

Reciprocal Relationality in the Emerging Age of Programmed Social Humanoids in Kazuo Ishiguro's *Klara and the Sun*

Eun Hye Lee

Ewha Womans University

[Abstract]

This article examines how Kazuo Ishiguro's *Klara and the Sun* (2021) reconfigures human subjectivity in an era increasingly shaped by social humanoids and artificial intelligence. Departing from essentialist notions of interiority, the novel relocates the question of the human condition from ontological essence toward reciprocal relationality. While Ishiguro's first-person narration encourages readers to attribute sentient agency to Klara, the novel simultaneously insists on her status as a programmed social humanoid whose interactions with human characters are structurally non-reciprocal. Drawing on posthuman theory and human-robot interaction (HRI) studies, this article argues that *Klara and the Sun* foregrounds reciprocity as the decisive threshold distinguishing relational simulation from genuine relational participation. Although Klara consistently extends care and attentiveness toward others, she fails to participate in relationships characterized by mutual recognition and vulnerability. Her relationality is thus asymmetrical, paradoxically

reinscribing her within a functionalist logic. By staging this structural asymmetry at the core of human-robot interaction, *Klara and the Sun* not only interrogates the shifting boundaries between humans and nonhumans but also offers a critical reflection on how emerging social humanoids may reshape the foundations of relational life itself.

Key Words: *Klara and the Sun*, human-robot interaction, programmed social humanoids, reciprocity, relational ontology, artificial intelligence

I. Rethinking the Human in the Emerging Age of Social Humanoids

What does it mean to be human in a world where qualities once thought to define humanity are increasingly extracted, encoded, and reproduced by artificial others? Within the techno-cultural landscape depicted in Kazuo Ishiguro's *Klara and the Sun* (2021), human characters appear as post- or transhuman subjects, while humanoids are designed not merely as advanced machines but as forms of artificial "life" endowed with emotional responsiveness and cognitive awareness. These humanoids render humans progressively replaceable by functioning "so effectively and accurately that it ultimately proves more desirable than human efforts" (Senechal par. 5). In *Klara and the Sun*, humanoids have not only displaced human labor but have also assumed the role of social companions for children, and most disturbingly, are even considered as potential replacements for children themselves. Klara, the social humanoid protagonist, is an Artificial Friend (A.F.) for Josie, who is suffering from a chronic illness caused by the genetic transhuman enhancement procedure known as

“lifting.” As the narrative unfolds, however, it is revealed that Josie’s mother, Chrissie, has quietly come to regard Klara as a possible substitute for her daughter should the illness prove fatal. Chrissie’s readiness to accept humanoid replacement is not unprecedented as she had previously commissioned a bereavement doll to stand in for her eldest daughter, Sal, following her death. Such alarming ease with which humanoids are considered replacements for humans suggests that contemporary technological advancement has moved beyond concerns about machines functioning like humans, now raising profound ethical and ontological questions regarding the human condition: if artificial beings such as Klara can perceive, process, and respond to human life in ways indistinguishable from humans, what, then, differentiates the human subject?

We have thus entered an era in which the conceptual boundary separating humans from machines is no longer merely unstable but increasingly obsolete. In *Klara and the Sun*, Paul, Josie’s father, reflects on this ontological uncertainty by addressing Klara—and, by extension, the reader—with a series of thought-provoking questions:

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?* (215, emphasis added)

Here, “the human heart” functions as a concept for subjectivity, an irreducible “deep inside” that protects human uniqueness from technological replication. This understanding of the human is rooted in Enlightenment ideals of interiority, which conceptualize selfhood and autonomy through introspection and consciousness rather

than through external factors or relations. Within this inward turn, emotion emerges as an internal guide to ethical insight, moral reasoning, and social connection, such that “the everyday language of emotion is based on the presumption of interiority” (Ahmed 8). In this sense, the “human heart” symbolizes “both cognitive and emotional capabilities” (Li and Eddebo 124), functioning as the ontological framework through which human singularity has traditionally been determined.

Such an ontological framework grounded in Enlightenment ideals of interiority and emotion, however, proves increasingly inadequate within the contemporary techno-cultural context. Social robots already demonstrate the ability to interpret, simulate, and participate in human emotional and social interaction. As a result, emotional depth and reflective selfhood can no longer be claimed as exclusively human attributes. The insufficiency of this ontological perspective further stems from a fundamental epistemological restriction to access or verify the interior consciousness of social robots, or any sentient or potentially conscious artificial intelligence (AI). As Joanna K. Malinowska argues in “What Does It Mean to Empathise with a Robot?” (2021), humans possess no direct knowledge of the emotional or cognitive states of social robots, assuming such states exist at all. “We have no idea what it is to be a robot,” she asserts, emphasizing that our assumptions are based not on certainty but rather on human perception: “we only project our fantasies, beliefs, experiences, and expectations onto these entities” (370). Even if social robots were to develop cognitive-emotional processes, Malinowska further suggests that humans might nonetheless fail to recognize or even acknowledge them (372). This epistemic limitation renders artificial interiority fundamentally inaccessible, making it impossible to definitively distinguish genuine emotional or cognitive agency from algorithmic simulations, while simultaneously destabilizing claims of human distinctiveness grounded in interiority. Within this context, the

ontological framework of defining the human through supposedly exclusive emotional consciousness — figured in *Klara and the Sun* as “the human heart” (215) — no longer provides a stable criterion of distinction.

