Wang, D. 2025. “From Engineering Ethics to Ethical Engineering: Leveraging AI for Governing Emerging Technologies” [in English]. *Filosofiya. Zhurnal Vysshey shkoly ekonomiki [Philosophy. Journal of the Higher School of Economics]* 9 (4), 47–67.

Dazhou Wang*

From Engineering Ethics to Ethical Engineering**

Leveraging AI for Governing Emerging Technologies

Submitted: Sept. 16, 2025. **Reviewed:** Nov. 11, 2025. **Accepted:** Nov. 11, 2025.

Abstract: Ethical Engineering (EtEn) is an emerging discipline that represents a paradigm shift from traditional Engineering Ethics (EnEt). Rather than focusing primarily on educating individual practitioners, EtEn aims to systematically embed ethical principles into the very fabric of technological systems and governance processes. This paper examines this fundamental transition from EnEt, which focuses on educating practitioners about normative principles, to EtEn, which treats ethics as a systematic engineering problem, focusing on translation of principles into executable governance tools. The study highlights AI’s dual role as both the primary domain requiring governance and a pivotal enabler for it, examining its potential to enhance ethical governance through improved algorithmic auditability, support for complex ethical decision-making, and cross-domain collaborative governance, while also addressing challenges like value alignment, bias mitigation, and technological reductionism. It identifies eight key issues that constitute the core research agenda for EtEn and argues that its development must be understood as an experimental, iterative process. This paradigm shift not only expands practical pathways for implementing EnEt but also offers novel methodological support for the ethical governance of emerging technology in the age of AI.

Keywords: Engineering Ethics, Ethical Engineering, Artificial Intelligence (AI), Emerging Technology, Value Sensitive Design, Moralizing Technology.

DOI: 10.17323/2587-8719-2025-4-47-67.

1. INTRODUCTION

Technological innovation in artificial intelligence (AI), biotechnology, nanotechnology, and neurotechnology is advancing rapidly, bringing unprecedented opportunities and formidable societal challenges. These technologies are complex, uncertain, and transformative, often stretching traditional ethical and governance mechanisms to their limits. Existing oversight models

*Dazhou Wang, PhD in Philosophy; Professor at the School of Humanities, University of Chinese Academy of Sciences (UCAS) (Beijing, China), dzwang@ucas.ac.cn, ORCID: 0000–0001–9586–4597.

Acknowledgements: This research was funded through the Philosophy and Social Sciences Major Program of Ministry of Education of China (23JZD006).

tend to be reactive and slow, struggling to anticipate, evaluate, and mitigate ethical risks in a timely manner.

For decades, the primary response has been engineering ethics (EnEt), which emphasizes educating engineers and technologists in ethical principles (Harris Jr. et al., 2013; Martin & Schinzinger, 2010). This approach relies on professional codes of conduct, case studies, and individual moral reasoning. While important, it operates mainly at the individual or organizational level, depending heavily on human judgment and self-regulation, which can be subjective, variable, and difficult to scale across global and distributed technology ecosystems. It often functions as an external constraint rather than an integrated component of the engineering process.

Currently, a new paradigm is emerging: Ethical Engineering (EtEn). While several books use the term "Ethical Engineering" (Hersh, ed., 2015; Schlossberger, 2023), they primarily operate within the established paradigm of EnEt, focusing on helping engineers understand and address ethical issues. In contrast, this paper articulates and defends a distinct conception of EtEn as a novel engineering science (Wang, 2023). Its core thesis is that EtEn represents a paradigm shift from advising individual engineers to systematically building ethics into engineering systems and processes, with AI serving as a key enabler for making responsible engineering achievable at scale. This direction is reflected in research exploring how to embed ethical considerations into autonomous systems (Wallach & Allen, 2009), develop tools for ethical risk reflection (Urquhart & Craigon, 2021), and extend value sensitive design (Friedman & Hendry, 2019; Friedman et al., 2013) across the AI lifecycle (Umbrello & van de Poel, 2021). While EnEt asks, "How can engineers behave ethically?", EtEn asks, "How can we design systems, processes, and tools that actively enable ethical outcomes?" This reframing shifts the goal from using moral philosophy to constrain practice to systematically embedding ethical reasoning into the fabric of technological systems and governance structures.

At the heart of this emerging paradigm lies AI — a domain fraught with ethical questions yet also a potent source of solutions to complex governance problems. While many scholars and practitioners focus on the ethical governance of AI development, a compelling strand of research emphasizes AI's capacity to function as an instrument of governance. This is reflected in diverse applications such as managing dual-use technologies (Ulnicane et al., 2023), countering cyber threats (Camacho, 2024), improving regulatory adherence (Jain et al., 2024; Padmanaban, 2024), and combating corrupt practices (Adobor & Yawson, 2023). A growing consensus suggests that AI can provide an integrative framework for navigating the intricate challenges at the intersection of technological innovation and societal evolution.

This paper aims to systematically develop the conceptual foundations of Ethical Engineering as a distinct paradigm. First, it outlines the conceptual transition from EnEt to EtEn, including a clarification of EtEn's unique position relative to value sensitive design (VSD) and the moralizing technology (MT) approach (Verbeek, 2006; 2011). Second, it examines AI's dual role in EtEn — as both its primary testbed and most powerful accelerator — and analyzes the strategy of "using AI to govern AI." Third, it identifies eight key challenges that define the research agenda for EtEn. Finally, it concludes by emphasizing the need for a reflexive, experimentalist approach to developing EtEn as a socio-technical governance system, one that requires interdisciplinary collaboration and cross-cultural sensitivity.

