The Morality of Artificial Friends in Ishiguro's *Klara and the Sun*

Jakob Stenseke

Lund University, Sweden

Abstract

Can artificial entities be worthy of moral considerations? Can they be artificial moral agents (AMAs), capable of telling the difference between good and evil? In this essay, I explore both questions—i.e., whether and to what extent artificial entities can have a moral status ("the machine question") and moral agency ("the AMA question")—in light of Kazuo Ishiguro's 2021 novel *Klara and the Sun*. I do so by juxtaposing two prominent approaches to machine morality that are central to the novel: the (1) view "from within," including the standard (or "metaphysical") perspective on moral agency, and the (2) view "from outside," which includes behaviorism, functionalism and the social-relational perspective. Importantly, while the story illustrates both views, it exposes the epistemological vulnerability of the first in relation to the practical and social reality imposed by the second. That is, regardless of what metaphysical properties the Artificial Friend Klara can be said to have (from within), her moral status as well as agency ultimately depend on the views of *others* (from outside), including the others' own epistemic beliefs about the nature of consciousness and personhood.

1. Introduction

Who should we care about and why? On what grounds do we attribute moral significance or agency to others? A common understanding is that every sentient being has, at some basic level, a *moral status*, which means that we have reason to care about it for its *own* sake (Jaworska & Tannenbaum 2021). Intuitively, if a being has the capacity to feel pain and pleasure, some non-trivial moral considerations about its well-being become relevant. We might dedicate effort to prevent it from feeling pain, or conversely, help it to achieve pleasure. At a minimal level, we do not intentionally cause it to suffer, or at least not without good reason. However, having a moral status does not make you a *moral agent* per se,[1] that is, a being or entity with properties or abilities that makes it able to *be* moral, e.g., to act in reference to what is "right" and "wrong" (Navari 2003). At the center of moral agency, if we follow the moral philosophy of Immanuel Kant, is the concept of *autonomy*, as it allows a rational agent to act according to their own self-imposed rules (in other words, be *self-legislative*), without the influence of factors that are external to herself (Kant 1785). On the other hand, if we follow the empiricism of

Journal of Science Fiction and Philosophy Vol.5 (2022)
© 2018 by Alfredo Mac Laughlin. ISSN 2573-881X
This article © 2022 by Jakob Stenseke. Published 2022 June 15.
Original publication under a CC-BY-NC License by the *Journal of Science Fiction and Philosophy*.

Hume, we find that reason cannot by itself guide the autonomous will, but is rather instrumentally the "slave of passions" (Hume 1740). Regardless of Kantian rationalism or the instrumentalism of Hume, moral agency resides in the foundation of our everyday moral practices, e.g., about whether someone is responsible, held accountable, worthy of blame, or allowed certain legal rights. It permeates and shapes our political and social reality: if we can decide *who* is worthy of our moral concern—of empathy, rights, dignity, and respect—we can also determine *who is not*. To that end, alienation or the dehumanization of "others" have served to justify and legitimize abuse and hierarchies of dominance, of imperialism and colonialism, slavery and racism, and the ongoing industrial-scale mass-slaughter of sentient non-human animals. Are artificial beings next in line?

Literature and philosophy are excellent venues to explore the possible, the yet to be, and how it might change and transform the human condition. After all, questions about morality and its intricate relationship to mind and behavior have for a long time been as prevalent in fiction as they have been in moral philosophy. In recent years, as intelligent and autonomous artificial systems enter and transform a growing number of human domains, the prospect of machine morality has once again become center stage in science-fiction as well as in academic discourse. Even if artificial entities were equipped with human-like cognition—emotions, desires, and the capacity for subjective experience—would they still be viewed as mere non-sentient tools for human ends, or might they at some point become something else?

Kazuo Ishiguro's novel *Klara and the Sun* (2021) describes a compelling vision of a possible, not-so-far-away future where Artificial Friends (AFs) are employed to alleviate the cold loneliness of modern life for children and teenagers. While the story explores a range of ethical issues related to emerging technology—including artificial slavery, genetic editing, societal impacts of automation—I will focus on its contribution to two distinct yet related debates on the morality of artificial entities: (i) whether and to what extent artificial entities can have a *moral status*, i.e., be worthy of moral consideration ("the machine question"), and (ii) whether and to what extent artificial entities can be *artificial moral agents* (AMAs), capable of behaving in reference to what is morally "good" and "bad" ("the AMA question"). I do so by juxtaposing two major approaches to machine morality that are central to *Klara and the Sun*: the view "from within" and the view "from outside." The first view attributes moral status (or moral patienthood) and agency in terms of necessary metaphysical properties or cognitive capabilities an agent can be said to possess; in particular, phenomenal consciousness (i.e., subjective experience), autonomy, and rationality. The second view follows the behavioristic paradigm, and attributes status and agency based on an agent's outward behavior (Danaher 2020). The view "from outside" is subsequently elaborated through the perspectives of functionalism and social-relationalism; the former centers on the functional role of moral behavior and cognition (Floridi & Sanders 2004), the latter emphasizes the social and relational aspect of moral practices (Coeckelbergh 2010b; Gunkel 2018).

