Embodied AI Ethics: A Phenomenological Critique of Current AI Explanation Systems

Wenrui Liang

School of Computer Science, The University of Sydney, Sydney, NSW 2050, Australia

Abstract

Current explainable artificial intelligence (XAI) systems face a fundamental challenge, which is the conflict between algorithmic explanations and embodied moral understanding. We examine two typical cases to reveal that human moral understanding possesses irreducible embodied characteristics. These characteristics cannot be fully captured by existing AI explanation technologies. Base on Merleau-Ponty's embodied phenomenology[1], we propose "Embodied AI Ethics" as a new theoretical framework. This framework shifts the focus from making AI more moral to protecting and enhancing human moral capabilities. We demonstrate how current XAI technologies systematically exclude these features. They do this through processes of disembodiment, decontextualization, and desubjectification. Based on these findings, we propose four design principles for embodied AI ethics. These principles provide theoretical guidance for developing AI interfaces that respect and support human moral understanding.

Keywords

Embodied AI Ethics, Phenomenology, Explainable AI, Human-Computer Interaction, Moral Understanding.

1. Introduction

In 2020, a 58-year-old African American woman named Deborah Williams felt something was wrong with her body. She went to the hospital for examination. The hospital's AI cancer screening system showed her diagnostic result as "87% benign probability." The doctor accordingly determined that her tumor was benign. However, this woman was diagnosed with advanced lung cancer six months later. She sadly passed away shortly after.

Another algorithmic tragedy befell 18-year-old African American youth Johnson. Johnson had a criminal history. He was judged by the COMPAS judicial risk assessment algorithm to have a 73% risk of reoffending. However, the judge at that time recalled observing sincere remorse in Johnson's eyes. In fact, Johnson not only did not reoffend but became a community volunteer.

When human intuition and algorithmic explanations conflict, how should we choose? This is one of the great challenges facing today's AI explanation systems. As humans, doctors and judges naturally possess a unique cognitive mode. This is moral understanding that emerges through direct bodily contact with the world. However, today's mainstream explainable AI does not possess this embodied moral foundation. Instead, it aims to emphasize technical transparency and algorithmic interpretability. The tragedies of Williams and Johnson tore open a small gap. Could AI be fraught with hidden dangers in the future due to the absence of moral ontology?

Accordingly, this paper proposes embodied AI ethics as a new theoretical paradigm. This paradigm aims to shift the focus from making AI more moral to protecting and enhancing human moral capabilities. The core insight is important. Moral understanding is essentially embodied, intersubjective, and contextually sensitive. These features make it impossible to be fully algorithmized. The entire paper takes Merleau-Ponty's phenomenology[1] as the moral

ontological foundation. It will revolve around the following key questions. What are the fundamental features of human embodied moral understanding? Why cannot these features be adequately supported by current AI explanation systems? How can we design human-computer interaction interfaces that protect and enhance such understanding? By answering these questions, we aim to reveal the deep limitations of current XAI technologies. We also aim to provide theoretical principles for developing AI interfaces that align with the essence of human understanding.

2. Related Work

2.1. Limitations of Current AI Ethics Research

Today's AI ethics research mainly has three approaches. The first is algorithmic fairness research, focusing on identifying and mitigating bias [2] & [3], 2016). The second is technology governance research, involving regulatory frameworks [4] et al., 2017). The third is philosophical ethics research, examining moral significance [5] et al., 2018). However, these perspectives' ethical examination of AI systems still belongs to external perspectives. This makes it difficult to explain how AI systems affect human understanding institutions themselves. The focus of recent research on the intersection of explainability and ethics design still lies in a specific area. This area is how to improve technical systems through deepening understanding of ethics. It is not about the impact and transformation of systems on human moral capabilities.

2.2. Explainable AI Research in HCI

In the HCI field, explainable AI research mainly focuses on user experience and trust building. SHAP (SHapley Additive exPlanations) is a game theory-based machine learning explanation method that explains model decisions by calculating the marginal contribution value of each feature to the model prediction results, assuming that complex decisions can be decomposed into a linear sum of various feature contributions. Although technical solutions like SHAP attempt to improve ethical rationality by enhancing technical transparency, user research results show that the expert acceptance of such technologies varies significantly. For example, medical AI research found that while 57% of studies showed SHAP could improve doctors' trust in AI systems, 29% found limited effects, with practitioners frequently expressing that SHAP outputs did not align with clinical experience [6] et al., 2020; [7] et al., 2019).

