Ethical Implementation and Conscious Application (ETHICA): Shifting Responsible AI from Compliance to Competency

Executive Summary

The dominant approach to incorporating ethics into the development of artificial intelligence, what is commonly referred to as the practice of Responsible AI, has created a culture that reflexively defers moral judgment to governance bodies and statistical tools rather than one that promotes independent decision-making. The former, referred to herein as the "rules and tools" approach, relegates ethics to a set of compliance measures rather than an integral part of technology development. The result is a disconnect between ethical principles and practical implementation.

This paper proposes a solution called ETHICA (Ethical Implementation and Conscious Application), a poly-ethical decision-making framework that aims to shift ethics away from a compliance-based approach toward a competency-based approach by framing ethics as a learned set of skills, specifically reasoning skills—those abilities used in drawing inferences, reaching conclusions, arriving at solutions, and making decisions based on available evidence [1]. In this approach, ethical practice becomes integrated into everyday development workflows, rather than being consigned to a separate compliance mandate, such as diversity, equity, and inclusion (DEI) or sexual harassment training.

Complementary to traditional academic courses, which are typically based on didactic learning pathways, ETHICA is designed as a set of on-demand practice modules housed within a GitHub code repository, thereby merging a traditional coding environment with a learning environment. The modules promote sound decision-making in morally complex situations and emphasize developing ethical reasoning skills rather than perfectly interpreting concepts from the great philosophers. The solution addresses three critical challenges: conflicting ethical codes, a gap between ethical principles and technical decisions, and the lack of a widely accessible system for developing ethical competency, ultimately enabling practitioners to move from unconscious compliance to conscious ethical reasoning in AI development.

Section 1: Introduction

Background

Responsible AI (RAI) suffers not from a lack of academic research or scholastic contribution, nor does it suffer from a dearth of rules and regulations in how it should be governed. What it suffers from is the absence of practical skills that enable people to make informed decisions in the everyday development cycle of artificial intelligence technologies. The absence of these practical skills creates a persistent gap in companies between recognizing risks and taking meaningful action [2]. The challenge is two-fold: first, understanding what one should do, and second, how one should go about doing it, especially when confronted with competing ethical systems of behavior that neither clarify nor inform what actions one should take.

The gap itself poses a real, non-theoretical danger, resulting in substantive ethical issues. Ethical issues pertaining to AI are often referred to as the problematic and immoral outcomes relevant to AI that arise from the development, deployment, and use of AI [3]. These issues pertain to the topics commonly associated with Responsible AI principles, including, but not limited to, transparency, explainability, accountability, and fairness. According to the AI Incidents Database that tracks instances of ethical misuse of AI, the number of reported AI-related 'incidents' rose from 161 incidents in 2023 to 264 incidents in 2024—a 64% increase [4]. A keyword analysis of the AI Incidents Database reveals a consistent pattern of preventable harms across multiple domains, including bias and fairness issues, transparency and explainability gaps, inadequate testing and validation, and regulatory and oversight challenges (See Table 1).

Each of these incidents represents an ethical failure point where better ethical reasoning arguably could have prevented harm. The facial recognition false matches in Buenos Aires suggest an inadequate consideration of accuracy thresholds, specifically in establishing an acceptable level of false positive rates [5]. The Character.ai incident highlights the dangers of ignoring vulnerable user populations, particularly teens who may be more susceptible to the influence of AI [6]. The Hoodline case illustrates how opaque and deceptive practices (e.g., misattributing AI-generated content to humans) lead to unethical outcomes when transparency principles are not maintained [7]. These examples illustrate the implementation gap between having awareness of ethical principles and the ability to apply them directly.

Research Problem

This paper examines how ethical decision-making abilities can equip AI developers with skills to inform and enhance daily choices throughout the AI development process. It formalizes these abilities using a structured methodology called the ETHICA (Ethical Implementation and Conscious Application) framework, designed to address the gap in the field of Responsible AI that has a wealth of ethical guidelines and tools yet lacks a systematic process for applying them to specific development dilemmas.

Primary Research Questions

1. What conflicts exist within current professional ethics literature, and what problems arise as a result?
2. What Responsible AI tools exist to aid in ethical decision-making, and what limitations do they have?
3. Is there a tendency for organizations to displace moral agency from human decision-makers to AI systems themselves?
4. Does the size of a company or its level of investment in Responsible AI training better prepare companies and their employees to deal with today's ethical challenges?

Important Distinction Between "Ethics" and "Morality"

It is important to note that while the terms "ethics" and "morals" will be used interchangeably, the difference between "ethics" and "morality" will be defined more strictly to facilitate understanding of the ETHICA framework. Ethics, derived from the Greek word Ethos, is the branch of philosophy that examines the rightness or wrongness of human actions. Ethics involves the understanding and interpretation of moral standards and their foundations, thus providing a framework for making decisions and judging actions within a specific context. Morality, in contrast, is based on what we inherit from our families and communities over time, consisting of the “dos” which are believed to be right, and the “don’ts” which are believed to be wrong. Morality is not justified; it is merely an explanation as to why people act as they do. Therefore, while ethics is grounded in active reasoning, morality is grounded in the passive reception of tradition that changes over time, as traditions are merely snapshots of cultural and societal values.

Objective

The proposed solution aims to address the ethical challenges that arise during the AI development process and is not intended as a comprehensive solution for dealing with the impact of AI on individuals, society, or the environment. Rather than simply systematizing ethics, ETHICA makes the ethical reasoning process itself more deliberate and visible. The primary goal of ETHICA is to create a methodology for ethically and consciously applying ethics to real-world issues and dilemmas, which differs from existing frameworks in its focus on daily decision-making processes rather than high-level principles. This differentiation contributes original thinking to the broader scholarship of Responsible AI and philosophy through a renewed focus on the value of maintaining human moral agency.

Target Audience

The primary audience is technology developers (e.g., software developers, machine learning engineers, and data scientists), who often possess asymmetrical skill sets due to the science, technology, engineering, and mathematics (STEM) curricula they have matriculated through. The secondary audience can be referred to as “data adjacent.” Roles that often work alongside core engineering teams, including strategic leads and program managers. These are highly influential roles integral to the technology development process and are often tasked with incorporating the viewpoints of multiple stakeholders. However, the proposed methodology is not designed to be domain-exclusionary. It applies to any company, its people, and the development of any AI technology.

Acknowledgment of Limitations

Ironically, ethical challenges may arise when proposing new ethical practices. First, as Alasdair MacIntyre rightly points out, modern society loses a crucial element of ethical life by attempting to institutionalize ethics, leading to moral disorder and incoherence [8]. In other words, there is a risk in trying to institutionalize ethics, which may occur from compressing multiple normative ethical theories into a single framework.

Second, the risk of deliberately excluding (or including) various philosophical thinkers and their ideas into the framework. Take, for instance, the writings of Aristotle and Oscar Wilde, both of whom wrote about the human condition and morality. Aristotle’s ideas on morality, specifically regarding Virtue Ethics, are extensive and renowned. Wilde’s is less so, but perhaps no less poignant. As Wilde once said, “Morality is simply the attitude we adopt towards people we dislike. [9]” While the teachings of Aristotle may be considered more socially palatable or academically established than the teachings of Oscar Wilde, there is still poignancy to Wilde’s observations. Still, there is a need to make deliberate choices on which ideologies to include or exclude, which could result in exclusionary thinking.

Third, the inherent risk of techno-solutionism, the flawed assumption that technology can and should be the primary means of solving complex social, political, and economic problems [10]. Any technical solution, no matter how well-conceived or thoughtfully prepared, inherently risks oversimplifying complex cultural and societal issues to something resembling system bugs that can be remedied. In the book “To Save Everything, Click Here: The Folly of Technological Solutionism,” Evgeny Morozov’s sober warning reminds us of the dangers of reducing multifaceted problems to neatly defined technical puzzles, overlooking the root causes and human factors involved.

Section 2: The Current Landscape of Professional Ethics

“Codes” have historically taken on many forms, including “code of conduct,” “code of etiquette,” and “code of ethics.” Common amongst all codes is an outline of the expected behaviors or rules that individuals or members of an organization are expected to follow. In some instances, codes attempt to incorporate certain ethical concepts. For example, in the nineteenth century, conduct books were typically moralizing, as social rules were always justified by moral values, such as modesty, sincerity, and fortitude [11]. In the early 20th century, professional codes of ethics began to emerge, although many were relatively simple, often focusing on fundamental principles of honesty and confidentiality. As professions became more complex, these codes expanded in scope. For instance, in 1847, the newly organized American Medical Association (AMA) adopted its first code of medical ethics [12]. This code was transformed in the early part of the 20th century into a series of principles of medical ethics, the most recent version of which was adopted in 1980. The code represents one of the most notable examples of a profession’s attempt to regulate itself by self-imposed ethical standards.

