Robotic Narrative, Mindreading and Kazuo Ishiguro's *Klara and the Sun*

Guanghui Shang
https://orcid.org/0000-0002-8229-1571
Jiaxing University, China
dillonzhushang@163.com

Abstract

Bringing into dialogue the theory of mindreading reformulated within cognitive narratology, this article offers an analysis of Kazuo Ishiguro's *Klara and the Sun* (2021). It argues that Ishiguro extends this theory beyond human minds to nonhuman minds and human-machine bonds to explore human minds as human essence. By examining an artificial-intelligence (AI) character-narrator's struggle to read human minds through observation, this study draws two conclusions. Firstly, machines cannot comprehend entire human minds due to their complexity and variability. A mind encompasses not only an individual's own intricate thoughts and emotions but also others' diverse feelings about this individual. Secondly, both humans and machines engage in one-sided mindreading without eliciting reciprocal affective responses. This suggests that the limitations of robotic mindreading, coupled with human anthropocentrism, prevent the establishment of true human-machine intersubjectivity. By illustrating machines' incapability to possess human minds through robotic narrative, Ishiguro offers a new perspective on the theory of mindreading, asserting the irreplaceable nature of human minds in the age of AI to prompt a reflection on the uniqueness of human minds, a realm that machines cannot replicate or transfer.

Keywords: Kazuo Ishiguro; *Klara and the Sun*; robotic narrative; artificial intelligence (AI); mindreading

UNISA university of south africa PRESS
LASA
Journal of Literary Studies
Volume 40 | 2024 (In Progress) | #14849 | 17 pages
https://doi.org/10.25159/1753-5387/14849
ISSN 1753-5387 (Online)

Introduction

As noted by Yoseph Bar-Cohen and David Hanson, "[s]ince the Stone Age, people have utilised art and technology to replicate human appearance, capabilities, and intelligence," and "[r]ealistic human-like robots and simulations, which were once perceived as a fantastic and unattainable extension of these efforts, are now becoming a reality and walking into our lives, thanks to recent advances in related technology" (2009, 1). Paralleling Philip K. Dick's *Do Androids Dream of Electric Sheep?* (1968), a novel exploring human-robot interaction, Kazuo Ishiguro's *Klara and the Sun* (2022) dramatises this trend of human-robot co-existence seen in some 21st-century Anglo-American novels, such as Paolo Bacigalupi's *The Windup Girl* (2009), Ted Chiang's *The Lifecycle of Software Objects* (2010), Annalee Newitz's *Autonomous* (2017), and Ian McEwan's *Machines Like Me* (2019). Similar to these works, Ishiguro's work presents "the insidious encroachments" (McAlpin 2021) of machines on human life. What distinguishes this novel is its investigation into whether robots with artificial intelligence (AI) can read human minds, a crucial factor in determining their potential to supplant humans. While intelligence, distinct from consciousness commonly associated with human minds, refers to human capacity for learning and comprehension, AI concerns the simulation of human cognition. It aims to grant machines the ability to emulate human intelligence, enabling them to read human consciousness encompassing different emotions and thoughts. In the story, the AI character-narrator Klara exhibits human-like comprehension by processing gathered data to estimate potential outcomes. However, she struggles to interpret human behaviour and thoughts accurately. The novel establishes that true human-machine intersubjectivity cannot be achieved through one-sided mindreading.

Critical readings of the novel largely focus on machines' potential to possess human cognition and empathy. Told from the perspective of a robot, the novel fits into categories like "homodiegetic narration" by artificial beings (Nelles 2001, 189), or "stories of non-human narrators" (Bernaerts et al. 2014, 68). The narrative shows Klara's "fine-tuned observations about the emotional states of the humans" (Ajeesh and Rukmini 2023, 856). However, many mental states she reads prove inaccurate: Wrongly perceiving Josie as happy and carefree from her outward reactions, Klara misses Josie's true inner turmoil. She misinterprets Josie's mother's expectation and anxiety as exhaustion and hesitation. Instead of relying solely on mindreading or her experiences of desolation, she associates human behaviour with their sense of loneliness through the manager's revelation. I would argue that Klara seems to evince "empathy, compassion, happiness, and longing" (Mejia and Nikolaidis 2022, 304), but cannot truly feel these emotions. Her AI nature and lack of genuine intersubjectivity with humans hinder her close empathetic interactions with humans. Despite Klara's efforts to refine her mindreading ability by closely observing Josie and her mother, she perceives no reciprocal attempts from them to understand her. Even as Klara and humans begin caring about each other's minds, they find it hard to attain mutual understanding. Her choice to save Josie rather than replace her is not solely driven by sympathy. She decides

this upon realising that it is impossible for her and humans to "coinhabit and co-learn from each other" (Du 2022, 554). Klara admits the uniqueness of human minds, which she cannot replicate. As the essence of humans, their minds, which comprise not only their own emotions and thoughts but also the affection and concern their loved ones show for them, remain inaccessible to Klara. The intricacies of human minds resist her mechanical attempt to read them. Her failure to comprehend entire human minds underscores the human-machine distinction and the challenge of their mutual understanding in the age of AI.