2. FROM ENGINEERING ETHICS TO ETHICAL ENGINEERING: A PARADIGM SHIFT

The field of technology ethics is undergoing a significant evolution, marked by a transition from the established framework of EnEt to the emerging paradigm of EtEn. This shift, facilitated by intermediary approaches like VSD and the theory of MT, represents a move from principle-based guidance toward system-based implementation.

2.1. CHARACTERISTICS OF ENGINEERING ETHICS

EnEt has long been an established field dedicated to addressing the moral obligations and dilemmas that engineers encounter in their practices. At its core, it is built upon several key tenets. Normative principles form the foundation of EnEt. Organizations such as the National Society of Professional Engineers (NSPE) and the Institute of Electrical and Electronics Engineers (IEEE) have developed codes of ethics that emphasize public health, safety, and welfare. Philosophical frameworks like utilitarianism and deontology also play a crucial role in guiding ethical reasoning and decision-making. While utilitarianism focuses on maximizing overall well-being, deontology emphasizes adherence to moral rules. More recently, the importance of virtue ethics is gaining increasing recognition in the field of engineering ethics.

Education and individual agency are also central to EnEt. It places strong emphasis on teaching students and professionals how to recognize ethical issues, analyze complex dilemmas, and make morally sound decisions. Through courses and training programs, engineers are equipped with the necessary skills to cope with the ethical challenges they may face in their careers.

Another characteristic of traditional EnEt is its case-based and reactive nature. Historical cases, such as the Challenger disaster, are often used as teaching tools (Elliot et al., 1993). These cases provide valuable lessons about the consequences of unethical practices. However, this approach is retrospective, as it primarily analyzes past events in order to prevent similar mistakes in the future.

EnEt also functions as an advisory and constraining framework. It sets out a set of rules and guidelines that engineers must follow to ensure that their work remains within socially acceptable boundaries. It acts as an external force that guides engineering practice, but it does not necessarily provide a comprehensive solution to all modern engineering challenges.

Despite its importance, the traditional paradigm of EnEt faces several modern challenges. One of the main issues is limited scalability. As technology develops at an ever-increasing pace, it becomes difficult to apply a one-size-fits-all set of ethical principles to a wide range of engineering projects. Weak enforcement is another problem. There is often a lack of strict mechanisms to ensure that engineers adhere to ethical guidelines. Moreover, the fast-paced nature of agile software and technology development is not well-matched with the relatively slow and static nature of traditional EnEt.

2.2. CHARACTERISTICS OF ETHICAL ENGINEERING

In contrast, EtEn is an emerging paradigm that builds on EnEt but differs from it in fundamental ways. It constitutes a distinct engineering discipline focused on designing systems that inherently facilitate ethical outcomes. EtEn is proactive and procedural. It seeks to integrate ethics into every stage of the development lifecycle, from research and development to deployment and decommissioning. By incorporating tools, processes, and checkpoints at each stage, it ensures that ethical considerations are not an afterthought but an integral part of the engineering process.

This new paradigm is also system-oriented. It considers the entire socio-technical system, including the interactions among humans, technology, and institutions. Instead of focusing solely on the actions of individual engineers, it looks at how the entire system functions and how ethical issues can arise from these complex interactions.

As is well known, a core challenge in EtEn is the technical translation of values. Abstract ethical principles such as fairness, transparency, and accountability need to be translated into concrete, measurable technical requirements (Wang, 2020). For example, ensuring fairness in an algorithm may require developing specific metrics and techniques to detect and correct biases. EtEn is enabling and empowering. Rather than merely constraining the behavior of engineers, it provides tools and resources that empower developers and governance bodies. Bias detection systems can help identify and mitigate biases in algorithms, while ethical checklists can guide engineers through the ethical decision-making process. These tools allow engineers to build more trustworthy products and services.

IBM has provided a specific example of EtEn. AI Fairness 360, an open-source toolkit developed by IBM, provides a standardized "toolbox" containing dozens of algorithmic fairness metrics (such as demographic parity and equalized odds) and bias mitigation algorithms (such as reweighting and adversarial debiasing). Engineers can seamlessly integrate AIF360 into their machine learning workflows by simply importing it like any other library. They can then use different metrics to calculate their model's fairness scores across various demographic groups (such as different genders or races), and experiment with different debiasing algorithms to determine which method most effectively enhances fairness while maintaining model accuracy.

This shift from EnEt to EtEn can be compared to the introduction of quality assurance (QA) in manufacturing (Feigenbaum, 2012). Just as QA introduced systematic processes to ensure that quality was built into products, EtEn aims to systematically integrate ethics into engineering system. It represents a new way of thinking about engineering, one that is more proactive, comprehensive, and better suited to the challenges of the twenty-first century.

2.3. COMPARISON BETWEEN ENGINEERING ETHICS AND ETHICAL ENGINEERING

The evolution from EnEt to EtEn signifies a shift from principle-based guidance for individuals to the systematic, engineering-based implementation within systems (Table 1). This transformation turns ethical considerations from an external constraint into an intrinsic element of technological design and development. This transition represents a necessary paradigm shift for addressing the limitations of traditional EnEt in the face of modern engineering's complexity and pace. EtEn offers a more holistic and proactive solution, enabling engineers to build technologies that are not only innovative but also ethical and trustworthy. As such, it defines a new frontier for the engineering sciences.