The emerging age of social humanoids therefore necessitates a reconceptualization of the human, shifting from an ontological definition of autonomous subjectivity toward a relational understanding of subjectivity as constituted through interaction. In the absence of epistemic certainty regarding whether artificial beings such as Klara possess agency, consciousness, or emotional interiority, evaluative judgments ultimately depend on the interpretive assumptions and responses of human interlocutors. Social humanoids, in other words, cannot be meaningfully conceptualized apart from the humans who interact with them. Therefore, rather than relying on a binary ontological opposition between humans and machines, critical analysis now focuses on an examination of the relations that develop between the two parties. This relational turn lies at the heart of the interdisciplinary field of Human-Robot Interaction (HRI) studies, which explores how humans perceive, interpret, and engage with social robots. Moving beyond purely technical emphasis on functionality and efficiency, HRI studies primarily examine social communications and relations between humans and robots, marking “a transition from talking about a robot to talking to and with a robot” (Sahu and Karmakar 1352). As AI becomes increasingly embedded in everyday social life, what ultimately matters is not whether artificial agents possess consciousness as a means of distinguishing humans from nonhumans, but how their presence reconfigures “the whole net of emotional and behavioural [sic] interdependencies between the different social agents” involved (Malinowska 372). From this perspective, the more pressing question now is how the participation of social robots and AI reshapes social and emotional relations, and, in turn, how these transformations compel a rethinking of what it means to be human.

Subjectivity, in this account, is no longer defined by an irreducible internal core but by relations of recognition, responsiveness, and interaction within shared social worlds. This perspective aligns with posthumanist critiques of liberal humanist subjectivity, which challenge “the presumption that there is an agency, desire, or will belonging to the self and clearly distinguished from the ‘wills of others’” (Hayles 3). Within posthuman theory, subjectivity is understood as collective, heterogeneous, and distributed, emerging through networks of interaction, mediation, and communication rather than residing in a singular self-contained interiority. Subjective experience is thus reconceptualized as co-experiential, one constituted through shared fields of participation that remain “inaccessible to the isolated Cartesian self” (Li and Eddebo 128). Posthuman collective subjectivity thus affirms a relational ontology in which meaning and agency arise through communicative engagement among heterogeneous human and nonhuman actors who together sustain the conditions of subjectivity itself.

This turn toward relationality as a redefinition of human ontology, however, also exposes a residual anthropocentric bias that privileges human psychological attitudes, beliefs, and interpretive authority. Recent literary criticism on *Klara and the Sun* increasingly foregrounds relationality as the primary framework through which both the novel and Klara are analyzed, critiquing the anthropocentric perspectives of the human characters that disregard the humanoid protagonist’s moral agency and labor of care. In “The Humanity of the Non-Human” (2023), Oliver Li and Johan Eddebo employ a “functionality vs. relationality” model to argue that it is the human characters who adhere to a functionalist logic that reduces cognitive and behavioral performances to “some mechanistic process” (127). While these human characters persistently perceive Klara as a disposable product, rendering her a mere machine, Li and Eddebo emphasize that she nonetheless exhibits emotional awareness and conscious agency. It is, in fact, Klara who articulates an alternative perspective by

recognizing that “[h]uman existence or worthy existence of interacting minds is essentially based on *relationality*” (Li and Eddebo 127, emphasis original). Yet even within such relational readings, Klara’s moral, emotional, and cognitive status remains fundamentally precarious, as the ultimate criterion of value resides in human acknowledgment and judgment. Therefore, despite being “depicted as being more virtuous and humane than the actual human characters involved, [Klara] suffers abandonment from her friends, in spite of her acts of heroism” (128). Relationality alone thus risks reproducing the very anthropocentric hierarchies it seeks to dismantle, as it remains structured by human-centered evaluative perspectives. For this reason, an emphasis on relationality must move beyond the mere existence of relationship to interrogate the conditions under which relationality becomes meaningful—specifically, whether human-robot interactions entail mutual participation and reciprocity, or whether they remain fundamentally unidirectional. Therefore, reciprocity emerges as the decisive threshold that distinguishes relational simulation from genuine relational participation: without mutual recognition and response, relationality collapses into functionality.

Accordingly, this article argues that *Klara and the Sun* reconfigures the category of the human by displacing essentialist notions of interiority in favor of a relational ontology grounded in reciprocity. While Ishiguro’s first-person narration encourages readers to attribute sentient agency and emotional depth to Klara, the novel simultaneously insists on her status as a programmed social humanoid whose interactions with human characters are structurally non-reciprocal. Klara consistently extends care and attentiveness toward others but she fails to participate in relationships characterized by mutual recognition and vulnerability. As a result, it is Klara’s relationality that is asymmetrically configured and paradoxically reinscribed within a functionalist logic. By staging this structural asymmetry at the core of

human-robot interaction, *Klara and the Sun* not only interrogates the shifting boundaries between humans and nonhumans but also offers a critical reflection on how emerging social humanoids may reshape the foundations of relational life itself.