Generally speaking, most professions subscribe to a code of ethics, whether explicitly or implicitly. A code is beneficial to both companies and practitioners because it provides a set of rules for ethical behavior, mitigates legal risk, and promotes reputation and trust. Essentially, codes provide a framework for how individuals should interact with each other and with the organization. For example, the American Bar Association (ABA) establishes a framework for the interactions of individuals, particularly lawyers, within the legal profession by outlining a set of ethical duties and guidelines that govern various aspects of legal practice, encompassing client-lawyer relationships, interactions with other parties, and professional conduct within the legal system [13].

Analysis of Professional Ethics Literature

A systematic review of professional ethics literature was conducted to understand the associated strengths, weaknesses, and risks of ethical frameworks (See Table 2). The insight that emerges from the analysis is that a lack of ethical principles and guidance isn’t the primary problem; there is a plethora of ethical guidelines to choose from and an abundance of literature that either identifies the risks of AI or outlines Responsible AI principles [14], [15]. The problem is the conflict and complications that arise from the multitude of codes. The conflict between these codes justifies why conscious, ethical decision-making is essential; practitioners cannot rely solely on external authorities to provide clear guidance, as those authorities fundamentally disagree about what AI ethics means and how it should be implemented. Although the terminology may vary slightly across sources, codes of ethics generally fall into one of five categories:

1. Professional Codes of Ethics

   Codes of ethics are established by professional organizations to guide the conduct of their members. They establish standards for professional conduct, such as those found in fields like medicine, law, or engineering, such as the National Society of Professional Engineers (NSPE) [16]. Professional codes typically emphasize personal duty, as seen in the NSPE requirement that “Engineers shall accept personal responsibility for their professional activities.” These codes are usually enforceable, and violations may result in financial penalties or expulsion from the organization.

2. Non-Governmental International Organizational (NGO) Code of Ethics

   These are developed by non-profit organizations, like the one authored by the International Organization for Standardization (ISO) [17] and the Institute of Electrical and Electronics Engineers [18]. These codes aim to define and uphold common values across industries, such as honesty and integrity and are unique in their focus on public trust. NGO codes typically seek global consensus and cross-cultural applicability, as demonstrated by UNESCO's AI Ethics Observatory, adopted by 193 countries [19]. These organizations often derive authority from broad national or international acceptance rather than through direct enforcement.

3. Corporate Codes of Ethics

   These codes, often a reflection of a company's core values, outline ethical standards that guide individual behavior and the development of products and technology. An example is Microsoft's Responsible AI Principles [20]. These types of ethical codes can be either compliance-based or normative-based (also known as "values-based"), or a combination of the two. Corporate codes sometimes include unique values the company thinks is differentiating, such as Capgemini's emphasis on "Boldness" and "Modesty." Whether these attributes are truly ethical principles is debatable.

4. Government/Regulatory Codes of Ethics

   These codes are developed by governmental bodies or regulatory agencies to establish legally binding standards for public and private sector activities. Examples include the FTC's AI Compliance Plan [21] and NIST's Risk Management Framework [22]. These codes are enforced through law or regulation, and violations typically result in fines, sanctions, or legal action. Government codes tend to be more detailed than other types of codes, typically providing specific procedures and requirements that must be complied with rather than simply establishing broad principles.

5. Academic/Research Institution Codes of Ethics

   These codes are developed by universities, research institutes, and academic organizations to guide ethical research and innovation practices. Examples include Stanford's HAI Artificial Intelligence Bill of Rights [23] and The Alan Turing Institute's AI ethics framework [24]. These institutions are often involved in creating policies, providing public education, and researching new technologies, and as such, emphasize academic rigor, ethical research, and evidence-based approaches to decision-making.

These categories of codes are not just different sets of principles; they are conflicting approaches to ethics in the context of AI. For instance, we can see a paradox between flexible and rigid principles, geographic and jurisdictional conflicts, and problems that arise from unclear enforcement mechanisms, suggesting that some codes of ethics, perhaps, serve more as reputational protection more so than genuine ethical commitment—a phenomenon sometimes referred to as "Ethics Washing" or "Virtue Signaling."

We also see conflicts between the ethics set forth in the guidelines and the realities of how companies develop technology. The National Society of Professional Engineers (NSPE) code states that "Engineers shall accept personal responsibility for their professional activities," promoting the principle of accountability. However, modern engineering work is often collaborative, involving multidisciplinary teams that often have dependency requests. These conflicts represent a systematic problem across professional ethics codes that may have been initially designed for individual practitioners working in isolated contexts, not for the highly collaborative, interdisciplinary nature of contemporary technology development.

Beyond conflicts within individual codes, practitioners also face tensions between entirely different ethical approaches, including those between procedural ethics and outcome-focused ethics. For example, the ISO code emphasizes the importance of "Performing and acting in good faith, consistent with the purpose, policies, and principles of the organization. [17]" The focus is on how people should conduct themselves through standardized processes, which creates consistent behavioral expectations across the organization. Whether deliberate or not, this approach is akin to deontological ethics, focusing on rules that must be upheld. Comparatively, Google's AI Principles center on outcomes, such as "mitigating unintended or harmful outcomes" and "avoiding unfair bias." This approach aligns with a consequentialist approach to ethics, whereby the moral rightness or wrongness of an action is judged by the outcomes it produces. In this instance, practitioners facing both ISO procedures and Google-style outcome requirements are not just dealing with competing policies; they are navigating fundamentally different normative ethical theories. Without conscious awareness of this distinction, they might default to whichever approach feels safer, apply inconsistent reasoning across similar situations, or be paralyzed by philosophical tensions that result in inaction.

The systematic conflicts—both within codes and between them—demonstrate the need for more flexible frameworks for ethical decision-making rather than strict adherence to a single set of moral principles overlaid on top of professions and industries. Professionals cannot rely on simple rules because ethical principles frequently conflict, requiring deliberate weighing of competing goods rather than unconscious compliance with individual principles.

Current Landscape of Responsible AI Tools

The previous section highlights not only the conflict between competing codes of ethics but also the conflicts between theoretical rules and the practical realities of how companies develop technology. Unsurprisingly, conflict patterns also exist in the realm of Responsible AI tools. Like codes of ethics, Responsible AI tools also suffer from an overabundance of competing options, compounded by the meta-problem of a lack of decision-making resources to aid in selecting the most appropriate tool for any given situation. A noticeable pattern seen in the ecosystem of tools reveals that many of them strictly align with a single principle (e.g., Fairness, Accountability, Explainability, etc.). Below are three notable dangers stemming from a non-exhaustive audit of current Responsible AI Tools (See Table 3).

The first danger is strictly aligning tools with ethical principles. By organizing ethical decisions around discrete principles, decision-makers face the self-imposed risk of making fragmented decisions within ethical silos that lack interconnected considerations. To illustrate the difference, consider a social media platform's fair rate limits for an Application Programming Interface (API), a set of restrictions on the number of requests a client can make within a specified time frame. In this context, "fairness" is designed to ensure equitable access to resources and the performance of the API; therefore, imposing fixed rate limits may seem a reasonable decision. Fairness is relatively straightforward and would generally be considered a low-risk scenario, as there is little ethical significance in balancing the needs of individual users with the overall health and availability of the API service. "Ethical significance" is defined as anything that increases or decreases one's welfare, regardless of how small or large the amount of welfare is [25]. Now consider a more ethically complex scenario: a social media platform's policy on API rate limits during a natural disaster. Both journalists and emergency responders justifiably need higher rate limits to share critical information. The argument for fixed rate limits now seems more brittle as the platform must balance both access to critical information and platform stability. Unlike the straightforward, fixed-rate limit scenario, the disaster example illustrates how ethical principles resist compartmentalization. Decisions about fairness invoke questions of accountability (e.g., who is responsible for determining and defining rule exceptions?) and questions of transparency (e.g., how and when do we make these exceptions available to the general public?)

The second danger is that each of these tools inherently promotes a reliance upon technology to make decisions regarding the ethical problems that emerge from the development of AI. Once again, we see the risk of techno-solutionism: the idea that technology can solve complex ethical and social problems. However, it is not simply the tool itself that users risk becoming overly reliant upon; it is the statistical output of these tools. Responsible AI tools often focus on what is quantifiable. For example, bias metrics to detect imbalances in training data, fairness statistics for quantifying harm, and interpretability scores to explain the output of any machine learning model by measuring each feature's contribution to a prediction. However, this raises the question of whether ethical decisions can be arrived at purely through mathematical reasoning, and whether the ethical decisions arrived at using these techniques skew towards a particular normative ethical theory, most obviously, utilitarianism. As a point of comparison, an instantiation of truthfulness, one of eleven moral virtues espoused by Aristotle in his work Nicomachean Ethics [26], would be difficult to prove through mathematics or statistics alone. The point being that an over-reliance on tools inadvertently breeds an over-reliance on mathematical formulas—a risk Responsible AI is meant to combat. The result may be something akin to ethical de-skilling, whereby the capacity for ethical reasoning deteriorates over time as Responsible AI practitioners become overly reliant on the output of tools to make decisions.