Aspect	Engineering Ethics	Ethical Engineering
Core Focus	Moral obligations and decision-making of individual engineers	Building technological systems and processes that facilitate ethical outcomes
Methodology	Principle-oriented: Based on normative ethical principles (e. g., utilitarianism, deontology) and professional codes of conduct.	System-oriented: Treats ethics as an engineering problem, implemented through designed processes, tools, and system specifications.
Primary Goal	Education and practice constraints: Cultivate engineers' moral reasoning abilities and constrain their behavior within socially acceptable boundaries.	Systematic implementation and empowerment: "Hardwire" ethics into the entire development lifecycle, providing developers with tools and methods to achieve ethical goals.
Temporal Orientation	Post-hoc reflection and reactive: Primarily involves teaching and reflection through the analysis of historical cases.	Proactive and forward-looking: Integrate ethical considerations from the initial R&D phase through deployment and decommissioning.
Core Challenge	Addressing ethical dilemmas faced by individuals, emphasizing personal professional responsibility.	The "translation problem": Converting abstract ethical principles (e. g., fairness, transparency) into concrete, measurable, and implementable technical requirements and system specifications.
Mechanism of Action	Acts as an external constraint: Functions as rules and guidelines that constrain engineering practice.	Acts as an internal enabler: Creates tools and processes (e. g., ethics checklists, bias detection tools) to actively empower developers and governance bodies.
Level of Concern	Individual level: Focuses on the agency and responsibility of the engineer.	System level: Concerns the entire socio-technical system of interactions between people, technology, and institutions.

Table 1. Engineering Ethics vs. Ethical Engineering

2.4. VALUE SENSITIVE DESIGN AND MORALIZING TECHNOLOGY AS BRIDGING CONCEPTS

The transition from EnEt to EtEn is logically and methodologically facilitated by two key frameworks: MT and VSD. They serve as indispensable conceptual and methodological bridges, respectively.

EnEt, while essential, often operates reactively, focusing on educating individual engineers to reason about dilemmas using normative principles and historical cases. However, it struggles to provide scalable, procedural methodologies for integrating ethics into modern technological development, which is increasingly fast-paced and complex. This gap is precisely where VSD and MT intervene, paving the way for EtEn.

Logically speaking, the theory of MT provides the essential philosophical bridge. It challenges the view of technology as a neutral tool by arguing that artifacts actively shape moral decisions and behaviors, suggesting they can even exhibit a form of "moral agency." This perspective fundamentally expands the scope of ethical concern from the individual engineer to the technology itself and its socio-technical context. It provides the fundamental "why" for EtEn: because technology shapes morality, we must consciously design that influence.

Concurrently, VSD provides the crucial methodological bridge. It moves beyond abstract deliberation by offering a structured, tripartite methodology (conceptual, empirical, and technical investigations) for proactively identifying and embedding human values into technical design. This process-oriented framework is a direct precursor to EtEn, which seeks to systematize and operationalize such processes across the entire development lifecycle. VSD translates values such as privacy and fairness into tangible design requirements, directly addressing the "translation problem" that lies at the center of EtEn.

Thus, we can conceptualize their relationship as a continuum from philosophical foundation to technical execution: MT (Philosophical Layer) → Value Sensitive Design (Methodological Layer) → EtEn (Engineering & Implementation Layer) (Table 2). MT justifies the need for EtEn; VSD provides a key methodological process for it; and EtEn is the engineering discipline that systematizes, executes and governs the integration of their insights. For example, VSD, through stakeholder analysis, can determine what "fairness" should mean in a specific context, and EtEn is responsible for technically implementing this defined "fairness" through algorithms and system architectures.

Dimension	Moralizing Technology	Value Sensitive Design	Ethical Engineering
Theoretical Level	Philosophical Foundation	Design Methodology	Engineering discipline & Technical Implementation
Core Focus	Revealing the moral relationship between technology and humans, explaining how technological artifacts mediate and shape human moral perception, decision-making, and behavior.	Providing a systematic design process to proactively incorporate human values into technology design and development.	Establishing a systematic engineering discipline for translating ethical principles into concrete, computable rules and algorithms embedded in technological systems and processes.
Central Question	Do technological artifacts themselves have moral significance? How do they actively influence morality?	How should we systematically consider and embed human values in design?	How should we systematically build and govern technical systems to ensure they produce ethical outcomes?
Theoretical Aim	Descriptive & Explanatory: Describing and explaining the phenomena and mechanisms of technological mediation effects.	Prescriptive: Providing a guiding framework and tools for “what should be done.”	Constructive & systematic: Providing the principles, tools, and standards for “how to systematically build and govern” ethical technology.
Understanding of “Ethics”	Ethics is a relational product emerging from human-technology interaction. Morality is “materialized” in the materiality of technology.	Ethics refers to human-centric values that need to be identified and coordinated (e. g., privacy, fairness, well-being).	Ethics is a system property achievable through engineering practices, involving formalizable, operationalizable rules and constraints.
Primary Methods	Philosophical speculation, case studies (e. g., speed bumps, ultrasound machines).	Tripartite iterative investigations (conceptual, empirical, technical); stakeholder analysis.	Formal methods, algorithm design, verification and validation, safety engineering, standard setting.
Focus on Agency	Emphasizes the “moral agency” of technology itself (non-human intentional influence achieved through function).	Emphasizes the agency of humans (designers, stakeholders), who actively make value choices and embeddings.	Emphasizes the agency of the engineering system and process to reliably achieve ethical outcomes.

continued on next page

continued from previous page

Dimension	Moralizing Technology	Value Sensitive Design	Ethical Engineering
Temporal Orientation	Reflective & Prospective: Analyzes existing technologies, and its insights also guide future design.	Proactive: Primarily applied at the beginning and throughout the technology design process.	Proactive & Concomitant: Implemented during the design phase and continuously executed during system operation.
Typical Applications	Explaining how social media mediates interpersonal relationships; explaining how autonomous vehicles alter concepts of responsibility.	Designing privacy-respecting web browser cookie notices; designing urban planning software that supports democratic deliberation.	Designing the full stack of governance for an AI system, from ethical checklists and bias detection in development to real monitoring and audit trails in deployment.