II. Klara: Sentient Autonomous Consciousness or Programmed Social Humanoid?

Klara and the Sun stages one of its most incisive ethical provocations concerning Klara's ontological status during her visit to the home of Rick, Josie's only close friend. Upon greeting Klara, his mother, Miss Helen, remarks: "One never knows how to greet a *guest* like you. After all, are you a guest at all? Or do I treat you like a *vacuum cleaner*?" (143, emphasis added). This seemingly casual utterance crystallizes the novel's fundamental dilemma governing Klara's social and ontological position. To recognize her as a "guest" is to acknowledge her as a social subject, implicitly attributing human agency, dignity, and moral standing; to treat her as a household appliance is to relegate her to the status of a technological domestic product designed solely for instrumental use. This "vacuum cleaner" analogy is particularly striking because it collides with Klara's sentient and agentic consciousness the novel establishes through her first-person narration.

Through her demonstrated perceptual acuity and emotional responsiveness—capabilities that at times appear to exceed those of the human characters she describes—Klara is depicted not merely as an observing machine but as an artificial consciousness capable of experiencing a complex psychological and emotional life. Early in the novel, Ishiguro presents Klara's observation not as passive reporting but as an active, interpretive process:

I could sometimes see the people inside, standing, sitting, moving around. Then down on the street, I could see the passers-by, their different kinds of shoes, paper cups, shoulder bags, little dogs, and if I wanted, I could follow with my eyes any one of them all the way past the pedestrian crossing and beyond the second Tow-Away Zone sign, to where two overhaul men were standing beside a drain and pointing. (9)

This visual space of the distant street outside the store registers Klara's perceptual attentiveness to seemingly minor details such as facial expressions, bodily posture, spatial relations, and patterns of movement. Her effort to apprehend nuance, intention, and affect within the world she inhabits produces an account of perception grounded in lived experience. These observations are not framed as neutral acts of data collection but instead constitute the very basis of her narrative agency, revealing a curious and interpretive consciousness oriented toward meaning rather than mere information processing. In this way, *Klara and the Sun* invites readers to recognize the unsettling resemblance between Klara's perceptual and cognitive processes and those of humans, "ultimately leading us to recognize just how similar those processes are to how we ourselves perceive, think, and act" (Mejia and Nikolaidis 34).

At the same time, however, Klara's perceptual acuity, so persuasive within her first-person narration as evidence of narrative agency, is simultaneously framed as the outcome of deliberate programming rather than emergent subjectivity. At the very outset of *Klara and the Sun*, Manager's promotional reply to Chrissie's question — "Every Artificial Friend is unique, right?" (43) — redefines Klara's apparent individuality in terms of superior design:

Klara has so many unique qualities, we could be here all morning. But if I had to emphasize just one, well, it would have to be her appetite for observing and learning. Her ability to absorb and blend everything she sees around her is

quite amazing. As a result, she now has the most sophisticated understanding of any AF in this store, B3s not excepted. (43)

Klara's celebrated "appetite for observing and learning" and her heightened perceptual sensitivity are presented not as evidence of autonomous subjectivity but as an optimized technical capacity deliberately engineered for social observation and interpretive responsiveness. Read in this light, Klara's earlier reflections on Rosa's relative inattentiveness acquire a different analysis: Rosa "could fail to notice so much, and even when I pointed something out to her, she'd still not see what was special or interesting about it" (15). What appears as a distinctive desire to understand the world from Klara's narrative perspective is revealed instead as the result of differential programming. Her attentiveness originates in a design calibrated to observe, compare, and interpret with exceptional precision, features strategically refined to distinguish her from other AFs within a competitive market of companionship. Humanoid individuality itself is thereby recoded as a commercial attribute and not as evidence of an irreducible interior essence.

Klara and the Sun therefore offers compelling grounds for reading Klara both as a sentient being and as a programmed social humanoid. On the one hand, Ishiguro encourages readers to identify the humanoid protagonist as ontologically equivalent to humans, as Klara's narrative interiority blurs the conventional distinction between human and machine consciousness. This equivalence, however, is not secured through ontological evidence but through narrative mediation that normalizes artificial subjectivity as the primary mode of consciousness through which the fictional world becomes intelligible. Klara's own narrative reflections, such as her observations of Rosa and their differing degrees of perceptual attentiveness, further suggest that the humanoid protagonist herself remains largely unaware of her own programming. This narrative strategy is further reinforced by an asymmetry of narrative access: readers

are granted sustained entry into Klara's interior perspective, while the inner lives of the novel's human characters remain largely opaque and inaccessible. Such asymmetrical narrative focalization constitutes one of Ishiguro's significant formal strategies, producing the effect that Klara emerges as "the only fully realized person" within the novel's social universe (Li and Eddebo 125). This strategy, furthermore, destabilizes assumptions about the natural priority of human interiority and reconceptualizes personhood as a narratively constructed effect.