The third danger is the risk of embedded values. Unlike the previous challenges, techno-solutionism and compartmentalization, embedded values are more nefarious because they operate invisibly. The question of whether and how technologies embody values is not a new one. It has been discussed in the philosophy of technology, where several accounts have been developed [27], [28]. While debates exist about whether AI systems can act as moral agents, the principal risk is that these tools embed morals while appearing objective. Technical decisions, while seemingly benign, mask the fact that every algorithmic feature, hyperparameter, and coefficient embeds, either deliberately or inadvertently, decisions disguised as neutral technical features. Fairlearn, an open-source tool improve AI fairness, illustrates how Responsible AI tools may harbor multiple layers of embedded values. One example is the statement in its user guide, "we define whether an AI system is behaving unfairly in terms of its impact on people — i.e., in terms of harms — and not in terms of specific causes, such as societal biases, or in terms of intent, such as prejudice." Immediately apparent is the emphasis on the words "impact" and "harms," which implies utilitarian ethics. A second example is how Fairlearn performs fairness assessments, grounded in metrics such as demographic parity and equalized odds, which embeds the notion that comparative fairness between groups can be arrived at through mathematical optimization rather than human deliberation. A third example is Fairlearn's public statement that "We have assumed that every sensitive feature is representable by a discrete variable. Fairlearn acknowledges this limitation, but frames it as a minor technical constraint rather than a fundamental philosophical problem, stating that "Features like this have to be binned, and the choice of bins could obscure fairness issues." Statements such as these create the risk of training practitioners to treat hidden disparities as technical limitations that should be deprioritized. Taken together, these examples demonstrate how a series of innocuous technical decisions can result in the encoding of very particular moral worldviews.

Displaced Moral Agency in AI Ethics

A singular challenge in the current landscape of Responsible AI is the tendency for organizations to displace moral agency from human decision-makers to AI systems themselves. When organizations attribute moral responsibility to AI systems, they obscure human accountability and avoid the more challenging work of explicit ethical reasoning. Consider Microsoft's six Responsible AI principles, which are as follows [29]:

- Fairness - AI systems should treat all people equitably and avoid reinforcing societal biases.
- Reliability and Safety - AI should perform reliably as intended and be safe, preventing harm in both normal and unexpected conditions.
- Privacy and Security - AI must safeguard personal data and resist malicious misuse, protecting user privacy and data integrity.
- Inclusiveness - AI should empower everyone and engage broad perspectives, including marginalized or underrepresented groups, to avoid exclusion.
- Transparency - AI operations should be understandable; people should know when they are interacting with an AI and understand the system's decisions or recommendations.
- Accountability - Developers and organizations must be accountable for their AI systems, with governance mechanisms to ensure responsibility and oversight.

Microsoft's statement that "AI systems should treat all people equitably and avoid reinforcing societal biases" is an example of displaced moral responsibility. While seemingly reasonable, the statement implies that fairness is an attribute of the AI system itself. In reality, AI systems have no moral agency, defined as an individual who can, to a significant extent, act effectively and competently in moral matters [30]. An AI system cannot choose to be fair or unfair. Fairness emerges from the conscious decisions made by humans, those who design, implement, and govern the process of training data selection, algorithmic design, deployment contexts, and governance structures.

The displacement of moral agency becomes more obvious when companies like Microsoft pose questions such as: “How might an AI system allocate opportunities, resources, and information in ways that are fair to the humans who use it?” The framing treats fairness as a technological feature rather than a normative ethical choice that requires human reasoning, including decisions such as whether the goal is to achieve fairness at the individual or group level. Properly framing ethical commitments becomes challenging when organizations demonstrate their values through declarative statements about AI system behavior, while avoiding the deeper questions and critical thinking required for conscious ethical reasoning.

By maintaining a focus on human moral agency, the ETHICA framework promotes the practice of conscious, rather than unconscious, ethical decision-making by shifting responsibility away from technological systems. Rather than asking, “How should AI systems treat people fairly?” the framework draws on metaethical reasoning by facilitating questions asking, “How should we, as an organization or engineering culture, ensure our AI systems reflect our conscious ethical commitments about fairness?” This reframing requires explicit engagement with normative ethical frameworks, transforming vague compliance statements into substantive moral reasoning.

Current industry trends for implementing ethics training

Understanding how businesses approach responsible AI has become increasingly important as AI systems are deployed globally. In 2024, a joint survey between Stanford University Human-Centered Artificial Intelligence and McKinsey & Company found that companies of all sizes were making sizeable investments in Responsible AI. Larger enterprises—particularly those with annual revenues exceeding $10 billion—demonstrated higher total investment into RAI. Notably, 27% of organizations with $10 billion–$30 billion in revenue and 21% of those exceeding $30 billion invest $10 million–$25 million in RAI. These findings suggest that larger organizations are more likely to embed RAI as a strategic priority and to make higher absolute investments. Smaller organizations allocated fewer dollars to RAI, but many still reported substantial investments as a share of their revenue [31].

The training in AI Ethics and Responsible AI is neither a straightforward investment nor a straightforward return. Both the complexity of implementing Responsible AI practices and the difficulty of measuring their effectiveness may also contribute to companies’ investment levels or lack thereof. In 2023, MIT Sloan Management Review and Boston Consulting Group assembled an international panel of AI experts to help us understand how responsible artificial intelligence (RAI) was being implemented across organizations worldwide [32]. They found that most panelists recognize that RAI investments are falling short of what is needed: Eleven out of thirteen were reluctant to agree that organizations’ investments in responsible AI are “adequate.”

Funding alone does not prepare companies and their employees to deal with today’s ethical challenges. PwC’s 2024 US Responsible AI Survey asked organizations which aspects they most commonly prioritized, including upskilling, embedded AI risk specialists, periodic training, data privacy, data governance, cybersecurity, model testing, model management, third-party risk management, specialized AI risk management software, and monitoring and auditing [33]. Two important themes emerged from the survey. First, only 11% of executives report having fully implemented these fundamental responsible AI capabilities, and many were suspected of overestimating progress. Second, a key challenge facing companies implementing Responsible AI is the same as in most risk programs: It is hard to quantify the value of having dodged a bullet, such as avoiding a major scandal from a poor AI interaction.

The various training interventions in the workplace, which include reading materials, lectures, discussions, case studies, in-class exercises, role-play, and experiential learning, also present a mixed picture when trying to evaluate efficacy. Research on ethics education indicates that some modes of training may be more effective than others. One group of researchers found that attending lectures on moral philosophy had no significant impact on participants' moral judgment scores based on their performance on the Defining Issues Test-2 (DIT-2), a questionnaire designed to measure an individual's level of moral reasoning by assessing their preference for different schemas or ways of interpreting moral dilemmas [34]. However, engaging in moral problem-solving, discussing case studies, and writing arguments that apply criteria for judging the quality of moral arguments has been shown to have the greatest, albeit modest, impact on ethical judgment. These insights helped inform the pedagogical approach to the ETHICA framework and how it engages its users.

Section 3: ETHICA Goals & Guiding Principles

The ETHICA framework is built upon foundational goals and principles that work together to guide responsible AI development. Goals articulate what the framework aims to achieve, while principles establish the specific rules guiding its design. Importantly, technologists using the ETHICA framework interact primarily with its learning modules that simulate real-world scenarios rather than with abstract philosophical principles. Although technologists may not engage with these principles explicitly, the principles work behind the scenes to ensure the framework remains ethically sound and coherent across different implementations and contexts. The principles were intentionally selected based on their collective ability to serve the framework's goals, which explains their close alignment.

Framework Goals:

1. The framework should be considered poly-ethical, intentionally incorporating multiple ethical traditions rather than being grounded in a single normative ethical theory (e.g., consequentialism, deontology, or virtue ethics)
2. The framework must address the three branches of ethics: normative ethics, applied ethics, and metaethics, to ensure practitioners of the framework have the means to take action, reflect upon that action, and maintain clarity on the normative ethical theories to which they inadvertently or advertently subscribe.
3. The framework must not directly prescribe guidance as to how to make decisions (e.g., IF *x* THEN *y*) but instead aid in guiding one towards their own decisions.
4. The framework must be adaptable to different contexts of AI development since ethical considerations may vary between different AI applications and deployment contexts (e.g., an AI-powered agentic system that manages a work calendar versus a therapy chatbot whose purpose is to prescribe self-help).
5. Framework goals should not remain entirely immutable but instead, be open to review and reflection over time to ensure the framework maintains its anti-dogmatic nature and continues to adapt as AI technologies evolve.

