriching feature of the volitional approach. Although the example discussed here is specific, this kind of "meta-analysis" — a reflection on what we are doing when we engage with AI — can and should be applied at every level, whether as programmers or as users. This level of reflection enables us to address ethical issues as fundamentally human concerns rather than as purely mathematical problems, emphasizing the moral competencies that must intertwine with technical and professional expertise.

6. Conclusions

In summary, this paper calls for a renewed focus on the concept of human agency in AI ethics education. By contrasting mechanistic and volitional models of agency, we have argued that a more comprehensive understanding of human moral experience is essential for preparing AI professionals to responsibly design, deploy, and oversee intelligent systems. Our argument rests on the idea that education on human agency and the demands of human oversight are deeply interconnected.

In particular, this work proposes volitional agency as the most appropriate paradigm for describing human agency — something the mechanistic approach is ill-equipped to capture. A volitional account of agency reminds us that moral reflection allows for a plurality of perspectives, and that such plurality should not be dismissed as a flaw. On the contrary, it offers an opportunity to strengthen critical thinking, especially when contrasted with the mechanistic model of agency that underpins most AI systems. Importantly, our aim is not to advocate for the implementation of volitional agency in AI systems, but rather for an AI ethics education that acknowledges the qualitative difference between human and machine agency. In doing so, it encourages a shift toward an ethics curriculum that prioritizes the human dimension as a necessary foundation for responsible AI development and oversight.

A stronger emphasis on moral reasoning and ethical reflection can contribute to more thoughtful and effective oversight practices. Making decisions about when, whether, and how to use an AI system - or evaluating the implications of its outputs in sensitive contexts - requires not only technical competence, but also a clear grasp of the ethical dimensions of human agency. Integrating this perspective into AI ethics education enhances ethical competence and strengthens the capacity for critical reflection in complex socio-technical environments. In this sense, fostering ethical reflection grounded in human agency directly supports the educational shift proposed in this paper—one that recognizes the qualitative difference between human and machine forms of agency as central to responsible AI development.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Grammar and spelling check and reward. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] L. Pham, B. O'Sullivan, T. Scantamburlo, T. Mai, Addressing digital and ai skills gaps in european living areas: A comparative analysis of small and large communities, Proceedings of the AAAI Conference on Artificial Intelligence 38 (2024) 23119–23127. URL: https://ojs.aaai.org/index.php/AAAI/article/view/30357. doi:10.1609/aaai.v38i21.30357.

[2] T. Scantamburlo, A. Cortés, F. Foffano, C. Barrué, V. Distefano, L. Pham, A. Fabris, Artificial intelligence across europe: A study on awareness, attitude and trust, IEEE Transactions on Artificial Intelligence 6 (2025) 477–490. doi:10.1109/TAI.2024.3461633.

[3] D. T. K. Ng, J. K. L. Leung, S. K. W. Chu, M. S. Qiao, Conceptualizing ai literacy: An exploratory review, Computers and Education: Artificial Intelligence 2 (2021) 100041.

[4] J. Su, D. T. K. Ng, S. K. W. Chu, Artificial intelligence (ai) literacy in early childhood education: The challenges and opportunities, Computers and Education: Artificial Intelligence 4 (2023) 100124.

[5] University of Helsinki and MinnaLearn (Reaktor), The elements of ai, https://www.elementsofai.com/, 2018. Online course.

[6] E. Parliament, Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828 (artificial intelligence act), 2024. URL: http://data.europa.eu/eli/reg/2024/1689/oj.