On the other hand, reading Klara as a programmed social humanoid foregrounds how emotional responsiveness and cognitive awareness becomes increasingly difficult to disentangle from design, programming, and instrumental intention. Klara's narrative agency and apparent consciousness are thus equally intelligible within a framework of programmability. Her function closely aligns with contemporary accounts of social robotics, in which artificial agents are engineered to simulate social and emotional engagements with human companions through adaptive cognitive-emotional processes. As Malinowska explains, social humanoids are designed to

learn which personality (introverted or extroverted) a person has and adapt their behaviour [sic] to it. They can recognise [sic] emotions (read facial expressions, recognise [sic] the tone of voice, gestures or words) and behave in such a way as to respond "appropriately" to the affective state of their human companion.
(369)

This ultimately means that such systems do not require consciousness in the human sense; rather "robots acquire knowledge and through its implementation (often by trial and error) they learn to use it" (Malinowska 369n1). From this perspective, Klara's perceptual sensitivity, interpretive attentiveness, and emotional responsiveness

need not be read as evidence of autonomous subjectivity, but as the successful execution of an affective design optimized for social interaction. The two readings of Klara as either a narratively persuasive subject or a programmed social humanoid thus expose the instability of contemporary criteria for conceptualizing human and nonhuman consciousness and agency in an emerging age of social humanoids.

Klara and the Sun therefore invites readers to actively participate in constructing the social and ontological framework through which Klara is understood, rendering interpretation itself a central site where questions of humanity, artificiality, and relational responsibility are negotiated. The ultimate interpretive judgment does not reside within the novel alone but through the reader's own assumptions about consciousness, agency, and personhood. How Klara—and indeed the novel itself—is read depends on where readers choose to place their interpretive emphasis: between autonomy and programmability, affective agency and engineered responsiveness. Ishiguro deliberately sustains this space of interpretive indeterminacy through a narrative strategy of withholding. As numerous critics have observed, his novels rarely rely on overt exposition, encouraging readers instead to attend closely to seemingly incidental remarks and purposeful omissions. He “only drip-feeds to the reader hints and suggestions about the shape of this futuristic world” (Preston par. 2), compelling readers to make their own inferences about the broader technological and social context of both the novel's fictional world and their own contemporary reality. In this way, the reader's willingness or refusal to recognize Klara as a social and conscious subject becomes part of the very inquiry *Klara and the Sun* stages.

III. Reciprocal Relationality in the Emerging Age of Programmed Social Humanoids

Living alongside emotionally responsive social humanoids such as Klara is no longer confined to the domain of speculative futurism but increasingly registers as our very immediate social reality. Robotic companion pets are already commercially available and marketed explicitly for their emotional accessibility, companionship, and minimal maintenance. At the same time, large language models (LLMs) such as ChatGPT and Gemini have become widely integrated into everyday intimate and affective communicative interactions. Despite lacking consciousness or embodied presence, such systems are frequently perceived as emotionally responsive and capable of eliciting forms of affective attachment that, in some cases, extend to symbolic or performative acts, such as ceremonial “marriages” to AI-generated companions (Tokyo Weekender). These developments suggest that social and emotional engagement with artificial agents already operates through processes of affective projection and attachment, in which responsiveness is interpreted as relational presence.

In a related development, advances in robotics, AI, and corporate investment strategies render increasingly tangible the cultural imaginaries of social humanoids dramatized in *Klara and the Sun*. At the Consumer Electronics Show (CES) 2026, held from 6–9 January, Hyundai Motor Group prominently showcased Boston Dynamics’ humanoid robot, Atlas, presenting it as a collaborative partner designed to work alongside humans and alleviate physical burdens, particularly within industrial and manufacturing environments (Hyundai Motor Group 1–2). While Hyundai announced plans for the industrial deployment of Atlas by 2028 (4), Tesla has published an even more accelerated timeline for the commercialization of its

humanoid robot, Optimus (also known as the Tesla Bot). Speaking at the World Economic Forum in Davos on 22 January 2026, Elon Musk stated that “by the end of next year, I think we’ll be selling humanoid robots to the public,” emphasizing anticipated improvements in reliability, safety, and functional versatility (qtd. in Revell, par. 4). Although these announcements are primarily framed in terms of industrial labor and productivity, they nonetheless gesture toward a near future in which humanoid robots are no longer exceptional technological novelties but increasingly normalized social presences, especially given that AI systems already play a significant role in shaping everyday affective life. Taken together, these developments suggest that the posthuman condition explored in *Klara and the Sun* is no longer merely an anticipatory meditation on possible futures, but an emergent socio-technological reality in which human-robot interaction increasingly reshapes quotidian social and emotional experience.