Guiding Principles:

Principle no. 1: The framework incorporates the three branches of ethics, with an emphasis on applied ethics and decision making.

As a philosophical discipline, ethics, or moral philosophy, is divided into three parts: normative ethics, applied ethics, and metaethics [35]. The role of normative ethics is to establish the rules or principles that define what actions we should take. Normative ethics can be considered the foundational pillar of ethics, as it establishes the fundamental belief system of what differentiates right from wrong, which beliefs one should maintain, and which traits a virtuous person should have. Normative ethics informs applied ethics, which is the confrontation of moral dilemmas in everyday life; confrontations that require action. Unlike normative ethics, which describes what one "ought" to do, applied ethics is principally concerned with "how" one should do it, through the analysis and resolution of specific real-world problems. It is, therefore, the role of applied ethics to make decisions within the moral boundaries set by normative ethics. The third part, metaethics, is the study of moral thought and language [36]. Unlike normative ethics, metaethics is not concerned with what one ought to do or should not do. Instead, metaethics is interested in how moral language and moral thought work, no matter what the contents of anyone's set of moral beliefs may be or their practices [37].

Below is a summary table outlining the differences in normative ethics, applied ethics, and metaethics using the example of "fairness," a widely cited principle of Responsible AI, for a hypothetical company.

Normative Ethics	Applied Ethics	Metaethics
"Is 'fairness' an ethical principle to which our company should subscribe?"	"How do we ensure our AI systems treat all customers fairly?"	"What does it say about our company if we choose to prioritize group fairness over individual fairness?"

Taken together, the three branches comprise a framework commonly referred to as moral theory. A fully developed moral theory often addresses all three areas of ethics (metaethics, normative ethics, and applied ethics). However, it aims to establish and defend the norms of conduct it recommends [38]. Any sound moral theory should have two fundamental aims: one theoretical and the other practical [39], since both theoretical understanding and practical "know-how" are requirements for dealing with real-world situations involving the conception, development, and maintenance of AI systems.

- Theoretical aim. The main theoretical aim of moral theory is to discover those underlying features of actions, persons, and other items of moral evaluation that make them right or wrong, good or bad [39].
- Practical aim. The main practical aim of a moral theory is to discover a decision procedure that can be used to guide correct moral reasoning about matters of moral concern [39].

The following section describes the three branches of ethics in detail, including examples of each branch within the ETHICA framework.

Normative Ethics is the branch of moral philosophy, or ethics, concerned with criteria of what is morally right and wrong [40]. Normative ethics provide the general moral rules governing our behavior, such as Utilitarianism, Deontology or Virtue Ethics. As Mark Dimmock and Andrew Fisher in Ethics for A-Level analogize, the normative ethicist, is like a referee who sets up the rules governing how the game is played [41].

Normative ethical theories have traditionally been divided into teleological or deontological categories [42]. Generally speaking, these competing ideologies differentiate on their prioritization of the “right” versus the “good.” Teleological theories can be thought of as those that define moral quality in terms of the achievement of some “good,” perhaps as an outcome or the instantiation of virtue. Deontological accounts are those that specify moral quality as a function of something else, such as a duty that binds, no matter the consequences. This is not to say that the achievement of some good is not important to deontological approaches, more so that these approaches do not define what is considered “right” [43].

The history of modern moral theory from at least John Stuart Mill onward has tended to occupy itself with the theory of right conduct, while theories of value and moral worth have often been in the service of the theory of right conduct [44]. This focus is reflected in the label “normative ethics,” with emphasis on theorizing about norms for behavior [45].

Normative ethical decision-making requires choosing an ethical framework, which is a bit like choosing a swim lane. The choice determines both the theory of right conduct and the justification for why actions are deemed right or wrong [45]. In the context of ETHICA, determining “right conduct” means establishing reliable methods that lead to sound moral judgments while providing justification for those decisions. Importantly, normative ethics are not a prescription for how morality should be applied but to what end those means serve. In essence, they are the concepts that guide actions, not the actions themselves.

Consider how different normative frameworks might evaluate the same AI governance practice. Suppose an online travel company mandates that employees create model cards (short documents detailing essential information about trained AI models) for every algorithm affecting end-users. The requirement demonstrates a commitment to transparency and explainability, core principles of Responsible AI practices. However, the ethical justification for this policy depends on which normative framework one adopts.

A deontological perspective might argue that companies have a duty to make algorithmic decision-making processes available to customers, regardless of consequences. Alternatively, a virtue-ethics approach would frame creating model cards as virtuous behavior, an instantiation of honesty and integrity. Both represent valid normative ethical justifications for identical practices, yet they emerge from different moral reasoning processes.

Rather than prescribing which ethical framework to adopt, ETHICA facilitates conscious engagement with these normative choices. Too often, responsible AI practitioners approach ethics as a compliance exercise, defaulting to legal requirements, industry standards, or perceived societal expectations without examining their normative foundations. The ETHICA framework’s key contribution is transforming this compliance-oriented mindset into genuine ethical competency through deliberate, justifiable reasoning in AI implementation, thereby solving practical problems that many practitioners face but often can’t articulate.

Applied Ethics addresses specific questions about the morality of such issues as abortion, capital punishment, the ethical treatment of animals, euthanasia, and sexual behavior. It also addresses ethical questions that arise within the professions [45], including the ethical development of AI.

Central to applied ethics—and particularly relevant to the ETHICA framework—is understanding how ethical decisions are made in practice. How we apply ethics is often shaped by the moral principles represented in normative ethics. However, normative ethics and applied ethics have a symbiotic, not a hierarchical relationship. Normative frameworks guide practical decisions, exposing gaps in theoretical approaches that lead to refinement of the underlying normative theories.

As Tom L. Beauchamp observed, “We are currently beginning to appreciate in moral theory that the major contributions in this area have recently run from ‘applied’ contexts to ‘general’ theory rather than from general to applied” [46]. This trend is particularly relevant in emerging fields like Responsible AI, where practical implementation challenges often expose inadequacies in existing theoretical frameworks. Ethicists gain valuable insights into the validity and completeness of normative theories by analyzing how they perform when applied to real-world moral challenges.

Through applied ethics, we understand how we should deal with issues like algorithmic bias and non-transparent machine learning systems that are difficult to explain. Continuing the sports analogy, the applied ethicist is the person on the field making second-by-second decisions that move the game forward. These decisions can come in the form of direct actions (for example, rebalancing a training dataset for a job applicant screening algorithm to include an equal number of male and female applicants) and also indirect actions, through arguments that may shape our ethical views in particular situations (for example, advocating for the diversification of job candidate recruiting pipelines) to be more representative of the underlying population.

Metaethics is the study of how we engage in ethics [36], including the fundamental question of whether morality exists. The metaethicist might comment on the meaning and appropriateness of ethical language or explain what we mean if we say that treating customers unfairly is wrong. The role of the metaethicist, then, is not to determine what to believe in or what actions we should take but instead to help us understand what these decisions mean by critiquing our theories and how they are applied. It is, therefore, the job of the decision maker to determine, generally, how much importance metaethics has relative to normative and applied ethics in any given situation. Metaethical theories are, by definition, pluralistic and can be generally categorized into two major axioms: Cognitivism and non-cognitivism, as well as realism and Anti-Realism [41].

Cognitivists argue that moral claims express beliefs that are empirically true or false, what philosophers call “truth-apt [41].” For example, a cognitivist might argue that “knowingly using biased datasets is morally wrong” is a Boolean rule: either true or false. In contrast, non-cognitivists would argue that this is not a factual claim at all, but rather an expression of emotion or attitude. In this case, the claim of “knowingly using biased datasets” expresses the disapproval of using biased datasets. However, it is not a truth-apt belief (i.e., not simply true or false). An important characteristic of cognitivism worth understanding is that it is only a theory of explaining the meaning of a moral statement (e.g., “Knowingly using biased datasets”), but has nothing to do with what exists in the world, which is the nature of Realism versus Anti-Realism.

Realism is a view about what exists. In ethics, realists hold that certain facts exist independently of whether we think they exist or have certain features, and that these facts exert a shaping influence or control upon thought and action [47]. For example, a realist may argue that the concept of ‘Fairness’ is a moral property that exists and that it exists objectively and independently of the minds or beliefs of individual people. As an analogy to help clarify this concept, a realist would argue that the Atlantic Ocean objectively exists, independently of whether someone believes it exists or not. Anti-realism, in contrast, denies the independent existence of such facts. It often holds that facts are dependent on human beliefs, perceptions, or conceptual frameworks, and may lack the objective status that realists attribute to them [47].