Table 2. Comparison between Value Sensitive Design, Moralizing Technology, and Ethical Engineering

In summary, MT and VSD are not superseded by EtEn but are foundational to it. A robust approach to the ethical governance of technology requires their integration: beginning with the perspective of MT to recognize the moral influence of technology; proceeding to the methodology of VSD to organize and guide the design process; and culminating in the systematic engineering discipline of EtEn to achieve the defined ethical goals reliably and at scale.

3. ARTIFICIAL INTELLIGENCE: THE "ACCELERATOR" AND "PROVING GROUND" OF ETHICAL ENGINEERING

AI, particularly Machine Learning (ML) and Deep Learning (DL), possesses unique technical characteristics that make it not only the most critical application domain for EtEn but also a core catalyst and enabling tool that drives the maturation of its methodologies and sharpens the precision of its theories.

3.1. ARTIFICIAL INTELLIGENCE AS THE "ACCELERATOR" OF ETHICAL ENGINEERING

As an enabling tool, AI technology can significantly enhance the effectiveness, precision, and scale of the EtEn toolbox, propelling it from manual, qualitative analysis into a new stage of automation, quantification, and systematization. AI-driven static and dynamic analysis tools can automatically scan codebases and training datasets to identify potential patterns of bias, security vulnerabilities, or privacy-leakage risks, thereby overcoming the limitations of traditional manual code reviews when dealing with complex systems comprising millions of lines of code. For instance, tools can automatically detect representational biases in datasets or evaluate a model's performance disparities across different demographic subgroups, providing engineers with quantified fairness reports.

Furthermore, EtEn emphasizes the prospective assessment of technological consequences. AI-Enhanced Agent-Based Simulation can construct highly complex virtual social environments to deploy and test algorithms or systems, observing the emergent behaviors and long-term ripple effects generated through interactions with vast numbers of simulated users. This allows engineers to "rehearse" potential unintended ethical consequences of a system, such as the formation of information cocoons, market manipulation, or the exacerbation of social discrimination, at a lower cost before real-world deployment.

Realizing "value alignment" also presents a major challenge: how to extract actionable design inputs from diverse and sometimes conflicting human preferences (Gabriel, 2020). AI techniques, particularly Inverse Reinforcement Learning and advanced interview analysis, can help systematically learn and infer underlying value preferences from human behavior, decisions, or feedback, formalizing them into reward functions or constraints, thereby offering a data-driven engineering path for value alignment. Explainable AI (XAI) is not merely a goal for enhancing model transparency; it is itself a crucial tool for implementing EtEn (Adadi & Berrada, 2018; Dwivedi et al., 2023). Only when a system's decision-making process can be explained and traced can engineers and auditors effectively diagnose it for unfairness, discrimination, or logical errors. For instance, SHAP (SHapley Additive exPlanations) is a leading technique in XAI that explains the output of any machine learning model by quantifying the contribution of each input feature to a single prediction. As a proto-ethical-engineering tool, SHAP has the potential to operationalize the ethical principle of transparency, provide the foundational capability for auditability, and ultimately transforms an opaque "black box" model into a system that can be interrogated and understood. Thus, developing and integrating XAI tools is a core part of building trustworthy, auditable AI systems and an indispensable component of EtEn methodology.

Additionally, ethical norms and social standards are constantly evolving. AI systems can be used for the continuous monitoring of deployed products, analyzing user feedback, public discourse, and operational data in real-time to automatically detect whether their behavior is beginning to deviate from established ethical guidelines or newly enacted laws and regulations, enabling dynamic compliance management and early warning.

3.2. ARTIFICIAL INTELLIGENCE AS THE "PROVING GROUND" OF ETHICAL ENGINEERING

As the core object, the complexity, uncertainty, and autonomy of AI technology pose unprecedented challenges for EtEn, which in turn powerfully drives the discipline's deepening and maturation. It forces the mathematization and operationalization of ethical principles.

Traditional EnEt often deals with principled but vague concepts. However, confronted with AI algorithms, we must provide mathematical definitions for concepts like "fairness": Is it equality of opportunity or equality of predictive outcomes? This pressing engineering necessity forces philosophers, legal scholars, and social scientists to collaborate with engineers in translating abstract ethical concepts into computable, optimizable, and trade-off-able engineering metrics. Without the challenge posed by AI, the refined discussions of "fairness" and "accountability" within EtEn would not have reached their current depth.

AI also highlights the importance of systematic ethics. AI systems are typically not isolated models but components embedded within vast sociotechnical systems. Their ethical impact is often not determined by a single algorithm, but is an emergent property arising from the interaction of multiple stages: data collection, feature engineering, model training, deployment environment, and user interaction. This forces EtEn to develop a system-level perspective and methodology, requiring ethical assessment and governance across the entire system lifecycle rather than focusing solely on the "materialization" during the design phase.

