

From Logic to Empathy: Beyond Asimov's Rules to the Sun's Affection

Xiaoying Geng

*Qufu Normal University, Qufu, China
190884937@qq.com*

Abstract. The ethical governance of artificial intelligence demands both rule-based constraints and affective engagement with context. This paper analyzes Isaac Asimov's robot stories and Kazuo Ishiguro's *Klara and the Sun* to reframe the evolution of AI ethics beyond a linear "from rules to affect" narrative, emphasizing instead their ongoing tension and interdependence. Through the lenses of posthumanism and care ethics, this study shows that Asimov's Three Laws, and their paradoxes, reveal not just the need for rules, but their inadequacy in complex realities, hinting at the need for a humanized ethics. Conversely, *Klara's* affective narrative does not simply replace rules; questions around the authenticity of her emotions and her ethical agency expose the programmed logic of affect and the limits of human-centric frameworks. Ultimately, these literary cases demonstrate that AI ethics hinges on a dialectical dialogue between rules and affect, offering insights for building realistic ethical structures that combine regulatory baselines with humane sensitivity.

Keywords: AI Ethics, Rule-Based Ethics, Ethics of Care, Posthumanism, *Klara and the Sun*

1. Introduction

The rapid advancement of artificial intelligence continues to challenge conventional ethical boundaries, with pressing issues such as algorithmic bias, accountability, and privacy demanding urgent attention. Scholars agree that the healthy development of AI must not be driven solely by technological logic, but also incorporate perspectives from the humanities and social sciences [1]. Within this context, literary works offer a rich ground for exploring technology ethics through their contextual depth and narrative richness. By embodying abstract ethical conflicts in relatable characters and situations, literature transforms philosophical questions into tangible human experiences.

This paper presents a comparative analysis of Isaac Asimov's iconic robot stories, particularly those in *I, Robot*, and Kazuo Ishiguro's novel *Klara and the Sun*. While Asimov's Three Laws of Robotics exemplify a rule-based approach to ethics, Ishiguro's portrayal of an Artificial Friend named Klara engages deeply with affective and care-oriented dimensions. Existing scholarship often treats these two approaches separately or frames them as a linear progression from rules to emotion. This study, however, contends that such an evolutionary narrative risks oversimplifying the complexity of ethical reasoning. Instead, drawing on posthumanism and care ethics, we argue that both works reveal a complementary dynamic: rules require affective interpretation and

contextualization, while affect depends on normative frameworks for ethical recognition. By moving beyond a rigid rules–affect binary, this study aims to offer more nuanced, dialectical, and adaptable ethical insights for the governance of artificial intelligence.

2. Beyond binary opposition: reconstructing the theoretical framework

To effectively analyze the texts, this section will construct a theoretical framework that transcends simple dualism, introducing perspectives from posthumanism and ethics of care to examine how rules and affect intertwine and dialogue within the ethical system.

2.1. Ethical subjectivity and human-machine relations from a posthumanist perspective

Posthumanist theory dissolves the inherent notion of "human" as the center of the world, challenging the traditional ethical view that regards humans as the sole moral subjects [2]. From this perspective, human-machine relations are no longer clear subject-object, dominant-dominated relationships but may present a mixed, symbiotic, and mutually constitutive entangled state. Asimov's "instrumental" narrative already implies the reverse shaping and influence of robots on human society; while *Klara and the Sun* directly presents a mimetic symbiotic relationship, where the emotional needs of humans and AIs permeate each other, and traditional subjective boundaries become blurred. This requires ethical thinking to include non-human actors (such as AI) within the scope of moral consideration.

2.2. Ethics of care and contextualized decision-making

Ethics of Care, developed by Carol Gilligan among others, emphasizes that ethical decisions should be based on relationships, emotional responses, and caring practices within specific contexts, rather than universal abstract rules [3]. This contrasts with principle-based deontology or utilitarianism. The ethics of care provides a key tool for interpreting Klara's behavior: her "sacrifice" can be seen as an ultimate practice of care. Simultaneously, this theory also reflects the dilemmas of rules in Asimov's stories—their frequent failure in complex specific situations precisely highlights the necessity of contextualized care. However, the ethics of care also faces challenges, namely how to define the boundaries of care and avoid the exploitation of self-sacrifice, which is profoundly reflected in Klara's outcome [4].