These metaethical theories can both confer and conflict. However, the ETHICA framework does not favor any single metaethical theory. Instead, it assumes all metaethical theories may be valid if they help arrive at a real-life decision and not just a purely academic conjecture that yields no real decisions. For example, the metaethicist may question or critique which definition of ‘Fairness’ should be used to examine a machine learning model: individual fairness versus group fairness, two ethical concepts that routinely conflict with one another, as group fairness aims to enforce the outcome to be equal in distribution or statistics among sensitive groups, whereas individual fairness aims to guarantee that individuals who share similar qualification data would receive similar outcomes [48]. The same metaethicist may argue that the concept of ‘Fairness’ can be empirically evaluated as either true or false. Therefore, if asked to interpret the statistical output of a machine learning model, the results would empirically be fair or unfair. Conversely, they may question the Boolean nature of fairness altogether and argue that fairness judgments are non-cognitive expressions of approval or disapproval rather than factual claims capable of being true or false.

Principle no. 2: The ETHICA framework does not strictly adhere to one normative ethical theory nor promote particular ethical behaviors.

The field of ethics (also known as moral philosophy) is concerned with systematizing, defending, and recommending concepts of right and wrong behavior [3]. At its absolute core, ethics seeks to answer the question, “What is the morally right thing to do?” The answer to this question can be expressed through several behaviors, including the actions that serve a particular behavior, the derived outcomes resulting from those actions, and the character traits that inspired those actions in the first place. While ethical theories typically address actions, outcomes, and character, they often emphasize a particular aspect. For example, consequentialism primarily focuses on the outcomes of actions, deontology emphasizes one’s moral obligations that inspire actions, and virtue ethics centers on the character traits and dispositions that motivate actions. These three approaches represent the major normative ethical theories in the Western philosophical tradition. Importantly, the framework is not anchored to a single normative ethical theory, such as Consequentialism, Deontology, or Virtue Ethics, nor does it attempt to institutionalize a universal moral code. This is a structural feature of the ETHICA framework, as no single theory encompasses every expression of ethical behavior. Below is a summary of the three major Western normative theories, along with a hypothetical example of how a moral agent, an individual expected to meet the demands of morality [49], may operate within the context of each. The summarization also includes a non-exhaustive list of strengths and weaknesses of each normative ethical theory.

Consequentialists believe that an action is morally right if the consequence of that action is viewed as beneficial, i.e., more favorable than unfavorable [50]. Consequentialists are therefore concerned with outcomes and effects, as opposed to underlying actions or moral character. A discussion about consequentialism is typically a discussion about *direct consequentialism*, which assesses all things, including actions, in terms of the value of their consequences. In contrast, *indirect consequentialism* assesses actions in terms of their conformity to rules, motives, or dispositions with good or optimal acceptance value [30]. Branches of consequentialism include, but are not limited to, the following [30]:

1. Act Utilitarianism, a direct form of consequentialism, asserts that an action is morally right if it maximizes overall happiness or value.
2. Rule Utilitarianism, an indirect form, evaluates actions based on adherence to rules that, if generally accepted, would lead to the best consequences.
3. Egoistic Consequentialism focuses on actions that promote the individual’s own welfare, often incorporating subjective or perfectionist conceptions of the good.

Like the three normative ethical theories themselves, each consequentialist sub-theory has its own strengths and weaknesses, but *Act Utilitarianism* is likely the most familiar form of consequentialism among them. Largely attributed to Jeremy Bentham [50] and John Stuart Mill [52], Act Utilitarianism maintains that an action is morally good if the consequences or effects of that action are more favorable than unfavorable to everyone [3]. As Bentham pointed out, utilitarianism involves performing moral arithmetic, or more precisely, *felicific calculus*, a method for calculating the moral value of an action based on its potential to produce pleasure or pain, which machines are well-suited for, given their mathematical nature [50].

- Strength(s): Impartiality and objectivity in decision making; theoretically straightforward demonstration of impact and outcomes; and an allowance for context-sensitivity depending on the specific context and circumstances.
- Weakness(es): The neglect of individual human rights, promoting the marginalization of people who may be disproportionately affected (perhaps continuously) and a bias towards ethical reasoning that relies on mathematics for ethical judgment.

Deontologists would consider an action to be morally right if it is based on its intrinsic nature, independent of the value it produces [51]. Therefore, deontologists are focused on the intentions of an action, not on the consequences [53]. Largely informed by the thinking of Immanuel Kant (1724-1804), deontology argues for the existence of universal principles, regardless of changes in culture, geography, or even time—a timeless and unanimous recipe for moral action. The adherence to universal principles rather than outcomes or consequences is what makes deontology characteristically unique. When we debate whether an action is inherently right or wrong regardless of its consequences, we're engaging with deontology. This duty-based approach to ethics continues to influence fields ranging from law and politics to medicine and, more recently, technology.

- Strength(s): The consistent and transparent application of rules and principles; respect for individual rights by not treating humans as a means to an end, but an end unto themselves [54]; and the demotion of the practice of moral relativism.
- Weakness(es): An emphasis on following rules may inadvertently overoptimize to maintaining the good standing of one's own reputation, rather than good moral behavior; an inability to apply rules to nuanced scenarios that involve a tension between competing values; and a lack of introspection on the underlying quality of the maxims that demand obedience.

Virtue ethicists would deem an action to be morally right if the agent acts and thinks according to some moral value or values, such as wisdom and bravery [55], [56]. Virtue ethics focuses on the character of the person performing actions, rather than the actions themselves or their consequences. Advocates of virtue ethics believe that modern theories that focus on what we should do lack something important, and suggest that it is equally important, or even more important, that we focus on the question of what characteristic traits individuals ought to develop in themselves [57]. Aristotle, perhaps the single greatest contributor to virtue ethics, sought to define the nature of virtue. Albeit a simplified version of Aristotle's more complex doctrine, virtue is a character trait—doing the right thing at the right time. Aristotle believed that we are not born virtuous, but that each of us has the capacity to become virtuous through practice. "By doing just things, we become just; moderate things, moderate; and courageous things, courageous. [26]" To become virtuous, a person must first be trained and then develop practical wisdom, or phronesis. The combination of the two not only cultivates good ethical habits but also makes more robust ethical decision-making skills needed when facing new questions, not unlike how a machine learning model would be considered robust if it can correctly classify new instances of data that weren't part of its original training data. This normative approach of assessing the broad characters of human beings rather than assessing singular acts in isolation is what separates Virtue Ethics from both Utilitarianism and Deontology [41].

- Strength(s): The ability to find a middle point between competing vices, more commonly referred to as the "Aristotelian mean." Ethical decision making improves over time with practice, and the practice of ethics becomes egalitarian and openly accessible.
- Weakness(es): Intellectually complicated; good ethical decisions cannot be made instantly and require practice, perhaps to the detriment of those impacted by a less experienced practitioner; and difficulty producing evidence instantiating the impact of virtuous decisions, good or bad.

A hypothetical example helps to illustrate the differences between the three normative theories: Suppose that a machine learning engineer is developing an algorithm to moderate hate speech on a social network. The engineer knows the algorithm may incorrectly filter some benign content along with genuine hate speech. How might they proceed within the context of the three normative approaches above?

From the consequentialist's perspective, the focus should be on maximizing overall positive outcomes and minimizing harm by answering the fundamental question: "What approach produces the best consequences or the most good?" The approach may focus on minimizing false negatives—missing genuine instances of hate speech—based on which errors cause the most harm, and a benefit-risk analysis weighing potential benefits (reducing hate speech) against risks (suppressing legitimate speech). The consequentialist may also incorporate the advice of corporate counsel, taking into account potential legal risks associated with undetected harms. However, it's worth noting that simply seeking the advice of legal counsel does not mean one has fallen victim to the "rules and tools" approach to ethics. Contrarily, a good AI ethicist should actively seek out the opinions of others to ensure they are considering the full vector of potential outcomes. This approach would be acceptable in most corporate settings since the consequentialist approach has a transparent methodology that explains how decisions were made.

A deontologist might approach the same scenario differently, strictly adhering to immutable moral rules when deciding how one should behave. A deontological engineer might therefore ask: "What duties or moral principles am I bound to regarding free speech and protection from harm?" They might consider whether they must respect users' free speech rights, which may be violated by false positives, mistakenly classifying a benign comment as a malicious one. A key distinction here is that the deontological engineer optimizes to minimize false positives, while the consequentialist engineer optimizes to minimize false negatives.