Moreover, application of AI has spawned an urgent global need for AI governance, directly promoting progress in EtEn regarding standard development, certification processes, and the creation of audit tool. AI acts as a "stress test," examining and accelerating EtEn's evolution from corporate self-regulation toward industry regulation and societal governance. As a good example, IEEE has launched the renowned IEEE 7000$^{\text{TM}}$ series of standards, particularly "IEEE 7000–2021: Standard Model Process for Addressing Ethical Concerns during Systems Design." It provides a concrete methodology for directly integrating ethical considerations into systems engineering processes, requiring the identification and managment of ethical risks early in the design phase.

AI and EtEn are in a profound symbiotic relationship. On one hand, AI is the most severe challenge and the most important "proving ground" for EtEn. With its extreme complexity and social impact, it exposes the shortcomings of traditional ethical thinking and urgently demands a rigorous engineering solution, thereby powerfully driving the emergence and development of EtEn. On the other hand, AI technology is also the most powerful "enabler" for EtEn. It provides powerful tools such as automated analysis, large-scale simulation, and preference learning, making the systematic implementation of ethical design, assessment, and monitoring possible and thereby promoting the implementation and evolution of EtEn methodologies.

Therefore, AI is both the primary object of governance and the core means for achieving the governance objectives. Such dual nature makes the field of AI the most active, cutting-edge, and methodology-intensive area for EtEn thought and practice. Advancing AI ethics is, in essence, the practice of building and developing EtEn itself. The two complement each other, working together toward the core goal of ensuring that those increasingly autonomous, powerful, and ubiquitous technologies can robustly, reliably, and responsibly serve human well-being.

3.3. ENHANCING ETHICAL ENGINEERING BY USING AI TO GOVERN AI

The accelerating sophistication and integration of AI across social and economic systems has ushered in an era of unprecedented potential, yet it also introduces profound ethical and governance challenges. Traditional oversight mechanisms, often reliant on slow and subjective human review, are increasingly inadequate for regulating autonomous, large-scale, and rapidly evolving AI systems. In response, a promising yet complex paradigm is drawing attention: the idea of using AI itself to govern AI. This approach represents a fundamental shift within EtEn, moving away from external, intermittent checks toward embedded, continuous, and automated oversight. By leveraging AI's capabilities to monitor, evaluate, and even correct other AI systems, we introduce a layer of reflexivity, a capacity for self-awareness and adaptation, that is essential for managing the ethical risks of advanced technologies (Gou et al., 2023; Madaan et al., 2023; Collin et al., 2023).

The appeal of AI-driven governance lies in its ability to operate at the speed and scale of the systems it monitors. Unlike human committees, AI supervisors can analyze millions of decisions in real time, detect subtle patterns of bias or malfunction, and respond instantaneously to deviations. This enables a shift from post-hoc auditing to proactive ethical assurance. Core to this approach are several technical pathways that embody this reflexivity. One is Ethics-by-Design, which involves formalizing ethical principles into computational metrics, such as fairness definitions or privacy constraints, that can be built directly into AI architectures. Another is real-time monitoring through multi-agent systems, where guardian AI agents observe the behavior of primary models, flagging anomalies such as discriminatory outputs or performance decay. Furthermore, these systems can enable dynamic self-correction, allowing AI to adjust its own operations in response to ethical breaches. Simulation tools add another dimension, permitting the forecasting of long-term societal impacts before deployment, while blockchain-based audit trails create immutable records for accountability.

However, using AI to govern AI is not a straightforward solution. It introduces a series of deep and potentially recursive ethical and technical complications. The most fundamental is the meta-ethical dilemma: Who decides which values embedded in the governance AI? Ethical norms vary across cultures and jurisdictions, and encoding a single universal standard risks cultural imposition or ethical simplification, a challenge that echoes the value operationalization difficulties highlighted in the previous section. Moreover, AI systems today lack the nuanced understanding required to interpret context-rich moral dilemmas; their strength lies in quantifying metrics, not in interpreting philosophical nuance. This technical limitation becomes especially salient in edge cases, where rigid rules may fail. Additionally, governance systems are themselves vulnerable to adversarial attacks because malicious actors may manipulate supervision mechanisms, bypass safeguards, or poison the training data of the guardian AI. Beyond these risks, there are practical barriers related to standardization and cost. Without interoperable frameworks and shared standards, AI governance may remain fragmented across regions and industries. Meanwhile, the high expense of developing advanced oversight tools could exclude smaller entities, widening the gap between ethical haves and have-nots.

Looking ahead, the future of reflexive AI governance will depend on coordinated efforts across multiple domains. Critical to this effort is the development of open-source tools and benchmarks that make ethical oversight more accessible and reproducible. Equally important is sustained interdisciplinary collaboration through which ethicists, social scientists, engineers, and policymakers must work together to ensure that governance systems are both technically robust and socially legitimate. An incremental implementation pathway is advisable, one that retains meaningful human oversight, particularly for high-stakes decisions, while gradually introducing greater automation as the technology proves reliable. Finally, international coordination will be essential to avoid regulatory fragmentation and to foster alignment on core ethical principles, even as technical approaches may vary.

4. KEY ISSUES THAT ETHICAL ENGINEERING MUST FOCUS ON

The preceding discussion of using AI to govern AI underscores a fundamental characteristic of EtEn: its inherently experimental nature. The idea that EtEn could provide definitive, universal solutions is a misconception. It is better understood as a continuous process of experimentation between technology and ethics (Wang, 2018; Van de Poel, 2020). Through iterative learning and adaptation, this process gradually aligns technological development with ethical principles and human well-being. However, as an emerging field, its path of development is far from smooth; a series of profound and complex core issues urgently require exploration and resolution. The extent to which these problems are solved will directly determine whether EtEn can transition from a theoretical concept to a mature practice, truly fulfilling its mission of shaping responsible technology. These challenges constitute the core research agenda for EtEn as a discipline.