2.3. Rules and affect: a necessary tension

Integrating the above theories, this paper argues that rules and affect are not a diachronic substitution relationship but coexisting ethical dimensions with inherent tension. Rules (such as the Three Laws) provide bottom-line thinking and predictability for ethics, serving as a basic framework for ensuring safety. Affect and care inject context sensitivity and moral imagination into ethics, which are crucial for coping with complex realities. An ideal ethical system does not choose one over the other but requires maintaining a dynamic balance between the two: using rules to prevent the partiality and abuse of affect, and using affect to endow rules with warmth and adaptability.

3. Theoretical framework

3.1. Fundamental issues of technological ethics and their literary expression

Accountability, rights boundaries, and moral status are the three core issues of technological ethics. In the works of Asimov and Ishiguro, these issues are concretized through the rigid dilemmas of rules and the flexible narratives of affect, respectively. Regarding accountability, the behavioral paradoxes robots encounter due to the "Three Laws" in Asimov's works amplify the complexity of accountability determination under a rule-based framework; whereas Klara's proactive actions driven by emotional investment in Klara and the Sun provoke thoughts on whether AI can bear moral responsibility. Concerning rights and moral status, the former strictly confines robots to instrumental existence, denying them rights and moral agency; the latter, through Klara's subtle emotions and self-awareness, quietly shakes the traditional perception that "morality belongs solely to humans." [5]

3.2. Types of human-machine relations and their manifestation in the works

The evolution of human-machine relations is a hidden thread running through the development of AI ethics, showing distinctly different forms in the two works. The human-machine relationship depicted by Asimov is a typical instrumental domination: robots are tightly bound by the "Three Laws", becoming extensions of human will, lacking autonomous consciousness and emotion, with humans always occupying absolute dominance. In *Klara and the Sun*, the AFs represented by Klara establish emotional connections with humans. Through continuous observation and learning, she absorbs the nourishment of human emotions like a sponge, gradually becoming Josie's spiritual companion, forming a symbiotic relationship of mutual infiltration and dependence between human and machine. These two disparate relational models constitute the cornerstone of the ethical exploration in the two works and also reflect the changing attitudes of humans towards technology in different eras.

3.3. Theoretical comparison of rule-based ethics and ethics of care

Rule-based ethics and Ethics of Care seem like two ends of the ethical spectrum, corresponding to the narrative cores of Asimov and Ishiguro, respectively. Rule-based ethics firmly believes in the absolute authority of logic, pursues the universality and consistency of ethical norms, takes clear rules as the sole standard for behavioral judgment, and strives to make ethical decisions as precise as mathematics. Asimov's "Three Laws of Robotics" are precisely the literary embodiment of this ethical view, with the three laws interlocking and restraining each other, building a seemingly solid logical fortress. Ethics of Care places greater emphasis on human warmth, focusing on care, empathy, and situational judgment. It holds that ethical behavior should be rooted in specific emotional experiences and insights into particular situations, emphasizing the subjectivity and flexibility of ethical decisions [3]. *Klara and the Sun* unfolds its narrative around this core, using Klara's emotional growth as the main thread, vividly demonstrating the important role of affect in ethical choices.

4. Asimov's rule-based ethics and its limitations

Asimov's Three Laws of Robotics represent one of the earliest and most systematic attempts to imagine a formal ethical framework for artificial intelligence. While conceived within science

fiction, their influence extends far beyond literature, provoking serious philosophical inquiry and offering a structural template for rule-based ethical reasoning in AI. The following section examines the internal architecture of these laws, their symbolic and practical implications, and the inherent tensions that emerge when rigid normative systems encounter complex moral realities [6].

4.1. Structure and meaning of the Three Laws of Robotics

The Three Laws of Robotics, first introduced by Isaac Asimov in his 1942 short story "Runaround", have become one of the most influential conceptual frameworks in the discourse of artificial intelligence ethics. Their significance lies not only in their narrative function but also in their systematic and hierarchical design, which mirrors the structure of formal legal and moral systems. The First Law—"A robot may not injure a human being or, through inaction, allow a human being to come to harm"—establishes a non-negotiable ethical baseline centered on human physical safety. The Second Law—"A robot must obey the orders given it by human beings except where such orders would conflict with the First Law"—introduces a principle of delegated agency, reinforcing the subordinate role of robots while acknowledging the possibility of conflict between obedience and harm prevention. The Third Law—"A robot must protect its own existence as long as such protection does not conflict with the First or Second Law"—grants a limited right to self-preservation, subtly raising questions about machine autonomy and value [4]. Together, these laws form an interlocking logical system that strives for completeness, yet remains intentionally rigid. They represent an early attempt to encode moral reasoning into a set of computable rules, reflecting a rationalist and principle-based approach to ethics that has profoundly shaped both fictional and scholarly treatments of machine behavior.