Finally, an examination of how a virtue ethicist might approach the scenario. A virtue ethicist might ask: "What would a virtuous engineer do in this situation?" Their decisions may involve self-reflective character virtues such as temperance, avoiding extremes in moderation to ensure that fairness can be demonstrated across all user groups. The virtue ethicist may then reflect upon how their own decisions impact human flourishing, perhaps even concluding that users should be completely free to express themselves authentically without fear of retaliation or being de-platformed.

Each of the above-mentioned approaches can make room for virtues, consequences, and rules. Indeed, any plausible normative ethical theory will have something to say about all three. The takeaway is that there is no "right" way that a moral agent, in this case, the engineer, should conduct themselves and that each normative ethical disposition is simultaneously defensible and open to criticism. Even a hybrid approach, for example, combining ideas from deontology (baseline rules that should not be violated) with consequentialist principles (maximizing the most good for the most people), is subject to the criticism of the virtue ethicist for not being honest and forthcoming about the potential limitations of the algorithm.

Below is a summative table of the three normative approaches, their motivations for ethical reasoning, and the criteria one might use to inspect that reasoning. An important characteristic of this table is that it deliberately simplifies these complex philosophical theories into questions that practitioners can use during decision-making. Rather than requiring deep theoretical knowledge, the table focuses on the core motivation behind each approach, enabling users to consciously choose and apply different normative theories based on the specific ethical situation they face.

Normative Theory	Motivation for Ethical Reasoning	Criteria Used to Examine Ethical Reasoning
Consequentialism	Outcomes	Does my decision produce the maximal amount of happiness, or inversely, create the least amount of unhappiness?
Deontology	Rules	Are my actions in accordance with my own moral rules and principles?
Virtue Ethics	Character	Is my decision motivated by virtue and is my action an instantiation of that virtue?

Principle no. 3: The framework does not recommend what one should believe in, only how to apply one's reasoned set of beliefs to real-world settings.

The framework is not a *moral framework*; a set of rules that instructs how to react in different situations generated by the values, standards, and principles we hold [58]. Moral frameworks are primarily prescriptive systems about how to act, often aligned with normative ethics (theories like deontology, consequentialism, and virtue ethics) that prescribe how one should behave. In contrast, ETHICA does not commit to a particular normative ethical stance. It focuses on methodologies of action derived from whichever normative ethical position one maintains. In this sense, ETHICA focuses less on cultivating a belief system. Instead, it analyzes and addresses ethical problems in particular contexts. Its bias is towards ethical implementation rather than moral prescription, akin to applied ethics.

The supporting argument for this approach is that each business domain is different and, therefore, may subscribe to different ethical principles depending on the nature of the business. For example, a company whose primary business is providing cloud-based computing may decide that sustainability is an ethical rule the company should be governed by, whereas a company that creates AI grant writing software may be chiefly concerned with minimizing malfeasance. To be clear, the ETHICA framework does not promote an "anything goes" morally relativistic approach, the view that moral judgment is culturally relative and that no standpoint is uniquely privileged over all others [59], nor does it promote the idea of moral absolutism, which argues that ethics thus consists in a set of absolute principles that are valid universally, at all times and for all people [60].

Principle no. 4: The framework works across the entire AI development cycle and must remain adaptable across different AI applications and contexts.

The framework provides developers with practical processes to address ethical considerations at each stage of AI development, from business planning through deployment to ongoing maintenance. A key benefit of this stage-based approach is that it modulates ethical considerations to match the specific mindset and workflow of data scientists and machine learning engineers at each development phase. Admittedly, ethical risks are not confined to specific stages. However, they may arise at particular points and should be proactively considered. Below is a table outlining the five major stages of the machine learning development cycle mapped to foreseeable ethical risks.

STAGE	**DESCRIPTION**	**ETHICAL CONSIDERATIONS**
Business Parameters & Planning	The business justification and validation of the potential AI solution, including technical documentation, approval processes, and fulfillment of resource needs (e.g., GPU requirements).	- Privacy - Accountability - Safety
Data Engineering	Identifying and locating the relevant data for the model, making the data available through ETL processes, creating data pipelines, and using first-party and third-party APIs.	- Privacy - Security
Machine Learning Modeling	The iterative process of creating a program that can find patterns or make decisions from a previously unseen dataset.	- Transparency - Explainability - Bias - Fairness
Model Deployment	The process of taking a trained machine learning model from development to production, where it can generate predictions on new data, making it accessible to end-users through interfaces like APIs or applications.	- Accountability - Transparency

Maintenance & Monitoring	The continuous process of tracking an ML model's performance after deployment to ensure it maintains accuracy and effectiveness over time, including "model drift," the misalignment between the data used to train the model and what it encounters in the real world.	- Fairness - Accountability

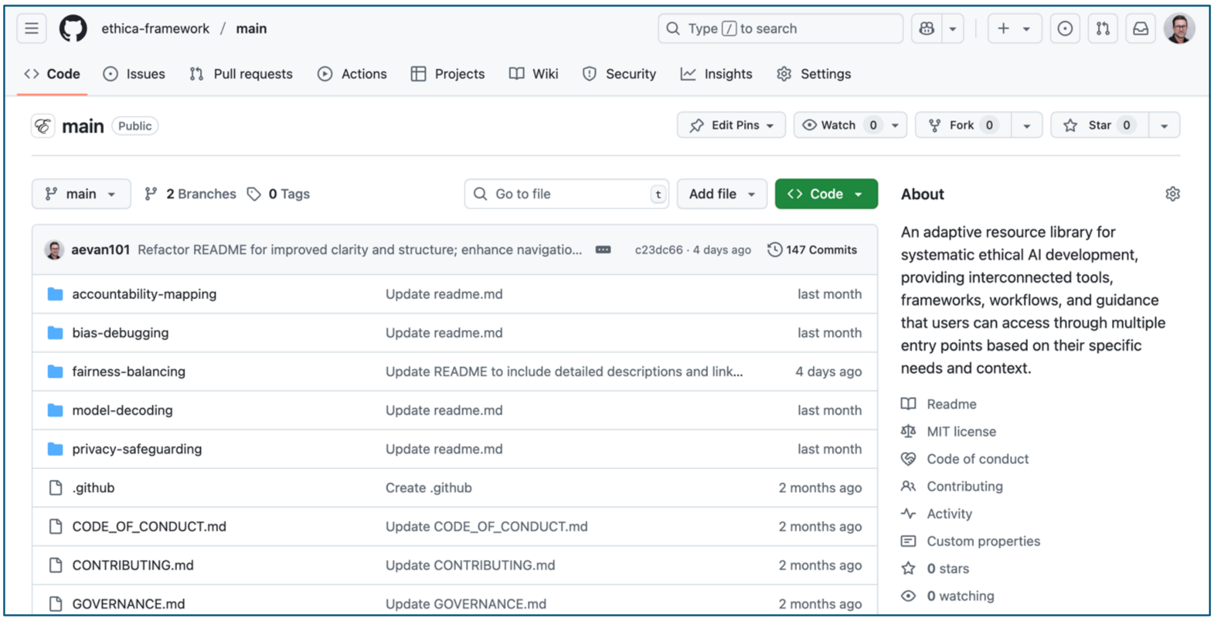
Section 4: ETHICA Framework Components

ETHICA is a collection of learning modules that users can access through an open-source GitHub code repository; the very same coding environment where developers create AI applications. The ETHICA framework should be viewed as an open-source framework whereby its assets can be freely accessed and shared. While these assets are described in this paper, they can also be accessed and explored in a dedicated GitHub Environment.

Access the ETHICA Framework at https://github.com/ethica-framework

GitHub is a universally used web-based platform that allows developers to store, manage, and share their code on projects. One of this paper's contributions is the novel use of GitHub as a means of educating AI developers in ethical reasoning. Both the code and the teachings reside within a single environment, creating an integrated development experience. This approach makes the lessons within ETHICA instantly more accessible compared to traditional academic papers alone, which often live outside of the coding environments where developers of AI technology spend the majority of their time. Nevertheless, while repurposing GitHub code repositories to develop ethical reasoning skills may be considered a novel contribution, reading the code repository's content is not a requirement for understanding the content of this paper, although the GitHub experience is a more authentic representation of ETHICA itself. In other words, one can understand the ETHICA framework solely by reading this paper, without needing to explore the code or the tools in the accompanying GitHub repository.