First, the Problem of Value Operationalization and Quantification. Ethical values such as “fairness,” “privacy,” “autonomy,” and “security” are inherently abstract, qualitative, and highly context-dependent. The primary task of EtEn is to provide engineers with a set of methods to transform these “soft” values into “hard” technical parameters that can be understood, encoded, measured, tested, and optimized. This translation poses significant challenges, as illustrated by the concept of “fairness”: the field of algorithms offers dozens of mathematical definitions, such as statistical parity, equality of opportunity, and individual fairness, each carrying distinct philosophical assumptions and legal implications. The critical question then becomes which definition should be selected for a given context, and on what normative grounds? Similarly, operationalizing “privacy” requires designing computable metrics, akin to a “loss function,” that can be balanced against other optimization goals like accuracy and latency. These challenges are further compounded when deploying technologies globally, where divergent cultural norms and legal frameworks demand adaptable implementations of core values.

Second, the Problem of Ethical Emergence in Complex Systems. The ethical issues of modern technological systems (e. g., smart cities, platform ecosystems, the Internet of Things) are often emergent properties arising from the interactions among system components, rather than simple summations of the attributes of individual algorithms or modules. This leads to significant difficulties in prediction and governance (Brey, 2012). A typical problem is "compound unfairness": a fair recommendation algorithm combined with a fair pricing algorithm may still produce systematic price discrimination or service exclusion against a particular group across the entire ecosystem. Such long-tail, cross-domain chain reactions cannot be fully anticipated at the design stage. Therefore, EtEn must move beyond the moralization of single technologies and develop theories and tools for system-level ethical simulation, real-time monitoring, and intervention. How to build digital twin environments that can simulate complex human-computer interactions to predict the ethical risks of technologies deployed in society will be a key focus of future research.

Third, the Lack of Value Trade-off and Decision-Making Frameworks. When fundamental conflicts arise between "accuracy" and "fairness," "efficiency" and "privacy," or "safety" and "autonomy," what framework should engineers use to make decisions? This is essentially a value judgment, yet it is an unavoidable daily issue in engineering practice. For example, to improve the safety of an autonomous driving system (protecting pedestrians), is it acceptable to sacrifice some passenger privacy (through more intensive in-car monitoring)? Who should have the authority to make this decision? The engineers, the company, regulators, or the public? EtEn cannot merely provide a set of potentially conflicting tools; it must develop structured, procedural frameworks for trade-offs and decision-making. This may include consensus-based prioritization, public participation mechanisms based on democratic deliberation, or clear legal rules. The absence of such frameworks places engineers under tremendous moral and professional risk.

Fourth, the Challenge of Auditability and Accountability. Whether a system is ethical cannot be determined solely by the developers' self-certification; it must be verified through independent, repeatable audits. However, there is currently a lack of widely accepted algorithmic audit standards, tools, and professional audit teams. Technical black boxes (especially deep learning models) make auditing exceptionally difficult. EtEn must promote the development of XAI tools and use them as the foundation for ethical audits. Simultaneously, legal accountability needs clarification: When an accident occurs, how is responsibility allocated among designers, developers, deployers, users, and even the algorithm itself? Establishing a clear chain from technical traceability to legal accountability is the institutional guarantee for the implementation of EtEn.

Fifth, the Boundaries and Risks of Standardization. Standardization is a cornerstone of engineering, yet the development of ethical standards presents a distinct double-edged challenge. While providing crucial guidance and stability for industry adoption, an overtly rigid or minimalist approach to standardization risks fostering "checkbox-ticking compliance," where meeting minimum requirements becomes the endpoint, inadvertently stifling moral imagination and superior ethical practices that exceed baseline norm. To navigate this tension, EtEn must actively advocate for and contribute to the design of dynamic, process-oriented standards. These should function less as static checklists and more as evolving frameworks that mandate continuous improvement.

Sixth, Cross-Cultural Global Governance Coordination. Technology is borderless, whereas values are regional. The differing ideas of China, the US, and Europe regarding data privacy, freedom of speech, and social governance have led to divergent paths in technological governance. The development of EtEn must confront the challenge of building global governance coordination. A completely fragmented governance system could lead to "ethical protectionism" and "regulatory arbitrage" (i.e., developing and testing in jurisdictions with the loosest standards), while imposing uniformity would ignore legitimate cultural diversity. Therefore, the key challenge for the future is how to form global consensus on core issues that prohibit a "race to the bottom" (e. g., lethal autonomous weapons), while exploring cooperative mechanisms like mutual recognition of certifications and cross-border data flows in other areas, seeking minimal consensus based on respect for diversity.

Seventh, Advanced AI Alignment and the Risk of Ethical Loss of Control. Facing the potential future emergence of superintelligence (AGI, Artificial General Intelligence), EtEn encounters its ultimate challenge — the Alignment Problem: How to ensure that the ultimate goals of an AI system far surpassing human capabilities remain fully aligned with humanity's complex, ambiguous, and dynamically evolving values? (Ji et al., 2023) This goes far beyond the current scope of algorithmic fairness, involving the internal modeling and calibration of motives, intentions, and values. Failure could be existential. EtEn must begin to prospectively consider these "long-term future" problems.