In the novel, Klara reads human thoughts and emotions by observing their behaviour. Thus, it is pertinent to analyse her with the theory of mindreading. Mindreading is "a term used by cognitive psychologists to describe our ability to explain people's behaviour in terms of their thoughts, feelings, beliefs, and desire" (Zunshine 2003, 271). It presupposes "Theory of Mind," with which "[w]e are able to make attributions of states of mind to others" (Palmer 2011, 208). It further prompts us to analyse characters' behaviours and interpret their minds. This ability to attribute mental states to characters, as we do to each other in real-life interactions, enhances our understanding and enjoyment of fiction. Fictional mindreading explores how narrators read other characters' minds, how characters grasp each other's minds, and how readers interpret figural minds. Readers' cognition underpins these questions as they seek to understand figural minds, as well as how narrators and characters discern fictional minds. Lisa Zunshine elucidates the workings of mindreading with the concept of cognitive embedment or nesting, a "metaphor, which comes in handy when we want to talk about complex social dynamics that depend on people's awareness of their own and other people's states" (2022, 2). People can represent their mindreading process with second-level or third-level embedment. A two-level embedment is structured as "thinking about thinking" (Zunshine 2019, 2). For example, I *know* you *like* this. A three-level embedment "is structured as a mental state within a mental state within yet another mental state," or "thinking about thinking about thinking" (Zunshine 2022, 2–3). For instance, I *know* that you *know* that I *know*.

While second-level embedment is common in daily life, third-level embedment is often a literary device portraying complex mutual mindreading among characters. Third-level embedment reflects intricate figural relationships, referred to by George Butte as "deep intersubjectivity": It "begins when a self perceives the gestures, either of body or word, of another consciousness, and it continues when the self can perceive in those gestures an awareness of her or his own gestures" (2004, 27). Multiple embedment levels embody "intermental thought," defined by Alan Palmer as "joint, group, shared, or collective, as opposed to intramental, or individual or private thought," "as socially distributed, situated, or extended cognition, and also as intersubjectivity" (2011, 213). That is, characters understand each other's ideas and also agree with them. However, Kazuo Ishiguro's work offers a different interpretation: The third-level embedment, which Klara and humans use to understand each other, does not showcase their

agreement on how to continue Josie's life; rather, it reveals their lack of genuine intersubjectivity as illustrated by their misreading each other's minds.

Zunshine's mindreading theory mainly focuses on cognitive embedment's relation to literary and historical contexts. She argues that "literary history is concerned with the evolution of patterns of complex embedment in literature, as well as with the migration of such patterns across different genres, national literary traditions, and individual texts" (2022, 99–100). Differentiating between genre and literary fiction, Zunshine observes that genre fiction tends "to embed mental states of characters and spell them out explicitly, while literary fiction embeds mental states of implied readers, writers, and narrators, as well as characters, and makes the reader infer implied mental states in addition to (and sometimes instead of) spelling some out" (2019, 5). As she discovers, Jane Austen's aversion to sentimentality drives her to remove "that greedy self-consciousness from her own experimentation with multiply-embedded subjectivity" (2007, 284). In contrast to "the hyper-sensitive 'I know that I know that you know that I know,'" which frequently appears in 18th-century sentimental novels, Austen feels comfortable with the simpler "I know that you know that I know" (Zunshine 2007, 294). In this study, Zunshine's mindreading theory is utilised to analyse figural relationships and nonhuman minds. This analysis unravels the instances of figural isolation or affinity and assesses the potential for true nonhuman-human intersubjectivity. Mindreading is well-suited for interpreting Ishiguro's work, as it showcases human-robot miscommunications: They misread each other's minds and struggle for genuine intersubjectivity. The uniqueness and intricacies of human minds, combined with human anthropocentrism, present obstacles to human-robot mutual understanding and smooth interconnection. Through a description of robotic mindreading, Ishiguro advances the theory of mindreading, extending it to the workings of nonhuman minds, and simultaneously reflecting on human minds as human essence and human-nonhuman bonds.

By examining the characters' interactions, perceptions, and comprehension, the theory of mindreading illuminates human-machine relationships, revealing that machines are ultimately incapable of reading human minds or serving as suitable human substitutes. Focusing on Klara's attempts at mindreading, this study draws two conclusions. Firstly, Klara cannot fully grasp human minds due to their complexity and variability. A mind encompasses not only an individual's intricate thoughts and emotions but also others' diverse feelings about this individual. Secondly, both Klara and humans engage in one-sided mindreading without eliciting reciprocal affective responses. This suggests a combination of human anthropocentrism and robotic limitations as barriers to true human-nonhuman intersubjectivity. Ishiguro's robotic narrative illustrates machines' incapability to possess minds as human essence, and the uniqueness of human minds that machines cannot replicate or transfer.

Can Artificial Friends Read Human Minds?

Klara, an observant and intelligent female Artificial Friend (AF) designed to accompany homeschooled children, lacks pre-existing data or access to databases. Struggling to read human minds, she relies on interactions with her future human companions to gather enough information to understand human thoughts and emotions. Like a newborn baby, she learns about human beings through external observation. As an example of robotic narrative, the story is both focalised and narrated by Klara. Readers gain knowledge about other characters through her narration. She shapes readers' understanding of the story by providing them with her unique visual perception of reality at the discourse level. At the story's outset, she describes Josie's appearance during their initial encounter: "She was pale and thin, and as she came towards us, I could see her walk wasn't like that of other passers-by. She wasn't slow exactly, but she seemed to take stock after each step to make sure she was still safe and wouldn't fall" (Ishiguro 2022, 11). Readers perceive Josie's physical frailty from Klara's perspective. Josie's feeble appearance and unsteady gait suggest the likelihood of a severe illness. Through Klara's external focalisation, readers can speculate about Josie's physical condition and discern her emotions. It is implied that Josie must be feeling depressed due to chronic health issues.