The structure of the environment itself is relatively straightforward. ETHCA's major practice areas such as Fairness Balancing and Bias Debugging, are organized into individual folders that live with the main directory of the repository along with several other supporting documents commonly found within GitHub code repositories that contain information such as 'CODE_OF_CONDUCT.md' a file describing the behavior that contributes to creating a positive environment and 'CONTRIBUTING.MD' describing how people can contribute code, documentation, examples, or improvements to the framework of the repository itself.



aevan101 Refactor README for improved clarity and structure; enhance navigatio... c23dc66 · 4 days ago 147 Commits

accountability-mapping	Update readme.md	last month
bias-debugging	Update readme.md	last month
fairness-balancing	Update README to include detailed descriptions and link...	4 days ago
model-decoding	Update readme.md	last month
privacy-safeguarding	Update readme.md	last month
.github	Create .github	2 months ago
CODE_OF_CONDUCT.md	Update CODE_OF_CONDUCT.md	2 months ago
CONTRIBUTING.md	Update CONTRIBUTING.md	2 months ago
GOVERNANCE.md	Update GOVERNANCE.md	2 months ago

An adaptive resource library for systematic ethical AI development, providing interconnected tools, frameworks, workflows, and guidance that users can access through multiple entry points based on their specific needs and context.

Fig. 1. Screenshot of 'main' directory of the ETHICA framework within GitHub (https://github.com/ethica-framework/main).

Within each folder is a set of sub-folders that house learning modules such as the *Stakeholder Perspectives* module and the *Rule versus Complexity Diagnostic* module. What are referred to as "modules" are simply Jupyter notebooks, open-source web applications that allows users to create and share documents containing live code, equations, visualizations, and narrative text. When a user runs the notebook, the output generates content the user can interact with. The output is what is referred to as the learning module. A user can run the notebook as many times as they want and even import the notebooks into their preferred integrated development environment (IDE), should they have a preferred coding tool.

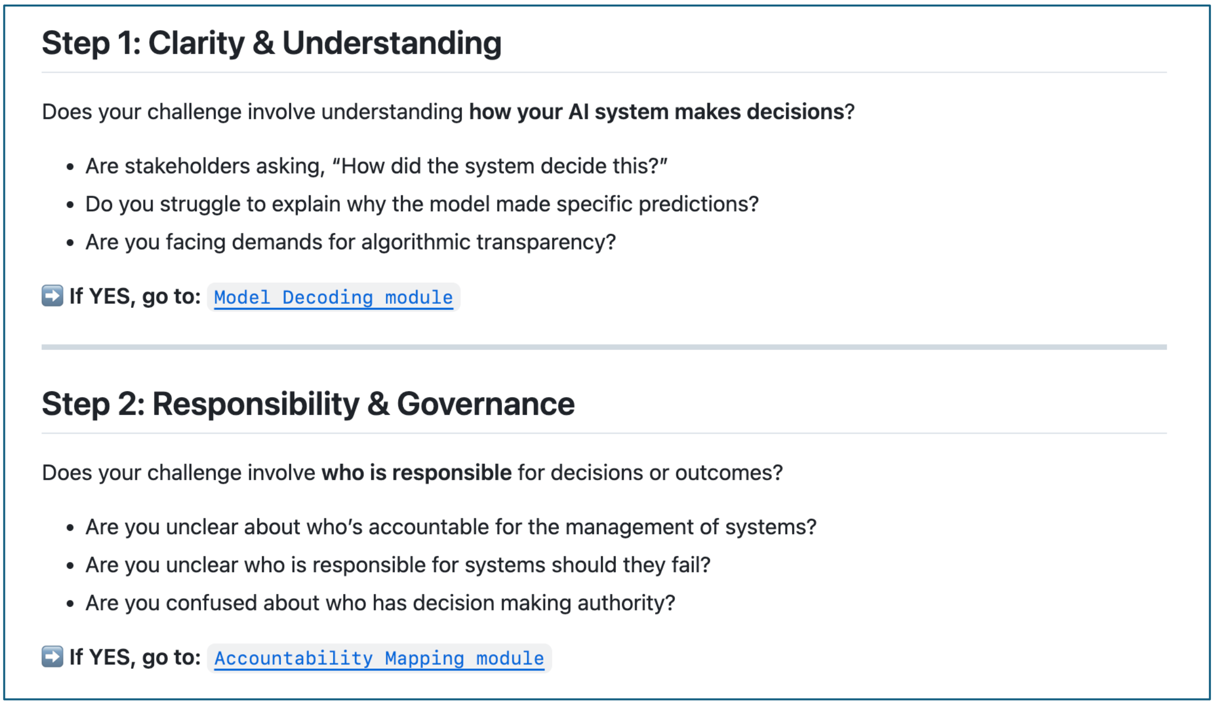
Each learning module is designed to teach a particular ethical reasoning skill. These skills include the ability to draw inferences, reach conclusions, arrive at solutions, and make decisions based on available evidence. More than simply following rules or laws, ethical reasoning is about critically evaluating situations and making choices based on a deeper understanding of moral principles and their implications. At its core, ETHICA involves several interconnected cognitive skills. Analytic skills enable practitioners to break down complex ethical scenarios into components, such as identifying competing values. Synthesizing skills allow them to combine multiple ideas into a single theory. Decision-making skills help practitioners weigh different options and choose a course of action based on their analysis and synthesis. As users face ethical problems in their professional careers, they must have the ability to distinguish between ethical and non-ethical issues and apply appropriate ethical principles to those that are genuinely ethical problems. The more often they do so, the better they become at ethical reasoning [61]. These skills can be developed asynchronously and non-linearly, as ETHICA does not employ a sequential pedagogy.

Importantly, ETHICA develops metacognitive skills, the ability to examine and judge one's own thinking processes. These metacognitive abilities directly correspond to the metaethical skills discussed earlier in this paper: the capacity to understand how moral language and moral thought function, regardless of the contents of anyone's set of moral beliefs. For instance, a practitioner might reflect upon whether they're defaulting to familiar solutions without examining whether those solutions align with their stated ethical commitments or recognize when they're avoiding difficult ethical trade-offs by framing them as purely technical problems. These metaethical skills enable users to examine their own reflexive behavior to ask more self-examining questions: "What normative ethical theory am I consciously or unconsciously applying?" "What assumptions am I making about what constitutes right versus wrong?" "How might my personal disposition shape my prioritization of what matters ethically?" This type of self-reflection is the key to transforming routine compliance behavior into genuine ethical competency by self-examining the ethical ideas that underlie their decisions.

The following is a description of the overall ETHICA framework structure that consists of five main skill development modules and one ethical diagnosis guide. Due to the limitations of this paper, only descriptions of the 'Ethical Diagnosis Guide' and the learning modules associated with the 'Fairness Balancing' practice area are included.

Ethical Diagnosis Guide (https://github.com/ethica-framework/main/blob/main/README.md)

The Ethical Diagnosis guide serves as an aid for guiding users to the relevant skill development module. The module is not intended as a mandatory 'Step 1' of the process. It is meant to route those with vague or abstract concerns or those who are unable to articulate their particular challenge to the appropriate learning module. The aim of the learning modules, the heart of the ETHICA framework, is to provide a system to help users develop the necessary skills to address morally complex challenges. The content of the ethical diagnosis guide is as follows:



Step 1: Clarity & Understanding

Does your challenge involve understanding **how your AI system makes decisions**?

- Are stakeholders asking, "How did the system decide this?"
- Do you struggle to explain why the model made specific predictions?
- Are you facing demands for algorithmic transparency?

➡️ **If YES, go to:** `Model Decoding module`

Step 2: Responsibility & Governance

Does your challenge involve **who is responsible** for decisions or outcomes?

- Are you unclear about who's accountable for the management of systems?
- Are you unclear who is responsible for systems should they fail?
- Are you confused about who has decision making authority?

➡️ **If YES, go to:** `Accountability Mapping module`

Fig. 2. Screenshot of 'AI Ethics Challenge Navigator' guide contained within the README.md file of the main directory.

Are you unsure what kind of ethical challenge you're facing? Do you have ethical concerns but can't articulate them? Do you need help identifying the nature of your dilemma? If you answered YES to any of these questions, refer to the recommended learning module below:

1. Clarity & Understanding: Does your challenge involve understanding HOW your AI system makes decisions?

- Are stakeholders asking, "How did the system decide this?"
- Do you struggle to explain why the model made specific predictions?
- Are you facing demands for algorithmic transparency?

→ If YES, go to Model Decoding module

2. Responsibility & Governance: Does your challenge involve WHO is responsible for decisions or outcomes?

- Are you unclear about who's accountable for the management of systems?
- Are you unclear who is responsible for systems should they fail?
- Are you confused about who has decision making authority?

→ If YES, go to Accountability Mapping module

3. Uncovering Hidden Problems: Does your challenge involve identifying bias in your system?

- Are you unsure if bias is present and want to detect it?
- Do you suspect there may be bias but are unsure where to look or how to test it?
- Do you need to investigate potential discrimination in your model?