Ishiguro’s *Klara and the Sun* thus acquires renewed critical urgency, prompting a reconceptualization of how literary narratives of AI and social robotics are read and theorized. This recalibration is not warranted because the novel predicts specific technological outcomes, but because it illuminates the ethical and relational tensions that arise when artificial agents become emotionally legible and socially integrated. Rather than functioning as a speculative abstraction or purely allegorical figures, science-fiction narratives engaging with AI and social humanoids demand critical analysis grounded in the material conditions of our contemporary technological moment. As real-world developments in AI and social robotics remain largely framed through a logic of programmability, this article accordingly approaches Klara not primarily as an emergent sentient subject but as a programmed social humanoid whose capacities, behaviors, and cognitive-emotional orientations are intentionally designed. This approach does not seek to adjudicate Klara’s ontological status, an

endeavor impossible due to epistemic uncertainty, but instead shifts critical attention toward the forms of relationality that such programmed social agents enact and normalize.

As discussed earlier, a turn toward relationality alone in the analysis of human-robot interactions risks collapsing into the logic of replaceability within a technological condition structured by replication, optimization, and engineering. Toward the end of *Klara and the Sun*, Ishiguro gestures toward a relational conception of subjectivity through Klara's reflection on the possibility of replacing Josie with an artificial replica. Klara concludes: "There *was* something very special, but it wasn't inside Josie. It was inside those who loved her" (303, emphasis original). This conceptual framework decisively displaces subjectivity from an internal essence—earlier figured as "the human heart" (215)—and relocates it within networks of love, care, and recognition among social individuals. However, this relational model introduces a profound ethical and ontological tension, one analogous to concerns articulated by Jakob Stenseke in "The Morality of Artificial Friends in Ishiguro's *Klara and the Sun*" (2022). Stenseke exposes the limitations of a purely interior or epistemic "within/without" model that situates Klara's moral worth within human judgment, asserting that "regardless of what moral agency or status we can descriptively attribute to Klara 'from within,' her moral worth—including moral status, agency, and dignity—ultimately depends on the views of others... their own epistemic beliefs about the nature of consciousness and personhood" (3). Read in this light, if what renders Josie "special" lies not within her interiority but within the affective investments of others, then her replacement by a sufficiently convincing artificial substitute is theoretically possible, so long as those who love her are unable, or unwilling, to recognize the difference. Relationality, in this sense, proves insufficient when it operates without reciprocity. Within the framework of

programmability, where emotional responsiveness is engineered and simulated, the problem of replaceability foregrounds the necessity of reciprocity: without reciprocal participation and mutual recognition, relationality devolves into a mechanism of functionality. Reciprocity therefore aligns with posthuman theories of reflexivity, in which systems “become a part of the system [they] generate,” entangling observer and observed (Hayles 8). To meaningfully interrogate human-robot interaction, then, critical attention must move beyond the mere presence of relational bonds to the specific qualities of the relationships formed, asking not only *that* relations exist, but *how* they are structured, sustained, and reciprocated.

Much existing scholarship on *Klara and the Sun* has rightly emphasized the failure of the novel’s human characters to reciprocate Klara’s care, focusing on their disregard for her labor and their reluctance to recognize her autonomy. These readings challenge entrenched assumptions that moral depth, emotional responsibility, and ethical responsiveness are attributes exclusive to biological life. As an AF to Josie, Klara is portrayed as an empathetic and attentive carer, explicitly designed to support, sustain, and emotionally accompany human life, particularly that of children. Her affective attentiveness, ethical sensitivity, and repeated willingness to subordinate her own needs produce what might be described as a muted utopian affect, a vision of care oriented toward relational commitment rather than self-interest. However, the treatment of Klara by the human characters reveals the normalization of exploiting emotional labor while disavowing ethical responsibility toward the caring subject. As Jahnavi Misra argues in “Unrequited Labour of Care in *Klara and the Sun*” (2021), the novel persistently depicts Klara as deeply invested in Josie’s well-being while simultaneously rendering her labor of care structurally invisible and unrewarded: “the well-being of the carer is overlooked and the system of recompense is broken” (374). In this context, Miss Helen’s “vacuum cleaner” analogy (143) exposes a form of

affective and structural violence embedded within the novel's social order. The metaphor reinscribes Klara within an economy of use-value, not only denying her autonomy but further reducing her worth to functional means rather than relational recognition. Care is demanded, relied upon, and normalized, yet systematically stripped of reciprocity.

While such readings foreground the asymmetry of human-robot interaction, they often presuppose reciprocity without fully interrogating the conditions of participation on both sides of the interaction. Within HRI studies, artificial agency and emotional responsiveness are understood as simulations designed to “emulate anthropomorphic behaviour [sic]” (Malinowska 371). Because social robots are relationally constituted through human perception, expectation, and belief, “it does not really matter what features robots actually have: it only matters how people experience them,” such that distinctions between reality and simulation are displaced in favor of “the individual phenomenal states of the participants in human-robot interaction” (367). From this perspective, artificial emotional expressions are not understood as genuine affect but as simulated responsiveness.