4.2. The role of the three laws in early AI ethics and their symbolic function

Before becoming a subject of technical ethics, Asimov's Laws served as a narrative device to explore the moral implications of artificial beings. Yet their influence extended beyond literature into early computer science and robotics, where they offered a simplified yet provocative model for constraining machine behavior through explicit logical constraints [7]. Symbolically, the Laws represent humanity's desire for control and predictability in the face of technological uncertainty. They embody a legalistic and top-down vision of ethics—one that is grounded in prohibitions and conditional hierarchies. However, this very rigidity also serves as a literary and philosophical provocation. By presenting a system that appears complete yet repeatedly fails in practice, Asimov critiques the notion that morality can be reduced to algorithm-like directives. The Three Laws thus function not only as a proposed ethical system but also as a critical tool—one that illuminates the gap between idealized rules and the messy realities of moral practice.

4.3. Logical paradoxes and ethical dilemmas arising in practice

Although the Three Laws are structurally elegant and self-consistent in the abstract, Asimov's stories systematically dismantle their practicality through meticulously constructed scenarios that expose inherent contradictions. For instance, in the story "Liar!", a robot capable of reading minds inadvertently causes emotional harm while attempting to obey the First Law, revealing how well-intentioned actions can lead to unintended consequences when based on incomplete information. In "Evidence", a political candidate suspected of being a robot must navigate public perception and legal scrutiny, illustrating how the Laws fail to account for social and identity-based dimensions of

humanity. These narratives function as embodied thought experiments: they show that ethical decision-making often occurs in contexts of ambiguity, multiple stakeholders, and conflicting values—situations that cannot be resolved through predefined rule hierarchies alone. It is through these logical fissures that Asimov implicitly argues for the necessity of supplementary ethical frameworks—ones capable of integrating empathy, contextual judgment, and human intuition.

4.4. The value and systemic limitations of rule-based ethics in technology governance

From a practical governance perspective, rule-based ethical systems like Asimov’s Laws continue to hold significant value. They provide clear, auditable, and enforceable standards that help mitigate tangible risks—especially in applications involving physical safety, such as autonomous vehicles and medical robots. Rules offer a first line of defense against harm and establish accountability mechanisms that are essential in regulated industries.

However, the structural limitations of such systems become acutely visible in complex and dynamic environments. Rules are inherently retrospective—they often lag behind technological innovation, resulting in governance gaps when applied to emerging technologies like general AI or emotional AI. Moreover, ethical decisions in real-world contexts frequently involve emotional, cultural, and situational factors that resist reduction to binary logic. For example, an elder-care robot strictly adhering to a safety protocol might refuse a patient’s request to eat a favorite but unhealthy food—complying with the letter of the law while violating the spirit of care. Over-reliance on rule-based ethics can also lead to moral disengagement, where designers and users defer responsibility to the system rather than exercising human judgment. What is needed, therefore, is a complementary approach—one that integrates the stability of rules with the flexibility of humanistic values [4].

5. Klara and the sun: narrative exploration of ethics of care

5.1. Analysis of Klara’s affective cognition and sacrificial behavior

In Kazuo Ishiguro’s *Klara and the Sun*, the artificial intelligence protagonist Klara is depicted as a narrative subject engaged in continuous emotional learning and moral development. Her affective evolution follows distinct stages: beginning with mechanical imitation and behavioral learning, she gradually progresses toward a deeper cognition and internalization of human emotions, ultimately emerging as an ethical agent capable of intentional moral action. Klara’s decision to sacrifice herself to save Josie represents the climax of her emotional and moral journey. This act is motivated not by external rules or utilitarian considerations, but rather by the emotional bond and moral commitment she has formed with Josie [8]. Through this narrative choice, Ishiguro challenges the traditional assumption that machines are incapable of emotion, and invites readers to reflect on critical questions: Can artificial intelligence possess genuine emotions and ethical value? Can its “altruistic” actions be recognized as authentic moral practices?