The embodied interaction theory developed based on the phenomenological tradition emphasizes the fundamental role of the body in human-computer interaction. However, current embodied interaction research mainly focuses on physical interaction levels, with less involvement in the embodied dimensions of moral understanding.

2.3. Research Gaps

Three important gaps emerge: First, existing HCI research lacks philosophical reflection on the nature of human understanding. Second, explainable AI research adopts functionalist perspectives, lacking ontological analysis of explanation phenomena. Finally, systematic application of embodied interaction theory to ethical design of AI explanation systems remains scarce.

3. Case Analysis: The Conflict Between Embodied and Algorithmic Understanding

3.1. Williams' Medical Case: Embodied Intuition vs. Algorithmic Reasoning

Williams' persistent chest discomfort lacked textbook symptoms. Her doctor attributed her complaints to anxiety based on the AI system's probability. This led to the unfortunate outcome.

This case exemplifies what Merleau-Ponty describes as "motor intentionality." This is the body's capacity to register meaningful changes before cognitive awareness. Williams' pre-reflective bodily awareness represents lived, experiential knowledge. This knowledge precedes analytical thinking. The fundamental epistemological limitation lies in a specific aspect. Virtual AI systems do not possess a priori knowledge. Thus, they cannot obtain the embodiedness that humans have as embodied beings. They can only process externally observable data through simulation.

3.2. Morrison's Judicial Case: Moral Perception vs. Statistical Risk

Judge Morrison considered Johnson innocent through eye contact. Meanwhile, the AI's calculated result was wrong. Morrison's moral perception comes from a kind of "intercorporeal resonance." This is the capacity to understand others' moral intentions through direct lived encounters. This intersubjective moral perception operates through pre-reflective attunement between embodied subjects. This makes authentic moral understanding possible. The judge's assessment was based on her capacity to understand the defendant's moral intentionality through lived encounters. This represents a form of knowledge that cannot be reduced to statistical patterns.

3.3. Common Patterns: Marginalization of Embodied Knowledge

Both cases reveal systematic marginalization of embodied knowledge: (1) Epistemological hierarchies: embodied knowledge is dismissed as "subjective" while algorithmic analysis is elevated to "objective truth"; (2) Institutional constraints: protocols are increasingly structured around algorithmic recommendations; (3) Temporal displacement: the demand for rapid decisions favors algorithmic speed over embodied understanding's temporal requirements.

4. Theoretical Framework: Merleau-Ponty's Embodied Phenomenology

4.1. Core Phenomenological Concepts

Merleau-Ponty's argues that embodiment is consciousness's basic structure and our primary mode of world-engagement[1]. The body-subject possesses "motor intentionality"—the capacity to respond meaningfully before conceptual analysis occurs. Pre-reflective understanding operates before explicit thinking, representing a primary way of engaging complex situations. Intercorporeality reveals that understanding emerges through bodily encounters between subjects rather than isolated mental processes.

4.2. Three Fundamental Features of Embodied Moral Understanding

4.2.1. Pre-reflective Moral Perception

Williams' sensation exemplifies pre-reflective moral perception—the body-subject's capacity to detect morally significant features before conceptual analysis. This perception is immediate, holistic, temporally extended, and non-linguistic. AI systems cannot access this dimension because they lack embodied world-engagement.

4.2.2. Intersubjective Moral Relations

Morrison's perception of Johnson's remorse illustrates intersubjective moral understanding—perceiving others' moral intentions through direct embodied encounters. This involves intercorporeal resonance and moral intentionality recognition in unique encounters between specific individuals.

4.2.3. Contextual Moral Sensitivity

Both cases demonstrate contextual sensitivity—responding to unique moral demands of particular situations. This involves situational particularity, temporal sensitivity, and value

integration. AI systems operate through standardization and cannot perceive moral uniqueness requiring special attention.

5. User Experience Analysis: Systematic HCI Evaluation of XAI Limitations

Our phenomenological analysis reveals fundamental user experience problems in current XAI systems. To systematically examine these issues from an HCI perspective, we developed a comprehensive framework for analyzing four key dimensions of human-computer interaction in AI explanation systems:

5.1. HCI Framework for Embodied AI Ethics

HCI Dimension	User Actual Experience	Current XAI Design	Experience Conflict	Williams Case	Morrison Case	Design Improvement Requirements	Measurable Indicators
Cognitive Load Dimension							
Holistic Understanding	Gestalt pattern recognition, intuitive grasp of the whole	Feature list analysis, requiring item-by-item understanding and integration	Holistic cognition vs. decomposed cognitive load	Doctors rely on intuition to perceive the overall condition, but SHAP requires analyzing each feature weight	Judges holistically judge defendant's character, but need to understand decomposed risk factors	Provide options to switch between overview and details	Understanding time, cognitive load scale, error rate
Information Processing Rhythm	Gradual understanding at personal pace, allowing pauses for thinking	One-time batch information output, high processing pressure	Natural rhythm vs. information flood pressure	Instant display of all feature importance, doctor information overload	Immediate display of complete risk assessment, judge has no time to digest	User-controlled phased information display	Processing time, pressure perception, decision quality
Professional Cognitive Matching	Conforming to domain professional thinking patterns and habits	Technology logic-driven, mismatched with professional thinking	Professional cognition vs. technical logic conflict	SHAP output doesn't align with clinical reasoning approach	Risk scores disconnect from judicial thinking logic	Domain knowledge-oriented information architecture	Professional matching score, learning curve
Trust Building Dimension							
Trust Foundation	Based on long-term experience accumulation and professional intuition	Based on algorithmic transparency and technical authority	Experience trust vs. technical trust conflict	Patients trust their bodily sensations, don't trust 87% probability	Judges trust professional judgment, question 73% risk score	Balance mechanism between experience and algorithmic trust	Trust scale, algorithm acceptance, usage willingness
Trust Verification	Established through result feedback and continuous verification	Established through explanation reasonableness and technical indicators	Result verification vs. explanation verification difference	Williams' final cancer diagnosis confirmed intuition was correct	Johnson's successful rehabilitation verified judge's accurate judgment	Integrate result feedback trust update mechanism	Prediction accuracy, long-term trust change trajectory
Trust Repair	Acknowledging errors, learning and adjusting from failures	Technical upgrades and algorithmic optimization	Humanized vs. technicalized repair methods	Doctors can admit misdiagnosis and improve, AI lacks this capability	Judges can reflect on wrong judgments, algorithms can only be passively optimized	Support interface design for error acknowledgment and learning	Error recovery time, trust rebuilding degree
Decision Support Dimension							
Decision Flexibility	Flexibly adjust decisions based on contextual uniqueness	Standardized procedures, prioritizing consistency	Contextual flexibility vs. standard consistency tension	Each patient situation is unique, requiring personalized treatment	Each case context is special, requiring individualized judgment	Context-aware adaptive decision support	Decision adaptability score, context matching degree
Decision Autonomy	Maintaining human dominance in final decisions	Tending to rely on algorithmic recommendations and automation	Human dominance vs. algorithmic dependence tension	Doctor decision-making power marginalized by AI recommendations	Judge judgment overly influenced by algorithmic assessment	Interface design strengthening human final decision authority	Autonomy scale, decision control sense, responsibility undertaking degree

HCI Dimension	User Actual Experience	Current XAI Design	Experience Conflict	Williams Case	Morrison Case	Design Improvement Requirements	Measurable Indicators
Decision Quality	Comprehensively considering multiple values and long-term impacts	Optimizing specific indicators and short-term effects	Value integration vs. indicator optimization conflict	Consider patient overall wellbeing vs. optimize diagnostic accuracy	Balance justice, mercy, social impact vs. reduce recidivism risk	Multi-value balanced decision support tools	Decision satisfaction, multi-dimensional result evaluation
Interaction Experience Dimension							
Interaction Proactivity	Active exploration, questioning, verification and dialogue	Passive reception of information and execution of recommendations	Active exploration vs. passive reception experience difference	Doctors cannot actively explore AI's reasoning process	Judges cannot deeply question algorithmic evaluation logic	Support hypothesis testing and exploratory interaction design	Interaction proactivity score, exploration depth, questioning frequency
Feedback Mechanism	Two-way dialogue, timely adjustment and improvement	One-way output, lacking feedback loops	Dialogue vs. one-way interaction experience	AI cannot respond to doctors' further inquiries	Algorithms cannot answer judges' in-depth questions	Integrate dialogue and feedback interaction mechanisms	Dialogue rounds, feedback responsiveness, interaction satisfaction
Personalized Adaptation	Adapting to personal style, habits and preferences	Standardized interface, lacking personalization	Personal adaptation vs. standardized interface experience	Cannot adapt to different doctors' diagnostic styles and habits	Cannot adapt to different judges' judgment styles and experience	Learnable and adaptive personalized interfaces	Personalization degree, adaptation effect, user satisfaction

5.2. Key Findings from HCI Analysis

This systematic analysis reveals that current XAI systems create fundamental mismatches between user needs and system capabilities across all four dimensions:

Cognitive load problems: Users face information overload and cognitive-professional mismatch, leading to increased decision time and reduced accuracy.