Eighth, The Ethics of Ethical Engineering Itself. Finally, we must maintain critical reflection on EtEn itself. This powerful "hammer" could also be misused. Who decides which ethics are to be "materialized"? Could it become a tool of techno-authoritarianism, enabling social control and value indoctrination through technological design under the guise of "being good for you"? Therefore, the development process of EtEn itself must be transparent, democratic, and responsible. It must incorporate a mechanism for self-criticism and self-correction, remaining vigilant against the risks of alienation it may introduce, and ensuring that it ultimately serves human well-being and social prosperity, not the specific interests of any single group.

These eight key challenges are both severe tests and defining opportunities for EtEn. They necessitate the development of what may be termed "hybrid knowledge," a deeply integrated, interdisciplinary framework where philosophers and ethicists clarify normative foundations, social scientists map contextualized value interpretations, and engineers co-develop corresponding formalisms and technical implementations— especially translating abstract ethical principles into concrete technical realizations through algorithmic formalization and system design. Only through such cross-boundary concerted efforts can EtEn overcome its numerous obstacles and evolve from a promising concept into a mature discipline and practical system capable of reliably guiding the course of technological development and ensuring it benefits humanity. The success of this endeavor concerns not only the fate of one discipline but the future of us all.

5. CONCLUSION

The growing complexity and pervasiveness of emerging technologies underscore the urgency of embedding values such as fairness, accountability, and transparency into engineering practice. Yet significant challenges remain, including a shortage of practical tools, limited interdisciplinary collaboration, and underdeveloped methodologies for ethical evaluation. These gaps highlight the need for a coordinated international effort to advance EtEn as a novel engineering discipline. The transition from EnEt to EtEn represents a necessary evolution in the governance of emerging technologies. While educating ethical practitioners remains essential, it is no longer sufficient. This shift demands systematic approaches that integrate ethical considerations directly into technological design and development.

AI is central to this transformation. Tools such as XAI, ethics-embedded algorithms, and predictive risk models enable the operationalization of ethics, making governance more scalable and proactive. However, technical solutions alone cannot address persistent challenges such as value alignment and algorithmic bias. A holistic approach is required — one that aligns with Fisher's (Fisher, 2019) concept of "socio-technical governance," which merges technical capabilities with robust social oversight, transcending purely technical or exclusively social models. Crucially, the vision of "using AI to govern AI" must be implemented within a human-in-the-loop framework, where automation augments rather than replaces democratic deliberation and human judgement. Ultimately, the goal is to couple advanced technological tools with deepened democratic engagement, forming an adaptive and reflective EtEn capable of addressing the ethical, legal, and social implications of emerging technologies. This paradigm not only expands practical pathways for implementing EnEt but also promotes responsible innovation in the age of AI.

It should be emphasized that EtEn is not a panacea but a socio-technical approach whose success depends on addressing its limitations through continued technical refinement and robust human oversight, and that effective EtEn depends on sustained collaboration among educators, researchers, engineers, policymakers, and civil society. Philosophers and ethicists should play a key role in propelling the development of ethical engineering rather than remaining in an ivory tower of pure discourse. Only through shared commitment and cross-sectoral dialogue can we ensure that technological advancement serves humanity's best interests, promoting not only economic growth but also equity, dignity, and collective well-being for generations to come.

REFERENCES

Adadi, A., and M. Berrada. 2018. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* 6:52138–52160.

Adobor, H., and R. Yawson. 2023. "The Promise of Artificial Intelligence in Combating Public Corruption in the Emerging Economies: A Conceptual Framework." *Science and Public Policy* 50:355–370.

Brey, P. 2012. "Anticipatory Ethics for Emerging Technologies." *NanoEthics* 6 (1): 1–13.

Camacho, N. G. 2024. "The Role of AI in Cybersecurity: Addressing Threats in the Digital Age." *Journal of Artificial Intelligence General Science* 3 (1): 143–154.

Collin, B., P. Izmailov, J. H. Kirchner, et al. 2023. "Weak-to-Strong Generalization: Eliciting Strong Capabilities with Weak Supervision." arXiv. Accessed June 1, 2025. https://arxiv.org/abs/2312.09390.

Dwivedi, R., D. Dave, H. Naik, et al. 2023. “Explainable AI (XAI): Core Ideas, Techniques, and Solutions.” *ACM Computing Surveys* 55 (9).

Elliot, N., E. Katz, and R. Lynch. 1993. “The Challenger Tragedy: A Case Study in Organizational Communication and Professional Ethics.” *Business & Professional Ethics Journal* 12 (2): 91–108.

Feigenbaum, A. V. 2012. *Total Quality Control.* 4th ed. New York: McGraw-Hill.

Fisher, E. 2019. “Governing with Ambivalence: The Tentative Origins of Socio-Technical Integration.” *Research Policy* 48 (5): 1138–1149.

Friedman, B., and D. G. Hendry. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination.* Cambridge (MA): The MIT Press.

Friedman, B., P. H. Kahn, A. Borning, and A. Huldtgren. 2013. “Value Sensitive Design and Information Systems.” In *Early Engagement and New Technologies : Opening up the Laboratory*, ed. by N. Doorn, D. Schuurbiers, I. van de Poel, and M. Gorman, 55–95. Dordrecht: Springer.

Gabriel, I. 2020. “Artificial Intelligence, Values, and Alignment.” *Minds & Machines* 30 (3): 411–437.

Gou, Z., Z. Shao, Y. Gong, et al. 2023. “CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing.” arXiv. Accessed June 1, 2025. https://arxiv.org/abs/2305.11738.