In the rest of the essay, Ishiguro's novel will act as the vehicle for driving the philosophical debate on moral status and agency in the context of artificial intelligence. First, the view "from within" is examined through the inner narrative lens of AF Klara, the humanoid robot who is appointed to care for the 14-year-old Josie. The inner view is then contrasted with the view "from outside," illustrated by how the other characters of

the novel treat and view Klara based on her outward behavior (behaviorism), her role (functionalism), and the relationships they develop to her (social-relationalism). While *Klara and the Sun* ingeniously explores strengths and weaknesses of the different views, I will argue that it more importantly exposes the epistemological frailty of the view "from within," especially in contrast to the pragmatic reality imposed by the view "from outside." That is, 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. The novel's innovative upshot is that this not only involves aspects of morality that others attribute to Klara based on her behavior and function, but also their own epistemic beliefs about the nature of consciousness and personhood.

2. Artificial morality from within

Klara and the Sun is set in a possible near future where machines have replaced classes of former "elite workers," and affluent children are genetically edited ("lifted") to improve their chances for success. At the center of the story is a new kind of being: Artificial Friends (AFs), humanoid robots bought by parents to be social companions for their teenagers. Each AF has its own unique personality so as to accommodate the various needs of the teenage child it is tasked to accompany. As a result, the roles of AFs are often far from clear, and vary greatly from child to child. While some AFs are perceived as convenient tools, cool toys, or some special kind of artificial pets, others take on more dignified and emotionally profound roles that include the sort of reciprocal responsibilities and duties indicative of supportive human friendships. AF Klara, our narrator, belongs to the latter category. Klara stands out among her fellow AFs due to her excellent observational capacities. In a sales pitch given to a potential customer, the Manager of the AF store says:

> 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. (Ishiguro 2021, 42)

From the passage, it is apparent that Klara—an AF of the B2 series—does not merely record everything like a simple camcorder, but she is also equipped with sophisticated reasoning capabilities that allow her to interpret and understand a complicated reality; on par or even surpassing a later series of AFs (the B3). As Klara curiously explores various phenomena in her increasingly expanding outer and inner world, she is able to make inferences based on observation (induction), reflect upon and produce new inferences based on what she already knows (deduction), and creatively come up with hypothetical and probable explanations even when essential information is missing (abduction, or "inference to the best explanation"). Far from a tabula rasa, Klara seems to possess a sufficiently rich vocabulary of the relevant semantic mappings between words and phenomena that are necessary to make sense of her experience through sentences. Importantly—as one might expect from an AF—her outstanding observational skills seem particularly tailored to grasp the intricate subtleties of human social interaction. For better or worse, she reads more into events and social exchanges than is apparent from the outside. Like the most sensitive humans, Klara is constantly

scrutinizing the behavioral subtext of vocal tone, eye gaze, and body language, endlessly theorizing about the intentions, beliefs, and desires of others.

Essentially, Klara has most, if not all, of the capacities that are both narrowly and broadly associated with the umbrella term "rationality." She has reasons for her beliefs, decisions, and actions. She knows what to do in order to achieve a certain end, a sophisticated so-called "practical" or "instrumental" rationality (Walliser 1989). Her ability to navigate and balance the present with complicated chains of hypothetical future events allows her to effectively work towards long-term goals. Her capacity to ascribe and understand mental states of others makes her an apt mind-reader, a capacity often called "theory of mind" or "intentional stance" (Dennett 1987). Of course, this does not make her immune from making mistakes. Like all beings situated in complex and dynamic environments, her actions do not always achieve the intended outcome. Like all fallible mortals, she occasionally draws erroneous conclusions based on imperfect information (for instance, based on the fact that AFs are powered by sunlight, Klara develops a spiritual attachment to the Sun, which she comes to believe has certain special powers).

More important for the purpose of this essay is that Klara exhibits the sort of rationality that is necessary—although not by itself sufficient—for moral agency. Beyond the mere practical rationality of effectively pursuing a certain end (seen as essential for moral agency by authors such as Kolodny and Brunero [2020] and Johnson [2006]), Klara has the sort of moral and empathetic sensibilities needed to exercise moral judgement.[2] The breadth of Klara's moral competence can be illustrated by the fact that she is able to follow the prescriptions of the three major normative frameworks: consequentialism, deontology, and virtue ethics. In reference to the first, Klara considers the consequences of her actions and how those consequences relate to the values and preferences she seeks to maximize (e.g., Josie's well-being). It is also clear that Klara is able to adhere to general and particular moral rules and follow the commands of her sovereigns (deontology). In line with virtue ethics, Klara continuously learns from her experiences, and fosters the traits and dispositions of her moral character that enable her to be an excellent AF (following the virtue-theoretic emphasis on *being* rather than *doing*).

But regardless of her astute rational capacities, would we consider Klara as a moral agent if she acted without any regard of her own preferences or free will? This brings us to the next key aspect of moral agency, namely *autonomy*. In the AMA debate, authors such as Friedman and Kahn Jr. (1992), Hellström (2013), and Himma (2009) have argued that autonomy is a necessary prerequisite for moral agency.