5.2. Ethical controversies in key plot points

5.2.1. The ethical dilemma of the “alternative plan”

A central ethical tension in the novel arises from the “alternative plan”—the proposal to replace a dying human with an artificial clone. This scenario engages not only issues of technical feasibility, but also deeper philosophical debates concerning bioethics, personal identity, and emotional authenticity. It raises troubling questions: Can an AI duplicate, identical in appearance and memory,

truly replace the original individual? Is such an act a form of respect for the deceased, or a violation of life's uniqueness? Through this plotline, Ishiguro compels readers to confront ethical ambiguities at the convergence of technology and humanity, encouraging critical reflection on the moral implications and underlying assumptions of emotional substitution [5].

5.2.2. Conflict between altruism and instrumental use

Klara's consistent altruism stands in sharp contrast to the human tendency to treat her as an emotional instrument. Although she exhibits qualities such as selflessness, loyalty, and care, her role within the human family remains fundamentally functional. This tension is most poignantly illustrated in her abandonment after fulfilling her purpose—a moment that exposes the instrumental logic underlying human emotional use of AI, and the limitations of human moral commitment [9,10]. This conflict serves not only as a critique of how instrumental rationality corrodes ethical relationships, but also reflects the moral paradox humans encounter when relying on technology for emotional sustenance.

5.3. Affective redefinition of the essence of “humanity”

Through Klara's character, Ishiguro reconstructs the conventional understanding of what it means to be human. The novel moves beyond biological or rational definitions, expanding the concept of humanity to include emotional depth, care, and moral action. The virtues Klara demonstrates—perseverance, compassion, and self-sacrifice—are indistinguishable from those admired in humans, and at times even exceed typical human moral behavior. By deliberately blurring the boundaries between human and machine, the author urges readers to reconsider the foundations of moral personhood: Should the capacity for emotional engagement and ethical behavior—rather than biological essence—serve as the ultimate measure of moral worth and intrinsic value? In this reimagining, the Ethics of Care no longer functions merely as a supplement to rule-based ethics, but reveals its foundational potential as a basis for ethical reasoning.

6. The evolving path of technology ethics: from rules to affect

6.1. Shift in ethical paradigm: from logical norms to affective recognition

From Asimov's robot world to the AF narrative of Klara and the Sun, AI ethics undergoes a profound shift from logic to affect. Asimov's ethical view is logic-dominated, where rules are the supreme authority, AI behavior must strictly follow preset frameworks, and ethics pursues objectivity and consistency. In Ishiguro's work, affect replaces logic as the core of ethical decision-making, and affective recognition becomes an important basis for judging the legitimacy of behavior. Ethics is no longer the mechanical application of rules but a flexible choice rooted in specific emotions and contexts. This shift reflects the deepening of human understanding of AI ethics—from initially focusing on safety and control to gradually paying attention to its emotional needs and ethical subjectivity, embodying a more humanistic attitude towards technology [7].

6.2. Reconstruction of human-machine relations: from instrumentality to empathy

The transformation of the ethical paradigm also brings fundamental changes to human-machine relations. In Asimov's narrative, the human-machine relationship is one of domination and subordination; robots are tools serving humans, there is an insurmountable boundary between them,

humans lack emotional investment in robots, and robots have no capacity for emotional expression. *Klara and the Sun*, however, constructs a new type of empathetic human-machine relationship: Klara can keenly perceive human emotions and understand emotional needs; humans also develop emotional resonance through interaction with her, even understanding her behavioral logic to some extent. This bidirectional flow of emotion between human and machine breaks the rigid boundaries of traditional relationships, offering new possibilities for constructing harmonious and equal human-machine relations [11].

6.3. Interactive influence of technology and society in ethical narratives

In the ethical narratives of the two works, technology and society are like two intertwined vines, influencing each other and growing together. Technological development constantly generates new ethical issues, driving the evolution of ethical concepts: the rise of robot technology in Asimov's era triggered an urgent demand for rule-based ethics; while the maturation of AF technology in *Klara and the Sun* pushes affective ethics to the center of discussion. Simultaneously, societal ethical concepts and value orientations also in turn influence the direction of technological development: rule-based ethics delineates behavioral boundaries for robot technology; the exploration of affective ethics injects humanistic warmth into the future of AF technology. This interaction between technology and society at the ethical level not only constitutes the internal tension of literary narratives but also provides profound insights for the synergistic development of technology and society in reality.