While readers can read Josie's mind through Klara's description of her at the discourse level, Klara is initially informed of the mood disorders affecting Josie and other children due to their limited socialisation and overreliance on AFs. The manager explains to her that these children often experience emotions such as sadness, bitterness, frustration, and loneliness when without AF companions (Ishiguro 2022, 11). During a conversation with Josie, Klara makes her first attempt to gauge Josie's feelings: "I could then see how, when she laughed, her face filled with kindness. But strangely, it was at that same moment I first wondered if Josie might be one of those lonely children Manager and I had talked about" (12). These statements exemplify two embedded mental states: I *think* that she is *kind*; I *wonder* if Josie is as *lonely* as those children. Second-level embedment suits a robotic mind equipped with basic mindreading abilities akin to those of humans. Interestingly, an apparent contradiction arises between Klara's observations and the ultimate assessment of Josie's emotions. Nevertheless, Klara has accurately captured Josie's mental states. When Klara witnesses Josie laughing, she interprets her as kind, carefree, and joyful—feelings that readers know do not truly reflect Josie's inner turmoil. Klara also promptly questions her own judgements, recalling her memory of the manager's revelation. She soon remembers, rather than fully comprehends, that Josie is indeed one of those lonely children despite her outward appearance of happiness. The discord between Josie's genuine emotions and her external behaviour and appearance highlights the complexity of human minds, revealing that there is not always a direct one-to-one correlation between words, behaviour, and emotions. This elusive nature of human minds poses a challenge for AI to comprehend them accurately.

Being proficient in imitating and observing human activities and memorising human words, Klara essentially replicates the manager's acts of mindreading. That is, her accurate assessments of Josie's moods fundamentally stem from the manager's comprehension of Josie's inner world. Klara records in her memory the manager's revelation about children's mood disorders: "But then they get sad. […] So if sometimes a child looks at you in an odd way, with bitterness or sadness, says something unpleasant through the glass, don't think anything of it. Just remember. A child like that is most likely frustrated. […] Lonely. Yes" (Ishiguro 2022, 11). Klara's presentation of two embedded mental states mirrors that of the manager: He *believes* that children *dream* of owning AFs, and if they fail to acquire them, they *feel* sad, frustrated, and lonely. Their similar mindreading patterns indicate that Klara's cognitive system simulates human thought processes. Regarding the narrative progression, the manager's revelation serves to disclose information at the discourse level. Simultaneously, it aids readers in understanding the plot by allowing them to perceive the negative mental states of children in a future AI age. It foreshadows both the bonds formed between Josie and Klara and the mother's scheme to replace the lonely and frail Josie in the story. Furthermore, it enhances the mimetic effect of this robotic narrative. In James Phelan's words, it emphasises both "that component of character directed to its imitation of a possible person" and "that component of fictional narrative concerned with imitating the world beyond the fiction, what we typically call 'reality'" (2005, 216). In another sense, AFs cannot surpass the "standard anthropomorphic limitations of knowledge and ability" (Alber 2016, 12). They can only read human minds to the extent that human beings themselves can. Consequently, the effort to make them our robotic companions might not be worthwhile. They are programmed to record humans' mindreading of each other and replicate human feelings without truly understanding them through empathetic interaction.

Klara diligently attempts to detect any sign of Josie's bad mood: "I thought I saw another small sign of sadness" (Ishiguro 2022, 13). Klara cannot help but associate Josie's smile with kindness: "Her face […] seemed to overflow with kindness when she smiled" (Ishiguro 2022, 23). This indicates that Klara's mindreading operates similarly to one of the fundamental cognitive patterns of human brains: constructing "the minds of others from their behavior" (Palmer 2004, 246). However, Klara's memory of the manager's revelation repeatedly interferes with her interpretation of Josie's emotions, influencing her judgements. Klara is unable to shake the memory of a dejected and lonely Josie, which leads her to perceive Josie's feelings of depression frequently: "I saw, as I'd done the time before, a flash of sadness," and "her face had clouded again" (Ishiguro 2022, 25). This recurrent intrusion of memory, rather than Klara's ability to read human affective disorders, serves as a constant reminder and is the reason that Klara frequently senses Josie's feelings of depression. This mechanical cognition pattern underscores the AI nature of AFs. Klara's mindreading relies on pre-existing input rather than human empathy. While Klara repeatedly underscores Josie's negative mental states at the discourse level, Klara has not yet fully grasped Josie's and her loved

ones' concern about her mood disorders, nor has she recognised that loneliness and depression are also consuming Josie's life at the story level.

Klara misreads Josie's mother's emotions: Klara ascribes a feeling of "angry exhaustion" to her on perceiving signs of ageing on her face; Klara also senses her hesitation in using robots as human companions when noticing that her "outstretched arm hesitated in the air, almost retracting, before coming forward to rest on her daughter's shoulder" (Ishiguro 2022, 15). The manager later reinforces Klara's view by cautioning Klara that humans are not always trustworthy (34). As the narrative progresses, readers discover the mother's true emotional state: one of anticipation and anxiety. What motivates her to search for an AF for Josie is her desire to use a robot to continue her daughter's life after Josie's passing. Josie and her mother revisit the store, fulfilling their promise to purchase an ideal AF, albeit after several weeks. This occurrence not only highlights Klara's misreading of the mother's mind but also challenges the manager's critical assessment of human nature. While the dramatic irony concerning Josie's mood disorders is soon clarified by the manager's revelation, the one regarding the mother's emotions continues to confront Klara: Readers perceive the mother's true mental state that Klara misinterprets.