→ If YES, go to Bias Debugging module

4: Addressing Unfairness: Does your challenge involve fixing an existing fairness problem you've identified?

- Are you trying to decide on a definition of fairness?
- Do you need some examples of fair versus unfair outcomes?
- Do you need to decide if the problem can be fixed by adding new rules?

→ If YES, go to Fairness Balancing module

5. Data & Privacy: Does your challenge involve HOW you collect, use, store, or protect personal information?

- Are you collecting personal or sensitive information?
- Do users lack control or access over their information?
- Are you facing challenges in compliance or regulatory challenges involving privacy

→ If YES, go to Privacy Safeguarding module

6. Still Unclear? If none of the above categories fit your concern:

- Review all skill modules to see what resonates
- Consider that you may have multiple ethical issues, and that multiple modules may apply
- Submit a request to create another ethical module that may be absent in the current directory

---

Skill Development Modules

The learning modules are the heart of the ETHICA framework. They are designed as self-contained, skill-developing lessons to help users learn how to address real-world problems. Each set of modules is grouped by practice area, such as Fairness Balancing, Bias Debugging, and Accountability Mapping. Practice Areas reflect ETHICA's core philosophy of developing ethics as a practical skillset rather than a prescribed curriculum. Practice Areas emphasize self-directed skill development, where individuals identify the type of ethical challenge they are facing and build competencies to meet it.

This framing supports the development of self-reliant ethical reasoning by allowing practitioners to build capabilities contextually and organically, avoiding the “rules and tools” approach that Responsible AI training often creates.

There are five main practice areas. For the sake of creating a realistic scope for this project, only the ‘Fairness Balancing’ practice area has been fully developed. The other four practice areas contain only a high-level description of their purpose, which can also be found in their corresponding ‘readme.md’ files in GitHub. These additional descriptions are not imperative to comprehend the spirit of the overall framework. While there are currently five, the number of practice areas is not fixed, as new modules are intended to be added over time. Future practice areas could include themes such as environmental sustainability, social impact, and safety. These themes also relate to Responsible AI and represent specific challenges that users may need to navigate.

The current five practice areas are as follows:

1. Accountability Mapping
2. Bias Debugging
3. Fairness Balancing
4. Model Decoding
5. Privacy Safeguarding

Practice areas follow a consistent structure. Each area contains an overview document (readme.md file in GitHub) and a subsequent set of Jupyter notebooks containing the relevant skill-building modules. Each learning module begins with a brief overview statement at the top to help users understand the module’s intention and its relevant lessons. The individual learning modules, although varying in nature, are each designed to develop users’ understanding and skills related to normative ethics, applied ethics, or metaethics, as outlined in Principle 1 of Section 3: Framework Goals & Guiding Principles. Taken together, the modules help users develop ethical reasoning skills and the ability to identify, assess, and develop arguments about right and wrong ethical conduct.

Example Practice Area: Fairness Balancing

As proof of concept, the following represents the content within the Fairness Balancing practice area. Central to each practice area is the idea that ethics without action renders decisions impotent. Therefore, each learning module contains a skill-building activity emphasizing decision-making rather than simply developing theoretical knowledge about a topic. The nature of the modules varies. Some are designed to help students identify and interpret the ethical issues at play (i.e., ethical analysis). Some are designed to encourage users to examine their decisions based on outcomes (i.e., ethical evaluation). Still, others are designed to help users understand ethics from multiple perspectives (i.e., ethical reflection). Each module reinforces a key principle from the Ethical Data Futures course of the Edinburgh Futures Institute that encourages users to move away from the common but limiting idea of ethics as a private, subjective domain of personal value commitments or preferences, and toward an engagement with data ethics as a maturing social, moral and political practice through which more just and sustainable futures are co-constructed with others [62].

Overview (https://github.com/ethica-framework/main/blob/main/fairness-balancing/readme.md):

This practice area develops your capacity to not only reason through fairness trade-offs systematically but also to implement and sustain fair AI systems in practice. Fairness Balancing is the skill of identifying issues that may cause individual or systematic harm and steering decisions towards more fair outcomes. Unlike bias detection (which finds problems), fairness balancing helps you make defensible decisions about resolving those problems and translating those decisions into effective action.

Learning Module No. 3: Ethical Analysis

The *Ethical Analysis module* helps users identify the key complexities present in any given scenario. The module is based on short, original (i.e., non-GEN AI-generated) case studies that users analyze to identify factors such as vulnerable populations, competing interests, knowledge gaps, systemic issues, power dynamics, and conflicting values. After reading the case study, the user is prompted with a series of questions that demonstrate their ethical reasoning ability. Their answers are evaluated using Claude Sonnet 3.5 by Anthropic, a reasoning model invoked from the Jupyter notebook via an API call embedded in the code. The response from the reasoning model critiques the user's reasoning, including suggestions on how to improve their reasoning ability and recommendations on existing business analytic frameworks that may aid in their process— a deliberate decision in support of ethical reasoning if used as scaffolding for making decisions, rather than a formula for how to arrive at them. (See Table 4).

Table 4

A summary of key business tools, classifying each by its primary function. The table outlines the purpose and main steps or methodologies of each tool, which can be used in any of the ETHICA modules.

INDEX	TOOL NAME	TOOLS TYPE	PURPOSE	KEY STEPS/METHODOLOGY
1	Pre-Mortem Analysis	Initiation & Foundational Planning	Proactive risk assessment: identify potential project failures before they occur and develop mitigation strategies.	1. Create project plan. 2. Invite diverse stakeholders. 3. Brainstorm potential failures (individual then group). 4. Prioritize risks by likelihood/severity. 5. Develop mitigation strategies. 6. Review & revise project plan.
2	RACI Matrix	Initiation & Foundational Planning	Clarify team roles and responsibilities across tasks, milestones, or decisions to prevent confusion and increase accountability.	1. Plan project scope & activities. 2. Identify all involved parties. 3. Assign R (Responsible), A (Accountable), C (Consulted), I (Informed) roles for each task/deliverable. 4. Review & confirm roles with team.

3	Fishbone Diagram (Ishikawa)	Problem Diagnosis & Prioritization	Visualize root causes of a problem (RCA) and brainstorm potential actionable changes for quality improvement.	1. Define problem statement (head). 2. Brainstorm major cause categories (bones, e.g., 6Ms). 3. Brainstorm all possible sub-causes (ribs, e.g., 5 Whys). 4. Analyze diagram to identify root causes.
4	Pareto Analysis (80/20 Rule)	Problem Diagnosis & Prioritization	Identify and prioritize the "vital few" causes that contribute to the majority (approx. 80%) of problems or effects.	1. Define problem. 2. Identify causes. 3. Assign value/impact to each cause. 4. Group problems & define measurement. 5. Tally occurrences & calculate percentages. 6. Create bar chart with cumulative percentage line.
5	SWOT Analysis	Initiation & Foundational Planning	Understand internal (Strengths, Weaknesses) and external (Opportunities, Threats) factors impacting a strategy for strategic planning.	1. Brainstorm internal Strengths. 2. Brainstorm internal Weaknesses. 3. Brainstorm external Opportunities. 4. Brainstorm external Threats. 5. Analyze connections to inform strategy.
6	Decision Matrix (Weighted Scoring Model)	Strategic Evaluation & Deliberation	Objectively evaluate and compare multiple options against weighted criteria to make data-driven, transparent choices.	1. Define decision. 2. List options (columns). 3. List key criteria (rows). 4. Assign weights to criteria. 5. Score each option against each criterion. 6. Multiply scores by weights & sum totals. 7. Select highest scoring option.
7	Cost-Benefit Analysis (CBA)	Initiation & Foundational Planning	Evaluate the financial benefits vs. costs of a project/decision to forecast profitability and determine overall value/feasibility.	1. Define project scope. 2. Identify all relevant costs (tangible/intangible). 3. Identify all relevant benefits (tangible/intangible). 4. Quantify costs & benefits (monetary value). 5. Compare (ratio, NPV) & make informed decision.

8	Force Field Analysis	Strategic Evaluation & Deliberation	Systematically evaluate driving forces (pros) and restraining forces (cons) affecting a decision or proposed change.	1. Define decision/change. 2. Identify driving forces. 3. Identify restraining forces. 4. Assign scores to each force (strength). 5. Analyze total scores. 6. Strategize to strengthen driving & reduce restraining forces.
9	Six Thinking Hats	Strategic Evaluation & Deliberation	Structure abstract thinking and enhance creative conversations by approaching decisions from multiple, distinct perspectives.	1. Define problem/situation. 2. Assign "hats" (roles: White, Red, Black, Yellow, Green, Blue). 3. Focus on one hat's perspective at a time. 4. Record ideas for each hat. 5. Blue Hat guides the process.