[7] European Commission AI Office, Living repository to foster learning and exchange on ai literacy, https://digital-strategy.ec.europa.eu/en/library/living-repository-foster-learning-and-exchange-ai-literacy, 2025.
[8] B. J. Grosz, D. G. Grant, K. Vredenburgh, J. Behrends, L. Hu, A. Simmons, J. Waldo, Embedded ethics: integrating ethics across cs education, Commun. ACM 62 (2019) 54–61. URL: https://doi.org/10.1145/3330794. doi:10.1145/3330794.
[9] A. Aler Tubella, M. Mora-Cantallops, J. C. Nieves, How to teach responsible AI in higher education: Challenges and opportunities, Ethics and Information Technology 26 (2024). URL: https://doi.org/10.1007/s10676-023-09733-7. doi:10.1007/s10676-023-09733-7.
[10] S. J. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed., Pearson, Hoboken, NJ, 2020. URL: http://aima.cs.berkeley.edu/.
[11] A. Jobin, M. Ienca, E. Vayena, The global landscape of ai ethics guidelines, Nature machine intelligence 1 (2019) 389–399.
[12] E. Commission, C. Directorate-General for Communications Networks, Technology, G. ekspertów wysokiego szczebla ds. sztucznej inteligencji, Ethics guidelines for trustworthy AI, Publications Office, 2019. doi:doi/10.2759/346720.
[13] L. Methnani, A. Aler Tubella, V. Dignum, A. Theodorou, Let me take over: Variable autonomy for meaningful human control, Frontiers in Artificial Intelligence 4 (2021) 737072.
[14] B. Green, The flaws of policies requiring human oversight of government algorithms, Computer Law & Security Review 45 (2022) 105681.
[15] S. Sterz, K. Baum, S. Biewer, H. Hermanns, A. Lauber-Rönsberg, P. Meinel, M. Langer, On the quest for effectiveness in human oversight: Interdisciplinary perspectives, in: The 2024 ACM Conference on Fairness, Accountability, and Transparency, 2024, pp. 2495–2507.
[16] N. Garrett, N. Beard, C. Fiesler, More than "if time allows": The role of ethics in ai education, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 272–278. URL: https://doi.org/10.1145/3375627.3375868. doi:10.1145/3375627.3375868.
[17] N. Kiesler, S. Opel, C. Thorbrügge, With great power comes great responsibility - integrating data ethics into computing education, in: Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1, ITiCSE 2024, Association for Computing Machinery, New York, NY, USA, 2024, p. 471–477. URL: https://doi.org/10.1145/3649217.3653637. doi:10.1145/3649217.3653637.
[18] N. Brown, B. Xie, E. Sarder, C. Fiesler, E. S. Wiese, Teaching ethics in computing: A systematic literature review of acm computer science education publications, ACM Trans. Comput. Educ. 24 (2024). URL: https://doi.org/10.1145/3634685. doi:10.1145/3634685.
[19] I. D. Raji, M. K. Scheuerman, R. Amironesei, You can't sit with us: Exclusionary pedagogy in ai ethics education, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 515–525. URL: https://doi.org/10.1145/3442188.3445914. doi:10.1145/3442188.3445914.
[20] S. J. Russell, Rationality and intelligence, Artificial Intelligence 94 (1997) 57–77. URL: https://www.sciencedirect.com/science/article/pii/S000437029700026X. doi:https://doi.org/10.1016/S0004-3702(97)00026-X, economic Principles of Multi-Agent Systems.
[21] M. C. Ramos, C. J. Collison, A. D. White, A review of large language models and autonomous agents in chemistry, Chemical science (2025).
[22] N. Mehandru, B. Y. Miao, E. R. Almaraz, et al., Evaluating large language models as agents in the clinic, npj Digital Medicine 7 (2024) 84. URL: https://doi.org/10.1038/s41746-024-01083-y. doi:10.1038/s41746-024-01083-y.
[23] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, X. Zhang, Large language model based multi-agents: a survey of progress and challenges, in: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI '24, 2024. URL: https://doi.org/10.24963/ijcai.2024/890. doi:10.24963/ijcai.2024/890.
[24] H.-L. E. G. on Artificial Intelligence, Ethics guidelines for trustworthy ai, 2019. URL: https://data.

europa.eu/doi/10.2759/346720.
[25] N. Cristianini, T. Scantamburlo, J. Ladyman, The social turn of artificial intelligence, AI & Society 38 (2023) 89–96. URL: https://doi.org/10.1007/s00146-021-01289-8. doi:10.1007/s00146-021-01289-8.
[26] J. von Neumann, O. Morgenstern, Theory of Games and Economic Behavior, Princeton University Press, Princeton, 1944.
[27] G. E. M. Anscombe, Modern moral philosophy, Philosophy 33 (1958) 1–19. URL: http://www.jstor.org/stable/3749051.
[28] J. Nida-Rümelin, N. Weidenfeld, Digital humanism: For a humane transformation of democracy, economy and culture in the digital age, Springer Nature, 2022.
[29] D. Edmonds, Would you kill the fat man?: The trolley problem and what your answer tells us about right and wrong, in: Would You Kill the Fat Man?, Princeton University Press, 2013.
[30] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, I. Rahwan, The moral machine experiment, Nature 563 (2018) 59–64.
[31] J. Dai, Position: beyond personhood: agency, accountability, and the limits of anthropomorphic ethical analysis, in: Proceedings of the 41st International Conference on Machine Learning, ICML'24, JMLR.org, 2024.
[32] A. Nie, Y. Zhang, A. S. Amdekar, C. Piech, T. B. Hashimoto, T. Gerstenberg, Moca: Measuring human-language model alignment on causal and moral judgment tasks, Advances in Neural Information Processing Systems 36 (2023) 78360–78393.
[33] N. Scherrer, C. Shi, A. Feder, D. Blei, Evaluating the moral beliefs encoded in llms, Advances in Neural Information Processing Systems 36 (2023) 51778–51809.
[34] J. Moor, Four kinds of ethical robots, Philosophy Now 72 (2009) 12–14.
[35] F. Fossa, Etica funzionale: Considerazioni filosofiche sulla teoria dell'agire morale artificiale, Filosofia (2020) 91–106. URL: https://ojs.unito.it/index.php/filosofia/article/view/5235. doi:10.13135/2704-8195/5235.
[36] M. S. Vaccarezza, Le Ragioni Del Contingente: La Saggezza Pratica Tra Aristotele e Tommaso D'Aquino, Orthotes, Napoli, 2012.
[37] G. Basti, La sfida etica dell'intelligenza artificiale e il ruolo della filosofia, Aquinas (2022) 299–322. URL: https://www.pul.it/cattedra/upload_files/13/Etica_IA.pdf.
[38] M. Anderson, S. Anderson, Machine Ethics, Cambridge University Press, 2011.
[39] S. Vallor, Technology and the virtues: A philosophical guide to a future worth wanting, Oxford University Press, 2016.
[40] P. Díaz, T. Onorati, I. Aedo, Using active learning and critical thinking to identify and apply ethical values in engineering education, in: 2024 IEEE Global Engineering Education Conference (EDUCON), 2024, pp. 1–10. doi:10.1109/EDUCON60312.2024.10578610.