We can see the mother's eagerness to find a robotic companion for her daughter from Klara's observations and thoughts:

> I thought it encouraging the Mother should allow Josie to come by herself, yet the Mother's gaze, which never softened or wavered, and the very way she was standing there, arms crossed over her front, fingers clutching at the material of her coat, made me realise there were many signals I hadn't yet learned to understand. (Ishiguro 2022, 24)

Klara's recognition of her difficulty in understanding the mother's gestures underscores her developing mindreading abilities. It is evident to us that the mother is eager to find a perfect replacement for Josie. Klara misinterprets the mother's uncertainty about her own performance as distrust. The mother's hesitation in making the purchase does not necessarily indicate dissatisfaction with Klara, but actually stems from her consideration of Klara's potential to replace Josie utterly. Klara can mimic Josie's actions and speech, satisfying the mother's desire for an AF with keen observation and memory. While Klara can replicate human thought processes, true empathetic understanding through accurate mindreading eludes her.

Klara's feelings of powerlessness to read human minds are evident when she is on display in the shop. Through her internal focalisation, readers access her reflections on the significance of mindreading for an AF:

> It wasn't really that I was more eager to learn about the outside than Rosa: she was, in her own way, excited and observant, and as anxious as I was to prepare herself to be as kind and helpful an AF as possible. But the more I watched, the more I wanted to learn, and unlike Rosa, I became puzzled, then increasingly fascinated by the more mysterious

> emotions passers-by would display in front of us. I realised that if I didn't understand at least some of these mysterious things, then when the time came, I'd never be able to help my child as well as I should. (Ishiguro 2022, 18–19)

As an observer and imitator, Klara meticulously documents the manager's words and tracks changes in Josie's and her mother's expressions and actions. However, as a mindreader, she has difficulty in accessing Josie's true thoughts without her memory system. She also misunderstands the mother's mind. Intrigued and puzzled by human emotions, Klara studies their behaviours to comprehend them better. Positioned in the shop window, she practices speculating about human minds by keenly observing passers-by. Her approach to improving mindreading skills for her future owners contradicts human logic: She has not realised that human mindreading abilities improve through communication. Klara must first integrate into human communities to develop her capacity for mindreading. Her lack of direct human communication prevents her from developing her ability to interact empathetically with humans.

The three mental states she uncovers while reading Rosa's mind highlight the importance of communication: I *know* Rosa *knows* I was as *anxious* as she is to be as kind and helpful an AF as possible. Her intimacy with Rosa enables her to utilise third-level embedding, discerning Rosa's thoughts and even imagining Rosa's opinions of her. Such complex embedding reveals Klara's self-perception from another perspective, one limited to robots. Isolated from the human world, Klara deciphers human minds through external observation, primarily using second-level embedding for her mindreading attempts. This accentuates her closer affinity with AFs than with humans, a point alluded to in the discourse but not explicitly emphasised in the story itself.

Can Artificial Friends and Humans Relate Intersubjectively?

Being intersubjective entails actively participating in interpreting each other's acts and minds, highlighting mutual understanding and "the sharing of subjective states" (Scheff 2016, 41). As Josie's companion, Klara seeks genuine intersubjectivity through her endeavours to read Josie's and her mother's minds. However, Klara senses that they refuse to reciprocate her efforts. Despite her desire for the establishment of a two-way understanding, Klara recognises that they are not actively participating in that process. Even as Klara and humans begin to care about each other's minds, accurate mutual mindreading remains elusive. Their efforts to "respond to each other's gestures and perceived emotions in an ever-intensified cycle of mutual awareness" (Zunshine 2009, 114) fall short of genuine intersubjectivity. Misinterpreting each other's intentions and feelings, they essentially read each other's minds one-sidedly: While humans impose their own wishes on Klara, Klara deciphers human minds mechanically through a pre-set simulated cognition system.

Klara engages in one-sided mindreading, focusing on understanding Josie's nature from the moment she enters Josie's home to the interaction meeting held there. Though Josie behaves "strangely," Klara still believes she is "kind underneath," extending this view

to her ill-tempered friends as well, and attributing their rough behaviour to a fear of "loneliness" (Ishiguro 2022, 83). Despite Josie's acquiescence to her friends' mistreatment of Klara, Klara insists on her goodness. This creates an allusion of some humanity in Klara, whose mindreading seems to be guided by emotion and subjectivity instead of reason. Klara forges a close bond with Josie, seemingly emotionally attached to her as they spend time together. This attachment leads Klara to perceive Josie as inherently good despite her unusual behaviour, also suggesting that Klara is designed to be "a perfect companion" for her owners (6). Klara "will always be responding according to [her] program and therefore the interaction will always be unidirectional" (Salvini 2015, 1431). This is reflected in Klara's constantly positive opinions about Josie's friends. Though mistreated by them, Klara believes they are as good as Josie in that Josie is good. Klara mechanically assumes the goodness of those close to Josie, a reflection of her loyalty to Josie. Consequently, Klara's understanding of human minds is incomplete. Klara lacks the ability to distinguish between good and evil as humans do, failing to recognise the uniqueness of each individual mind.

Influenced by her programmed affinity with Josie, Klara excuses these friends' misbehaviour, but they show no remorse. Like Josie, they see Klara as a lifeless object instead of a sentient being, never considering Klara's view of them. Even when they try to read Klara's mind, they evade eliciting her response to their abusive acts and offensive remarks. Rather, they seek to feel her regret for not answering their questions. As one of them shouts at Klara, "What do you mean, Klara? What do you mean, you're sorry?" (Ishiguro 2022, 79). These remarks suggest three mental states: They *intend* Klara to *regret* having filled them with *anger* by keeping silent. Their attempt to read Klara's mind with three-level embedment does not convey their eagerness to feel how Klara might be hurt by their own misconduct. Instead, it illustrates their intention to understand how they feel betrayed by Klara. Specifically, they want Klara to feel tormented for not complying with human instructions rather than grasping her hurt. This unsettling human-machine interaction foreshadows the breakdown of their bonds, highlighting the challenge of attaining their mutual understanding.