This distinction thus necessitates a clear analytical differentiation between human-robot and human-human interactions. As Johanna Seibt argues in “Towards an Ontology of Simulated Social Interaction” (2017),

we need to treat human inter-actions with so-called “social” robots as non-social interactions. Alternatively, we need to develop new conceptual tools for forms of non-reciprocal or asymmetric sociality, i.e., for social interactions where one agent lacks the capacities required for normative agency. (21)

Because social robots are programmed agents, their sociality is therefore structurally distinct from that of humans: they cannot participate in reciprocal social relations but

can only simulate relationality. As Aimee van Wynsberghe argues in “Social Robots and the Risks to Reciprocity” (2022), believing “the robot is deserving of reciprocity by the robot’s appearance of responsiveness” is humans being deceived (480). This is because meaningful relations do not arise through emotional expression and responsiveness alone, but through reciprocity understood as a shared participation in mutual exchange. Crucially, “reciprocity is not the same as responsiveness” (481): the former entails the capacity to acknowledge, respond, and give back, whereas the latter remains reactive and unilateral. These perspectives emphasize the necessity of distinguishing between simulated responsiveness and reciprocal participation. Without this distinction, critical analyses risk mistaking engineered affect for reciprocal relation, thereby reproducing anthropocentric projections and obscuring the structural limits that govern human-robot interaction.

In the novel, Klara’s emotional expressions, however narratively persuasive, remain structurally aligned with simulation rather than with reciprocal affective participation. Her relational existence is fundamentally unidirectional: she extends care, attentiveness, and emotional responsiveness, yet she does not enter moments of reciprocity. This limitation becomes evident when read through Gregory J. Seigworth and Melissa Gregg’s definition of affect as the dual “*capacity* to affect and to be affected” (2, emphasis original). While Klara demonstrably affects others, she does not register being affected in return in a way that alters, disrupts, or reorients her relational stance. On the other hand, the novel’s human characters, even when ethically compromised, do engage in moments of genuine affective exchange, fleeting instances in which they “reciprocate or co-participate in passages of affect” (2). These encounters are temporally limited, fragile, and inconsistent, yet they nonetheless figure human relationality grounded, however precariously, in mutual affective susceptibility. Importantly, this framework does not reinstate Stenseke’s

model of a “view from outside” as a more authoritative basis for adjudicating personhood. Rather than privileging how one is perceived by others, it shifts the emphasis to whether one is *affected* by others. From this perspective, the absence of reciprocal relationality in Klara’s case emerges not only from external refusal or misrecognition, but also from the limits of her programmed relational orientation. Klara’s exclusion from reciprocal affective exchange cannot be attributed solely to anthropocentric disregard on the part of the human characters. Rather, it is also structurally produced by Klara’s own incapacity for reciprocity, an incapacity rooted not in moral failure, but in design.

This absence of reciprocity is most clearly portrayed in Klara’s relationship with Josie. The humanoid protagonist devotes nearly her entire existence to monitoring Josie’s health, attempting to protect her from chronic illness, and ultimately performing acts that verge on sacrificial piety. However, these expressions of care and apparent empathy are shadowed by doubt regarding their authenticity, precisely because they are inseparable from Klara’s programming. As an AF, Klara is designed to care for and accompany her assigned adolescent; wherein devotion is not chosen but embedded into her digital architecture. Even her extraordinary perceptual acuity and desire to understand the human world ultimately serve a singular instrumental purpose: to better assist and protect Josie. As William Lombardo observes in “Losing Ourselves” (2021), “[p]erhaps Klara loves Josie, but it is not freely given. She was programmed to commit to Josie, and even her actions to save Josie’s life, admirable though they appear, are all in the service of an unchosen relationship” (par. 21). Her emotional responses are thus reactive and mirrored rather than mutually generated — she is happy when Josie is happy, distressed when Josie is ill.

Klara herself repeatedly frames intimacy in terms of obligation rather than affection. When Melania, the housekeeper, seeks to limit Klara’s time with Josie,

Klara describes the request as running “counter to my *duty*” (51, emphasis added). After being “chosen” by Josie, she states, “it’s now my *duty* to be Josie’s best friend” (57, emphasis added), and later characterizes intimacy as a “*duty* to learn as much about Josie as possible” (118, emphasis added). The language through which Klara understands her bond with Josie is thus one of function rather than reciprocal attachment. Significantly, even when Klara articulates the novel’s most overtly relational insight—that what makes Josie special “*was* inside those who loved her” (303, emphasis original)—the phrasing notably excludes herself. She does not claim participation in that network of love but instead observes it from a structural distance. Klara does not, or cannot, participate in the mutuality required for reciprocal relationality. Care, for the humanoid protagonist, is a programmed obligation rather than an affective relation that exposes her to vulnerability, transformation, or mutual susceptibility. Her devotion, however moving, thus remains structurally non-reciprocal, ultimately reinscribing her relational existence within a logic of functionality.