Harris Jr., C. E., M. S. Pritchard, and M. J. Rabins. 2013. *Engineering Ethics: Concepts and Cases.* Wadsworth: Cengage Learning.

Hersh, M., ed. 2015. *Ethical Engineering for International Development and Environmental Sustainability.* London: Springer.

Jain, V., A. Balakrishnan, D. Beeram, et al. 2024. “Leveraging Artificial Intelligence for Enhancing Regulatory Compliance in the Financial Sector.” *International Journal of Computer Trends and Technology* 72 (5): 124–140.

Ji, J., T. Qiu, B. Chen, et al. 2023. “AI Alignment: A Comprehensive Survey.” arXiv. Accessed June 1, 2025. https://arxiv.org/abs/2310.19852.

Madaan, A., N. Tandon, P. Gupta, et al. 2023. “Self–Refine: Iterative Refinement with Self-Feedback.” arXiv. Accessed June 1, 2025. https://arxiv.org/abs/2303.17651.

Martin, M. W., and R. Schinzinger. 2010. *Introduction to Engineering Ethics.* New York: McGraw-Hill.

Padmanaban, H. 2024. “Revolutionizing Regulatory Reporting through AI/ML: Approaches for Enhanced Compliance and Efficiency.” *Journal of Artificial Intelligence General Science* 2 (1): 57–76.

Schlossberger, E. 2023. *Ethical Engineering: A Practical Guide with Case Studies.* Boca Raton: CRC Press.

Ulnicane, I., T. Mahfoud, A. Salles, et al. 2023. “Experimentation, Learning, and Dialogue: An RRI-Inspired Approach to Dual-Use of Concern.” *Journal of Responsible Innovation* 10 (1).

Umbrello, S., and I. van de Poel. 2021. “Mapping Value Sensitive Design onto AI for Social Good Principles.” *AI and Ethics* 1:283–296.

Urquhart, L. D., and P. J. Craigon. 2021. “The Moral-IT Deck: A Tool for Ethics by Design.” *Journal of Responsible Innovation* 8 (1): 94–126.

Van de Poel, I. 2020. “Design for Value Change.” *Ethics and Information Technology* 23 (1): 27–31.

Verbeek, P.-P. 2006. “Materialising Morality: Designing Ethics and Technological Mediation.” *Science, Technology and Human Values* 31:361–380.

———. 2011. *Moralizing Technology: Understanding and Designing the Morality of Things.* Chicago: University of Chicago Press.

Wallach, W., and C. Allen. 2009. *Moral Machines: Teaching Robots Right from Wrong.* Oxford: Oxford University Press.

Wang, D. 2018. “Toward an Experimental Philosophy of Engineering.” In *Philosophy of Engineering : East and West*, ed. by C. Mitcham, B. Li, B. Newberry, and B. Zhang, 37–50. Berlin and Heidelberg: Springer.

———. 2020. “Towards Responsible Engineering: Interpretation and Implementation of Ethical Codes.” *Chemical Engineering Higher Education* 37 (3): 1–7.

———. 2023. “Towards Ethical Engineering: Artificial Intelligence as an Ethical Governance Tool for Emerging Technologies.” *Computer Sciences and Mathematics Forum* 8.

Wang D. [Ван Д.] From Engineering Ethics to Ethical Engineering [От инженерной этики к этической инженерии] : Leveraging AI for Governing Emerging Technologies [использование ИИ для управления перспективными технологиями] // Философия. Журнал Высшей школы экономики. — 2025. — Т. 9, № 4. — С. 47–67.

Дачжоу Ван
д. филос. н., профессор, Школа гуманитарных наук Университета Китайской академии наук (Пекин); orcid: 0000–0001–9586–4597

От инженерной этики к этической инженерии

использование ИИ для управления перспективными технологиями

Получено: 16.09.2025. **Рецензировано:** 11.11.2025. **Принято:** 11.11.2025.

Аннотация: Этическая инженерия (ЭтИн) — это формирующаяся дисциплина, представляющая собой смену парадигмы по сравнению с традиционной инженерной этикой (ИнЭт). В отличие от подхода, ориентированного в первую очередь на обучение отдельных специалистов, ЭтИн ставит целью системное внедрение этических принципов в саму структуру технологических систем и процессов управления. В статье анализируется переход от ИнЭт, ориентированной на обучение специалистов нормативным принципам, к ЭтИн, рассматривающей этику как инженерную задачу системного уровня, предполагающую трансляцию этических принципов в исполняемые инструменты управления. В исследовании подчеркивается двойственная роль ИИ как и основной области, требующей регулирования, и ключевого средства для его реализации, а также анализируется его потенциал для совершенствования этического управления через повышение проверяемости алгоритмов, поддержку сложного этического принятия решений и междисциплинарное коллаборативное управление. Одновременно рассматриваются такие вызовы, как обеспечение согласования ценностей, снижение рисков предвзятости и опасность технологического редукционизма. Работа выделяет восемь ключевых проблемных областей, формирующих ядро исследовательской повестки ЭтИн, и утверждает, что развитие этой дисциплины следует понимать как экспериментальный, итеративный процесс. Сформулированный парадигмальный сдвиг не только расширяет практические возможности реализации инженерной этики, но и предлагает новые методологические решения для этического управления развивающимися технологиями в эпоху ИИ.

Ключевые слова: инженерная этика, этическая инженерия, искусственный интеллект (ИИ), перспективные технологии, ценностно-ориентированное проектирование, морализирующие технологии.

DOI: 10.17323/2587-8719-2025-4-47-67.