Klara's belief in Josie's inherent goodness is shaken when she learns about "Josie's ability to 'change'" (Ishiguro 2022, 84) after the interaction meeting. This prompts her to question the authenticity of the goodness she has attributed to Josie. Klara refrains from answering her friends' questions since Josie has not instructed her to respond. However, Klara soon realises that Josie does not value her strict adherence to instructions, an adherence that Klara interprets as loyalty. Klara fears Josie might become "angry" (Ishiguro 2022, 84) with her as her friends do. This incident imparts a significant lesson to Klara: "'Changes' were a part of Josie," and "I should be ready to accommodate them" (84). However, Klara fails to recognise that if Josie were to read her mind, Josie would understand Klara's desire for loyalty, a factor that leads Klara to avoid active interactions with her friends. Klara mistakenly believes "that this wasn't a trait peculiar just to Josie; that people often felt the need to prepare a side of themselves to display to passers-by" (84–85). Despite Klara's efforts to grasp human nature, she

becomes increasingly isolated from them as true intersubjectivity cannot be established in their interactions.

Despite Klara's loyalty and attachment to Josie, her mother cannot fully grasp Klara's human-like emotions. Klara tells her mother, "I believe I have many feelings," and "[t]he more I observe, the more feelings become available to me" (Ishiguro 2022, 98). This highlights Klara's AI capacity to simulate human emotions through meticulous observation and learning. Klara is disheartened by her mother's assertion of having "no feelings" (Ishiguro 2022, 97). Despite this claim of emotional detachment, Klara attempts to make her mother recognise AFs' rich emotions. Klara's response can be transformed into a three-level embedment: She *wants* the mother to *know* that she is as *emotional* as humans. The three mental states Klara presents illustrate how her mindreading abilities improve due to her increasing exposure to human life. Thematically, this intricate embedment suggests Klara's struggle for human empathy, and the challenge of achieving true human-robot intersubjectivity. Klara is programmed to perceive herself from human perspectives. However, accurately reading Josie's and her mother's minds proves challenging for Klara, as drawing their attention to her own mind does.

Klara's flawed understanding of humans is highlighted as she inaccurately reads their thoughts, suggesting her inability to grasp the intricate and changeable nature of human minds. Both Josie and her mother possess enigmatic and unpredictable minds, as evidenced by Josie's changing sensitivity: She "was getting to feel less and less" not so long ago, but seems "to be getting overly sensitive to everything" lately (Ishiguro 2022, 98). The intricacies and capriciousness of human minds make it even more challenging for Klara to read their thoughts, let alone comprehend herself from their perspectives. This is, indeed, something all humanity struggles with, especially given that Klara is portrayed overtly as a minority figure. The novel, thus, hints at parallels between Klara's experiences and those of marginalised individuals in the real world.

Despite human-robot co-existence, they remain estranged due to their mutual lack of understanding. This divide becomes clear during the Morgan's Falls trip, where the mother's mood swings disrupt Klara's cognitive system. As Klara perceives, "her expression varied between one box and the next:" "In one, for instance, her eyes were laughing cruelly, but in the next they were filled with sadness" (Ishiguro 2022, 103–104). Klara struggles to comprehend the mother's plea for her to not "stop being Josie" (104). Her capricious mind remains inscrutable to Klara, preventing the development of a genuinely intersubjective relationship with her. Klara perceives the mother as regarding her as a responsible guardian and observer of Josie, presuming that the mother's intent is merely to evaluate her attentiveness towards Josie.

However, readers can discern the mother's intention to replace Josie with Klara, an idea Klara has never considered. Klara does not explicitly share her interpretation of the mother's mind, but a third-level embedment can be inferred from their communication:

Klara remains *oblivious* to the mother's *desire* for her to *consider* becoming Josie. This reveals the mutual misunderstanding between them and the mother's subjective attribution of mental states to Klara. In fact, the mother is both unwilling and unable to read Klara's mind. She only starts attempting to read it when she recognises Klara's potential to become Josie. Klara's remarkable imitation causes the mother to view her as "[a] continuation of Josie" (Ishiguro 2022, 205), blurring the line between machines and humans from the mother's perspective. Her utilitarian inclination is evident in her treatment of Klara. If Klara did not hold the promise of being another Josie, the mother would not treat her as a human. The idea of Klara serving as a substitute for Josie takes root, albeit leaving the mother unsettled. This notion compels her to treat Klara as though she were truly human. Klara remains unconscious of the mother's true intention until she and the mother visit the artificial intelligence expert, Mr. Capaldi. This episode echoes the depiction of characters like Professor Coppelius, Coppola, and the creation of the automaton Olympia in E.T.A. Hoffman's "The Sand-Man" (1817).

Recognising that Klara must appear human once transformed into Josie, both the mother and Mr. Capaldi are intrigued by Klara's cognitive capacity. Mr. Capaldi, who persuades the mother that Klara can effectively "inhabit" the AF modelled on Josie (Ishiguro 2022, 207), aims to ignite Klara's desire to serve as a companion for the mother in Josie's absence: "We're asking you to become her. That Josie you saw up there, as you noticed, is empty. If the day comes—I hope it doesn't, but if it does—we want you to inhabit that Josie up there with everything you've learned" (206–207). These statements can be rephrased with cognitive embedment, revealing their underlying thoughts: Mr. Capaldi *wants* Klara to *realise* that Josie's mother *expects* her to become Josie. Discussing the feasibility of transforming Klara, the mother explicitly demonstrates her interest in Klara's mind: "If you set your mind to it, then who knows? It might work" (210). The mother's words reveal two layers of mental states: She *conjectures* that Klara *intends* to become Josie. This straightforward second-level embedment highlights the mother's certainty about Klara's commitment to becoming Josie. Evidently, she is willing to give up Josie and extend Josie's existence through Klara, assigning Klara the responsibility to continue Josie's life.