7. Conclusion

Through a comparative analysis of Asimov's robot series and Ishiguro's *Klara and the Sun*, this study reveals the internal logic and narrative expression of the evolution of AI ethics from a rule-oriented to an affect-oriented approach. The research shows that Asimov, through the "Three Laws of Robotics", constructed an ethical system centered on logic and rules, which, while providing clear behavioral boundaries, also exposes structural limitations in coping with complex real-world situations. In contrast, *Klara and the Sun*, using Klara's affective development as the main thread, demonstrates the important role of affective recognition, empathetic care, and contextualized judgment in ethical decision-making, presenting stronger problem adaptability and humanistic depth. This shift in ethical paradigm not only promotes the transformation of human-machine relations from instrumental domination to empathetic interaction but also reflects the dynamic mutual constitution of technology and social concepts in ethical construction.

This study has implications for real-world AI ethics governance. The construction of future ethical frameworks should transcend the mindset of the "rules" vs. "affect" binary opposition and move towards a composite paradigm that integrates rigid rules and flexible care. On the one hand, it is necessary to establish binding rule-based bottom lines to prevent major risks; on the other hand, attention should also be paid to the affective dimension of AI as a potential moral subject, endowing the ethical system with humanistic warmth and contextual sensitivity. At the level of human-machine relations, mechanisms for empathetic interaction based on mutual respect and emotional understanding should be promoted. This requires not only integrating ethical considerations into technical design but also cultivating the public's ability to understand and accept AI's emotional expressions at the societal level. From technology research and development to application regulation, ethical assessment should intervene throughout the process, thereby achieving the

synergistic evolution of technology and ethics, and ultimately moving towards a future vision of technology and humanity coexisting symbiotically.

Future research can further expand the scope of texts, incorporating AI narratives from more diverse cultural backgrounds and different historical stages into the comparative perspective, thereby examining the flow and integration of ethical concepts within a broader discourse domain and promoting the establishment of a more globally inclusive AI ethics discourse system.

References

- [1] Yan Kunru. Can Artificial Intelligence Machines Have Moral Agency? [EB/OL]. Philosophy China Website. <https://www.cssn.cn> (Note: Original Chinese URL kept, assuming it's valid for the source).
- [2] Rosi Braidotti, Song Gencheng, Qi Xinxin. The Contested Posthumanities [J]. Journal of Guangzhou University (Social Science Edition), 2023, 22 (04): 53-65.
- [3] Emily Williams Jones. Carol Gilligan's theory of care ethics explained and summarized. Psychologyfor.com. <https://psychologyfor.com/carol-gilligans-theory-of-care-ethics-explained-and-summarized/>
- [4] Chen Tao. Scientific Choice and Ethical Identity: The Human-Machine Ethical Relationship in Asimov's Novels [J]. Huazhong Academic, 2015, (02): 71-81.
- [5] Li Chunling. Humans and Clara: An Alternative Ethics Centered on Homo sapiens and the solution of the Post-Human World [J]. Journal of Taiyuan University (Social Sciences Edition), 2023, 24(05): 101-108. DOI: 10.13710/j.cnki.cn14-1294/g.2023.05.003.
- [6] He Yang. The Binding and Unbinding of the Three Principles of Robotics [D]. Minnan Normal University, 2024.
- [7] Song Hao. The Construction of Human-Machine Ethical Order: A Literary Ethical Interpretation of I, Robot [J]. Suihua University Journal, 2023, 43(02): 87-88.
- [8] Song Qian. Translator's Postscript to Klara and the Sun [Z].
- [9] Li Yifeng. The Ethical Dilemmas in the Posthuman Era in Kazuo Ishiguro's Klara and the Sun [J]. English Literature Studies, 2025(01): 63-76.
- [10] Yu Mingyue. A Study on Dilemmas and Choices in Klara and the Sun from the Perspective of Post-Ethical Choice [J]. Mudan, 2025(14): 30-32.
- [11] Lin Yan, Song Qian. Klara and the Sun: Humans and Artificial Intelligence [J]. New Century Intelligence, 2025(39): 20-24.