Trust issues: Conflicts between experiential and algorithmic trust sources cause long-term trust erosion and reduced system adoption rates.

Decision support failures: Lack of contextual flexibility and human autonomy protection undermines decision quality and user satisfaction.

Interaction limitations: Passive, standardized interfaces prevent effective exploration and personalization, reducing user engagement and effectiveness.

6. Design Implications: Four Principles of Embodied AI Ethics

6.1. Core Design Principles

6.1.1. Embodied Understanding Preservation Principle

AI systems should protect the space for embodied moral understanding rather than replace it. Systems must clearly distinguish between "technical analysis" and "human judgment" domains, mandate embodied human engagement at critical points, and actively discourage over-reliance on algorithmic recommendations.

6.1.2. Intersubjectivity Facilitation Principle

Since AI cannot participate in authentic intersubjective relations, systems should facilitate moral dialogue between humans. This includes identifying situations requiring intersubjective engagement, supporting multi-party participation, and protecting time for direct human encounters.

6.1.3. Contextual Sensitivity Maintenance Principle

AI systems should provide contextually adaptive frameworks rather than uniform approaches. This requires offering situationally adapted explanations, recognizing different levels of moral

complexity, and maintaining the uniqueness of each case rather than treating situations as instances of general categories.

6.1.4. Moral Capacity Enhancement Principle

Technology should serve human moral development rather than replace moral thinking. Systems should provide moral reflection tools, identify learning opportunities, and encourage moral growth over time rather than providing predetermined conclusions.

6.2. HCI Design Guidance

Medical AI interfaces: Implement hierarchical organization prioritizing patient information over technical analysis, mandate direct patient interaction periods before AI analysis access, and include tools supporting multi-perspective decision-making.

Judicial AI interfaces: Present defendant information maintaining moral agency recognition, include structured moral reflection opportunities engaging judges' embodied judgment, and ensure algorithmic analysis enhances rather than replaces direct encounters.

6.3. Evaluation Framework

Evaluation must transcend technical metrics to include moral sensitivity development, contextual judgment capacity, and intersubjective dialogue ability. Success indicators should assess relationship quality and moral encounter authenticity, requiring longitudinal studies tracking moral capacity changes over extended periods.

7. Discussion

7.1. Theoretical Contributions

This research introduces "Embodied AI Ethics" as a new framework that prioritizes the protection and enhancement of human moral agency rather than making AI systems more ethical. The systematic application of Merleau-Ponty's phenomenology to AI explanation analysis demonstrates the value of phenomenological methods for understanding human-computer interaction dimensions invisible to traditional approaches[1].

7.2. Limitations and Future Directions

Theoretical scope: While focusing on Merleau-Ponty's phenomenology, other philosophical traditions may provide additional insights[1]. Cultural boundaries: Analysis primarily draws on Western backgrounds—cross-cultural applicability requires investigation. Empirical validation: Design principles need systematic empirical validation through user studies and longitudinal research. Implementation challenges: Translating theoretical principles into specific technical implementations presents significant challenges requiring interdisciplinary collaboration.

8. Conclusion

This research is based on Merleau-Ponty's theory[1]. We establish "Embodied AI Ethics" as a new theoretical framework for AI explainable systems. Through phenomenological reduction, we first demonstrate that algorithms do not possess embodied characteristics. Therefore, they cannot reproduce human moral understanding. Consequently, today's AI systems have fundamental incompatibilities. Their assumptions about understanding and explanation are incompatible with embodied human moral judgment. However, through our new theoretical framework analysis, we can derive four design principles. These principles point toward human-AI collaboration that leverages technological capabilities while protecting human moral capabilities.

References

[1] M. (2012). Phenomenology of perception (D. A. Landes, Trans.). Routledge. (Original work published 1945)

[2] S., & [10], A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671-732.

[3] A. D., & [2], S. (2018). The intuitive appeal of explainable machines. Fordham Law Review, 87, 1085-1139.

[4] S., Mittelstadt, B., & [4], L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International Data Privacy Law, 7(2), 76-99.

[5] L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689-707.

[6] S., Wang, B., Zhu, M., & Zhang, J. (2020). Effectiveness of explainable AI in medical diagnosis: A systematic review. Journal of Medical Internet Research, 22(8), e19340.

[7] S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. Machine Learning for Healthcare Conference, 359-380.