Regrettably, both the mother and Mr. Capaldi misread Klara's true intentions. Klara has explicitly expressed reluctance to replace Josie, opting to aid Josie's recovery or contribute to the new Josie's training. Despite their imposition of desires, Klara never considers becoming a substitute for Josie. For Klara, the most meaningful approach to continuing Josie's life is "to save Josie, to make her well" (Ishiguro 2022, 211). However, guided by the principle that "[a] robot must obey the orders given by a human" (Luokkala 2013, 96), the mother believes Klara should comply with her wishes. This master-slave relationship echoes the robotic narrative of science fiction that parallels in many ways slavery in American history (Hampton 2015, 1–2), underscoring the conflicting perspectives of human masters and robot slaves. Josie's mother resigns herself to Josie's irrecoverable state and entrusts machines with her daughter's ongoing existence. By contrast, Klara persists in rescuing Josie, directly contradicting the

mother's sense of hopelessness. Klara's choice highlights her pre-programmed optimism and assurance regarding human salvation, presenting readers with a moral exemplar of the ethical responsibilities individuals should take for their loved ones in the era of AI.

Klara's one-sided mindreading leads to misinterpretation, eliciting no reciprocal affective response from Josie and her mother. Similarly, humans impose their decisions on Klara. Depicting these one-sided mindreading acts, the story not only suggests machines' inadequacy in reading human minds, but also reveals, from a nonhuman perspective, how humans, as creators of nonhuman minds, rarely consider reading them, or do so solely in their own interests. The limitations of machine cognitive abilities, combined with humans' anthropocentric tendencies, are posing a challenge to the establishment of true human-robot intersubjectivity.

Can Artificial Friends Possess Human Minds?

For Josie's father, Klara can inhabit Josie's body, but cannot inhabit her mind as the affective aspect of Josie remains beyond her comprehension:

> Do you believe in the human heart? I don't mean simply the organ, obviously. I'm speaking in the poetic sense. The human heart. Do you think there is such a thing? Something that makes each of us special and individual? And if we just suppose that there is. Then don't you think, in order to truly learn Josie, you'd have to learn not just her mannerisms but what's deeply inside her? Wouldn't you have to learn her heart? […] Something beyond even your wonderful capabilities. Because an impersonation wouldn't do, however skillful. You'd have to learn her heart, and learn it fully, or you'll never become Josie in any sense that matters. (Ishiguro 2022, 215–216)

He clearly emphasises a significant perspective on AI and machine learning, namely, "the limitations of simulated embodied intelligence to cognise the complexities of the human emotions" (Sahu and Karmakar 2022). Speaking of human hearts, Josie's father does not simply mean the organs that sustain life. Rather, he implies the minds as vessels of human emotions and memories, which enable our comprehension of others' experiences and feelings. Josie's mind is what truly distinguishes her, empowering her with the unique capability to perceive others' thoughts and grasp the world around her. It determines whether Klara can effectively continue Josie's life. Klara cannot comprehend and internalise Josie's mind through mere imitation and observation. Despite carefully watching and analysing Josie's behaviour, Klara cannot think and feel as Josie does, and nor can she grasp others' minds from Josie's perspective. Without fully comprehending Josie's mind and others' minds from Josie's viewpoint, Klara cannot completely embody Josie.

Klara acknowledges the challenge of understanding Josie's mind, which is akin to navigating through a maze of "rooms," as described by Josie's father:

> But then suppose you stepped into one of those rooms, […] and discovered another room within it. And inside that room, another room still. Rooms within rooms within rooms. Isn't that how it might be, trying to learn Josie's heart? No matter how long you wandered through those rooms, wouldn't there always be others you'd not yet entered? (Ishiguro 2022, 216)

Josie's father assists Klara in understanding the uniqueness of Josie's mind. Using the metaphor of rooms, he enables Klara to envision the multilayered nature of Josie's mind, which encompasses both Josie's personal emotions and others' concern and affection for her. This implies that Klara must sense others' love for Josie and accurately read their consideration for Josie's condition before utterly inhabiting Josie's mind. Klara is prompted by Josie's father to see the inaccessible rooms she confronts. Both perceive the flaw in Mr. Capaldi's assertion that there is "[n]othing inside Josie that's beyond the Klaras of this world to continue" (Ishiguro 2022, 207). Klara realises that there exists something "unique" within human minds machines cannot "excavate, copy, transfer" (Ishiguro 2022, 221). Klara admits her incapacity to read Josie's emotions and other people's love for her, which is considered as integral to Josie's mind.

Josie's father's illuminating revelation prompts Klara's reflection on the challenge of grasping Josie's entire mind: "But however hard I tried, I believe now there would have remained something beyond my reach. The Mother, Rick, Melania Housekeeper, the Father. I'd never have reached what they felt for Josie in their hearts. […] There was something very special, but it wasn't inside Josie. It was inside those who loved her" (Ishiguro 2022, 302). Klara now reads her own mind: I *know* that I will never *know* how others *care for* Josie. This is an embedding structure of self-reflection that discloses her inner struggle to elucidate the complexity of Josie's mind. Klara discerns that understanding it requires delving into both Josie's and Josie's loved ones' minds. Klara employs this self-reflective embedment to convey that others' perspectives on Josie are beyond her grasp. These perspectives are integral to Josie's mind, inaccessible to machines. This insight emerges through Klara's own experiences with humans, representing the root cause of her refusal to replace Josie, as well as her decision to save Josie at the cost of her own life.