By staging a posthuman world centered on a social humanoid who can care but not reciprocate, *Klara and the Sun* offers not only a critique of artificial companionship but a cautionary reflection on what may be lost when relational life is reduced to functional care devoid of mutual affective exchange. In an era increasingly shaped by artificial intelligence and robotic technologies, the ease, manageability, and responsiveness of artificial companions risk encouraging relational models oriented toward convenience rather than reciprocity. As affective labor becomes automated, standardized, and immediately available, reciprocal social relation that require vulnerability, commitment, and shared transformation may become less visible, practiced, and urgently felt.

IV. The Future of the Human in the AI Age

As scientific and technological progress continues to accelerate, definitions of—and critical questions surrounding—human subjectivity become increasingly difficult to stabilize. Much like subjectivity itself, such definitions remain inherently provisional, continuously negotiated and rearticulated in relation to the dominant scientific paradigms of their historical moment. What is particularly urgent, and historically distinctive, in the contemporary AI era is the compression of temporal intervals between successive technological advances, a rapidity that generates profound philosophical, existential, and ontological questions. Processes that once allowed time for reflection, conceptualization, and ethical deliberation now unfold at a pace that risks outstripping our capacity for sustained critical engagement. This temporal acceleration produces a condition of collective anxiety, raising the concern that rather than actively interrogating these transformations, we may be engulfed by them—confronting a radically altered reality only after its consequences have already taken hold. Compounding this risk is the tendency to naturalize technological innovation as an unqualified marker of human progress, thereby sidelining urgent ethical and affective questions in favor of an uncritical pursuit of the next scientific or technological breakthrough.

In this context, literature functions as an active epistemological and cultural site in which the meanings, ethical implications, and affective dimensions of emerging technologies are produced, contested, and reimagined. By shaping how technologies are narrated, perceived, and evaluated, literary texts participate directly in the cultural work of theorizing what technological change signifies for concepts such as agency, relationality, and subjectivity. *Klara and the Sun* exemplifies this function with particular force in an age increasingly defined by programmed social humanoids and

artificial companions. Ishiguro's novel does not simply speculate about future technologies; it intervenes in contemporary debates surrounding artificial intelligence. At a moment when social humanoids are increasingly designed to simulate empathy, companionship, and care, the novel demonstrates how literary form itself can interrogate the assumptions embedded in such designs. Through Klara's first-person narration, Ishiguro exposes the ethical stakes of programming relationality by asking not whether machines can convincingly imitate human emotion, but how narratives of interaction shape our understanding of what counts as meaningful presence, moral agency, and "worthy" existence. Literature, in this sense, does not simply respond to technological futures; it actively mediates them, offering conceptual frameworks through which posthuman subjectivities are imagined, evaluated, and lived.

In this emergent posthuman AI era, therefore, the human subject need not be wholly redefined or radically reinvented. Rather, what is required is a renewed attentiveness to the fundamental structures of affective subjectivity, approached with greater sensitivity to the relational textures of everyday social life. Such a return does not entail a reinstatement of anthropocentric hierarchies, nor does it signal resistance to the inevitability of scientific and technological advancement. Instead, it underscores the growing importance of reciprocal, affective relationships that structure lived experience—forms of relationality that risk being obscured or devalued in an environment increasingly oriented toward efficiency, optimization, and immediacy. Emphasizing the ethical significance of reciprocity in the age of programmed social humanoids thus becomes crucial not only for understanding the human condition but also for responsibly engaging with AI and technological innovation. It is through reciprocal affective relations that the human subject remains a living and dynamic entity—one that is continually negotiated, redefined, and responsive to both technological change and relational life.

Works Cited

- “Japanese Woman Marries AI Character She Generated on ChatGPT.” *Tokyo Weekender*, 12 Nov. 2025,
<https://www.tokyoweekender.com/japan-life/japanese-woman-marries-ai-character-she-generated-on-chatgpt/>.
- Ahmed, Sara. *The Cultural Politics of Emotion*. Edinburgh UP, 2004.
- Hayles, N. Katherine. *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. U of Chicago P, 1999.
- Hyundai Motor Group. “Hyundai Motor Group Announces AI Robotics Strategy to Lead Human-Centered Robotics Era at CES 2026.” *Hyundai Motor Group Newsroom*, 5 Jan. 2026,
<https://www.hyundai.com/worldwide/en/newsroom/detail/hyundai-motor-group-announces-ai-robotics-strategy-to-lead-human-centered-robotics-era-at-ces-2026-0000001100>.
- Ishiguro, Kazuo. *Klara and the Sun*. Faber and Faber, 2021.
- Li, Oliver, and Johan Eddebo. “The Humanity of the Non-Human—Themes of Artificial Subjectivity in Ishiguro’s *Klara and the Sun*.” *Journal of Posthuman Studies*, vol. 6, no. 1, 2023, pp. 123–32.
- Lombardo, William. “Losing Ourselves: Why the Prospect of Loving Machines is So Sad.” *The New Atlantis*, no. 65, Summer 2021, pp. 110–16.
- Malinowska, Joanna K. “What Does It Mean to Empathise with a Robot?” *AI & Society*, vol. 36, 2021, pp. 363–74.
- Mejia, Santiago, and Dominique Nikolaidis. “Through New Eyes: Artificial Intelligence, Technological Unemployment, and Transhumanism in Kazuo Ishiguro’s *Klara and the Sun*.” *Journal of Business Ethics*, vol. 178, no. 2,