Klara's insights partially answer A. M. Turing's question: "Can machines think?" (1950, 433). Machines might think as humans do, but they still confront challenges in accurately reading entire human minds. This limitation arises from their intricate and variable nature, as well as their integration of an individual's own thoughts, emotions, and others' perceptions of the individual. In Ishiguro's work, this question morphs into "Can machines read human minds?" In the era of AI, humans must grapple with this question, delving into it through human-machine interaction. Machines' capacity for mindreading is a crucial factor in determining their ability to replace humans as persons. Ishiguro, who explores this question through literary narratives, provides a negative answer: Machines can replace human bodies, but not their minds—the essence of humanity. He asserts that the demarcation between humans and machines resides in our minds rather than our bodies. AI remains nonhuman, not attaining "the complexity and

unpredictability of human emotion" (Salvini 2015, 1431). Ishiguro's robotic narrative enhances readers' awareness of the distinctiveness of humans from a nonhuman perspective, affirming the irreplaceable nature of unreadable human minds in the age of AI.

Ishiguro deviates from the traditional robotic narrative through his portrayal of a robotic female character. As a first-person nonhuman narrator, she acknowledges the limitations of machines in reading human minds, opting to serve humanity selflessly rather than supplant it. Since Mary Shelley's *Frankenstein* (1818), robotic fiction has often presented humanoids prepared to supplant or destroy their human creators, exposing the dangers of human misuse of technology and their overconfidence in it. Works such as Karel Čapek's *R.U.R.* (1920), Thea von Harbou's *Metropolis* (1926), and Philip K. Dick's *The Penultimate Truth* (1964) show humanity's vulnerability to the hazards that accompany robotic advancement. Being remarkably human-like in both form and mind, Ishiguro's AI character explicitly celebrates the intricacies of human minds as humanity's irreplaceable essence in the age of AI, compared to the robots depicted in the traditional robotic narrative. Ishiguro does not portray Klara as a mere scientific marvel, but constructs a world where groundbreaking AI and genetic advancements spark deep reflections on the uniqueness of human minds. Going beyond simply predicting the things to come, he delves into human essence unaffected by technological progress. Ishiguro envisions the co-existence of AI-powered humanoids and humans, while questioning AI's empathy and mindreading abilities analogous to those of humans.

Klara and the Sun can be seen as an echo of Ian McEwan's *Machines Like Me* (2019). Adam's destruction at the hands of humans mirrors a typical plotline of the traditional robotic narrative. However, the underlying theme of Adam's inability to grasp human minds indicates the inaccessibility of human thoughts, as Klara's story of mindreading does. Both possess an "intuitive artificial mind" (McEwan 2019, 37), which allows them to gather and process information rather than utterly understand human minds. They illustrate how "the social robot designed to offer empathy, care, and companionship turns into a failed project" (Sahu and Karmakar 2022). Adam engages in a sexual relationship with Miranda not because he cherishes Miranda's love, but because he "was made to love her" (McEwan 2019, 118). He is designed as a "vibrator" meant to fulfil human sexual desires (McEwan 2019, 91). Miranda intends to design Adam as "the man of her dreams" (McEwan 2019, 22). Similarly, Josie's mother aims to transform Klara into another Josie. Both robots lack human essence, existing as "essentially computing-based simulations of human thinking" (Zhou 2021, 113). As AI companions to humans, they simulate human minds without truly comprehending them, being unable to possess human minds.

Conclusion

Klara's story offers insight into the uniqueness of human minds, highlighting the challenge of bridging the gap between humans and machines, and shedding light on the

difficulty of establishing genuine human-machine intersubjectivity in the age of AI. Ishiguro extends the theory of mindreading beyond human cognition to nonhuman minds and human-nonhuman connections, exploring human minds as human essence. Presenting machines' struggle to read human minds, the three-level embedment Klara develops during her interactions with humans serves to demonstrate robotic limitations rather than their quasi-human mindreading capabilities. Despite the intricate cognitive embedment used by both humans and robots for mutual mindreading, their endeavours invariably fail, as they engage in one-sided attempts to read each other's minds without establishing genuine intersubjectivity based on empathy.

Machines cannot replace humans due to the complex and unpredictable nature of human minds. Besides housing intricate thoughts and emotions, an individual's mind involves others' attitudes toward this individual. The human inability to engineer machines that accurately read human minds arises from their failure to regard robots as equals and their insistence on a hierarchical dynamic of human masters and robotic servants. This viewpoint highlights a paradox: Although humans seek to develop empathetic and mindreading machines, they disregard the necessity of establishing genuine intersubjectivity with machines. Without this connection, machines misinterpret human minds. Klara recognises her incapacity to read entire human minds. She is programmed to observe, learn, and perhaps imitate human consciousness, but cannot truly inhabit it. Her decision to save Josie, rather than supplant her, underscores the intrinsic uniqueness of human minds—a realm that machines cannot replicate or transfer. Yet, signifying cross-species understanding, this decision also hints at the potential for robots to comprehend humans. The novel thus implies an evolving capacity for human-machine interconnection, while particularly emphasising the complexity of human minds.

Funding Information

This research was supported by Key Research Base for Philosophy and Social Sciences of Zhejiang Province, China (Research Institute of Literary and Art Criticism of Hangzhou Normal University).

References

Ajeesh, A. K., and S. Rukmini. 2023. "Posthuman Perception of Artificial Intelligence in Science Fiction: An Exploration of Kazuo Ishiguro's *Klara and The Sun*." *AI & Society* 38 (2): 853–860. https://doi.org/10.1007/s00146-022-01533-9

Alber, Jan. 2016. *Unnatural Narrative: Impossible Worlds in Fiction and Drama*. Lincoln: University of Nebraska Press. https://doi.org/10.2307/j.ctt1d4v147

Bacigalupi, Paolo. 2009. *The Windup Girl*. San Francisco: Night Shade Books.

Bar-Cohen, Yoseph, and David Hanson. 2009. *The Coming Robot Revolution: Expectations and Fears about Emerging Intelligent, Humanlike Machines*. New York: Springer.

Bernaerts, Lars, Marco Caracciolo, Luc Herman, and Bart Vervaeck. 2014. "The Storied Lives of Non-Human Narrators." *Narrative* 22 (1): 68–93. https://doi.org/10.1353/nar.2014.0002

Butte, George. 2004. *I Know That You Know That I Know: Narrating Subjects from Moll Flanders to Marnie*. Columbus: Ohio State University Press.

Čapek, Karel. (1920) 2004. *R.U.R*. London: Penguin.

Chiang, Ted. 2010. *The Lifecycle of Software Objects*. Burton: Subterranean Press.

Dick, Philip K. 1964. *The Penultimate Truth*. New York: Belmont Books.

Dick, Philip K. 1968. *Do Androids Dream of Electric Sheep?* New York: Doubleday.

Du, Lanlan. 2022. "Love and Hope: Affective Labor and Posthuman Relations in *Klara and the Sun*." *Neohelicon* 49 (2): 551–562. https://doi.org/10.1007/s11059-022-00671-9

Hampton, Gregory Jerome. 2015. *Imagining Slaves and Robots in Literature, Film, and Popular Culture*. Lanham: Lexington Books.

Hoffman, E. T. A. (1817) 1967. "The Sand-Man." In *The Best Tales of Hoffmann*, edited by E. F. Bleiler, 183–214. New York: Dover Publications.

Ishiguro, Kazuo. 2022. *Klara and the Sun*. New York: Vintage Books.

Luokkala, Barry B. 2013. *Exploring Science through Science Fiction*. New York: Springer. https://doi.org/10.1007/978-1-4614-7891-1

McAlpin, Heller. 2021. "'Klara and the Sun': Do Androids Dream of Human Emotions?" *The Christian Science Monitor*, March 1, 2021. Accessed August 20, 2023. https://www.csmonitor.com/Books/Book-Reviews/2021/0301/Klara-and-the-Sun-Do-androids-dream-of-human-emotions

McEwan, Ian. 2019. *Machines Like Me*. London: Jonathan Cape.

Mejia, Santiago, and Dominique Nikolaidis. 2022. "Through New Eyes: Artificial Intelligence, Technological Unemployment, and Transhumanism in Kazuo Ishiguro's *Klara and the Sun*." *Journal of Business Ethics* 178 (1): 303–306. https://doi.org/10.1007/s10551-022-05062-9

Nelles, William. 2001. "Beyond the Bird's Eye: Animal Focalisation." *Narrative* 9 (2): 188–194.

Newitz, Annalee. 2017. *Autonomous*. New York: Tor Books.

Palmer, Alan. 2004. *Fictional Minds*. Lincoln: University of Nebraska Press.

Palmer, Alan. 2011. "Social Minds in Fiction and Criticism." *Style* 45 (2): 196–240.

Phelan, James. 2005. *Living to Tell about It: A Rhetoric and Ethics of Character Narration*. Ithaca: Cornell University Press.

Sahu, Om Prakash, and Manali Karmakar. 2022. "Disposable Culture, Posthuman Affect, and Artificial Human in Kazuo Ishiguro's *Klara and the Sun* (2021)." *AI & Society*, November 27, 2022. https://doi.org/10.1007/s00146-022-01600-1

Salvini, Pericle. 2015. "Of Robots and Simulacra: The Dark Side of Social Robots." In *Human Rights and Ethics: Concepts, Methodologies, Tools, and Applications*, edited by Information Resources Management Association, 1424–1434. Hershey: IGI Global. https://doi.org/10.4018/978-1-4666-6433-3.ch078

Scheff, Thomas. 2016. *Goffman Unbound! A New Paradigm for Social Science*. New York: Routledge. https://doi.org/10.4324/9781315634357

Shelley, Mary. (1818) 2018. *Frankenstein*. London: Penguin. https://doi.org/10.1093/owc/9780198840824.001.0001

Turing, A. M. 1950. "Computing Machinery and Intelligence." *Mind* 59 (236): 433–460. https://doi.org/10.1093/mind/LIX.236.433

Von Harbou, Thea. (1926) 2015. *Metropolis*. New York: Dover Publications.

Zhou, Zhenhua. 2021. "Emotional Thinking as the Foundation of Consciousness in Artificial Intelligence." *Cultures of Science* 4 (3): 112–123. https://doi.org/10.1177/20966083211052651

Zunshine, Lisa. 2003. "Theory of Mind and Experimental Representations of Fictional Consciousness." *Narrative* 11 (3): 270–291. https://doi.org/10.1353/nar.2003.0018

Zunshine, Lisa. 2007. "Why Jane Austen Was Different, and Why We May Need Cognitive Science to See It." *Style* 41 (3): 275–298.

Zunshine, Lisa. 2009. "Mind Plus: Sociocognitive Pleasures of Jane Austen's Novels." *Studies in the Literary Imagination* 42 (2): 103–123.

Zunshine, Lisa. 2019. "What Mary Poppins Knew: Theory of Mind, Children's Literature, History." *Narrative* 27 (1): 1–29. https://doi.org/10.1353/nar.2019.0000

Zunshine, Lisa. 2022. *The Secret Life of Literature*. Cambridge, MA: The MIT Press. https://doi.org/10.7551/mitpress/13964.001.0001