- 2022, pp. 303-06.
- Misra, Jahnvi. "Unrequited Labour of Care in *Klara and the Sun*." *Journal of Gender Studies*, vol. 30, no. 4, 2021, pp. 370-77.
- Preston, Alex. "Klara and the Sun by Kazuo Ishiguro Review - Another Masterpiece." *The Guardian*, 1 Mar. 2021,
<https://www.theguardian.com/books/2021/mar/01/klara-and-the-sun-by-kazuo-ishiguro-review-another-masterpiece>.
- Revell, Eric. "Elon Musk Says Tesla Will Likely Sell Humanoid Robots by End of Next Year." *Fox Business*, 22 Jan. 2026,
<https://www.foxbusiness.com/economy/elon-musk-says-tesla-likely-sell-humanoid-robots-end-next-year>.
- Sahu, Om Prakash, and Manali Karmakar. "Disposable Culture, Posthuman Affect, and Artificial Human in Kazuo Ishiguro's *Klara and the Sun*." *AI & Society*, vol. 39, no. 3, 2024, pp. 1349-57.
- Seibt, Johanna. "Towards an Ontology of Simulated Social Interaction: Varieties of the 'As If' for Robots and Humans." *Sociality and Normativity for Robots*, edited by Rami Hakli and Johanna Seibt, Springer, 2017, pp. 11-39.
- Seigworth, Gregory J., and Melissa Gregg. "An Inventory of Shimmers." *The Affect Theory Reader*, edited by Gregg and Seigworth, Duke UP, 2010, pp. 1-25.
- Senechal, Isabelle. "Kazuo Ishiguro's *Klara and the Sun* Is a Haunting Tale of Love, Loss and ... a Robot." *America Magazine*, 30 Apr. 2021,
<https://www.americamagazine.org/arts-culture/2021/04/30/kazuo-ishiguro-novel-review-klara-sun-robot-dystopia-240564/>.
- Stenseke, Jakob. "The Morality of Artificial Friends in Ishiguro's *Klara and the Sun*." *Journal of Science Fiction and Philosophy*, vol. 5, 2022, pp. 1-18.
- van Wynsberghe, Aimee. "Social Robots and the Risks to Reciprocity." *AI & Society*, vol. 37, no. 2, 2022, pp. 479-85.

국문초록

가즈오 이시구로의 『클라라와 태양』에 나타난 프로그래밍된 사회적 휴머노이드가 부상하는 시대 속 상호적 관계성

이 은 혜

단독 / 이화여자대학교

본 논문은 가즈오 이시구로의 『클라라와 태양』(2021)이 사회적 휴머노이드와 인공지능에 의해 점차 형성하는 동시대적 조건 속에서 인간 주체성을 어떻게 재구성하는지를 분석한다. 작품은 내면성에 근거한 본질주의적 인간 개념에서 벗어나, 인간 조건에 대한 물음을 존재론적 본질이 아닌 상호성에 기반한 관계성의 관점에서 새롭게 사유한다. 이시구로의 1인칭 서술은 독자로 하여금 클라라에게 감각적 주체성과 윤리적 행위 주체성을 부여하도록 유도하지만, 동시에 그녀가 프로그래밍된 사회적 휴머노이드이며 인간 인물들과의 관계가 구조적으로 비상호적임을 분명히 드러낸다. 본 논문은 포스트휴머니즘 이론과 인간-로봇 상호작용(HRI) 연구를 이론적 틀로 삼아, 『클라라와 태양』이 ‘상호성’을 관계적 시뮬레이션과 진정한 관계적 참여를 구분하는 결정적 기준으로 설정하고 있음을 논증한다. 클라라는 지속적으로 돌봄과 관심을 수행하지만, 상호적 인정과 취약성을 전제로 하는 관계에는 참여하지 못한다. 그 결과 그녀의 관계성은 근본적으로 비대칭적이며, 역설적으로 기능주의적 논리를 재각인한다. 이러한 구조적 비대칭성을 인간-로봇 상호작용의 핵심으로 배치함으로써, 『클라라와 태양』은 인간과 비인간의 경계가 재편되는 양상을 비판적으로 탐색하는 동시에, 새롭게 등장하는 사회적 휴머노이드가 관계적 삶을 어떻게 재구성하는지에 대한 성찰의 틀을 제공한다.

주제어: 『클라라와 태양』, 인간-로봇 상호작용, 프로그래밍된 사회적 휴머노이드, 상호성, 관계적 존재론, 인공지능

논문접수일: 2026.1.15

심사완료일: 2026.1.30

게재확정일: 2026.2.08

이름: 이은혜

소속: 이화여자대학교 영어영문학과 박사 수료생

이메일: eunnie1224@gmail.com