Is Klara autonomous? Throughout the novel, we are given evidence that could support the view that Klara is autonomous in a profoundly human sense, but also evidence that indicates the opposite. At the beginning of the story, Klara's entire existence is constrained to the interiors of a store populated with other AFs. She is eager to fulfill her ultimate purpose: to be a good friend to whatever child she is chosen to accompany. After some time, Klara encounters Josie, a fragile teenager who suffers from an unspecified but life-threatening illness. They immediately connect, and after some time, also Josie's mother becomes convinced that Klara would be a suitable companion for Josie. But before Klara is purchased from the AF store, Josie dignifies Klara's autonomy with the possibility of saying no, making it clear that she only wants Klara to join her on the premise that Klara *really* wants to:

> 'You will come, right? If Mom says it's okay and everything?' I nodded encouragingly. But the uncertainty remained on her face. 'Because I don't want you coming against your will. That wouldn't be fair. I really want you to come, but if you said, Josie, I don't want to, then I'd say to Mom, okay, we can't have her, no way. But you want to come, right?' (Ishiguro 2021, 23)

There is no doubt what Klara wants. Although she seems open to the possibility of being selected by any child—not considering her own agency in the matter—she is secretly hoping for Josie to return after their initial encounters, worrying that someone else might pick her before that occurs, or even worse: that she is not picked at all. It appears as if the agreement of their companionship is, at some profound human level, reciprocal. But most AFs have no saying at all in the matter, as the Manager says: "It's for the customer to choose the AF, never the other way round" (Ibid, 32).

However, one might ask to what extent Klara and other AFs *wish* or *desire* things based on their own free will, or whether they are merely pre-programmed to do so. After all, it is hard to see how AFs could be commercially or practically viable if they were disobedient, both in terms of internal desires and outward behavior. It would be rather contrary to the purpose of artificial friends if they, even on occasion, did not agree—or did not wish—to enter into companionships with teenagers. It is likely that obedience is, at some rudimentary level, hardcoded in every AF on safety grounds, e.g., to prevent AFs from turning against their owners. At the same time, their compliance in combination with their human-like social cognition brings about an intriguing master-slave dialectic to the heart of all human-AF relationships. If an AF is programmed to accept and care for you, no matter how you treat them, it does not constitute a fair and equal relationship. Ultimately, AFs are only successful as friends insofar as they are perceived as such through the lens of their owners. Depending on their masters, they might be treated as disposable toys or as irreplaceable family-members. In the hands of malice, they could potentially be punching bags for dominance and abuse. Like a sophisticated pet, their traits and behaviors seem to be adaptive to whatever their masters expect them to be. Their autonomy follows the same design: AFs are autonomous within the boundaries of what their masters authorize.

But in the case of AF Klara, it is important to distinguish between two aspects of autonomy: (i) having the *capacity* for autonomy, and (ii) being *granted* autonomy by someone or something (e.g., some individual, a group of people, or a society). The former can be seen as the cognitive prerequisite for the latter (e.g., to be autonomous in terms of one's abilities), whereas the latter is the normative negotiation of in what capacity and in respect to whom the former applies. Following von Gerber (2014), this relational concept of autonomy can be modeled as a relationship between three components, where "one entity (X) is autonomous in relation to another entity (Y) regarding a right or a capacity (Z)" (Ibid, 5). For instance, a self-driving vehicle (X) is only autonomous in terms of having the capacity to drive from point A to B (Z) independent of human control (Y). In the context of international politics, a state (X) is autonomous in relation to other states (Y) regarding their capacity for self-governance (Z). However, human beings possess a vast number of capacities, some of which are generally applicable across a range of domains: moving, using tools, communicating, planning, reasoning, and reflecting. In most contemporary liberal societies, adult humans are more or less free to do and pursue whatever they want, using the means they have, as long as it does not violate the freedom of others or harm them (Mill 1859). Furthermore, even if one were

not granted central liberal tenets such as freedom of speech, one would still retain freedom of thought. When we discuss human autonomy, it is therefore more common to bundle up the various capacities it supposedly involves into more abstract placeholders, such as moral autonomy (ability to deliberate and act according to one's own moral law), personal autonomy (the capacity to independently decide for oneself), and political autonomy (having one's decisions authorized within a political context). Another route, common to debates in bioethics, is to distinguish between "ideal" and "non-ideal" autonomy, the former starting from hypothetical models (e.g., Kantian autonomy) and the latter from actual conditions in the real world (Marceta 2019). But even the non-ideal versions of autonomy are often built upon some underlying notion of general competence. For instance, Beauchamp and Childress' non-ideal version of autonomy rests on the assumption that everyday choices of "generally competent" agents are autonomous (Beauchamp & Childress 2001). In turn, the actions of such choosers are autonomous if they are performed "(1) intentionally, (2) with understanding, and (3) without controlling influences that determine their action" (Ibid, 104).

So, what can then be said about the autonomy of AF Klara? Regarding the true depth of Klara's capacity for autonomy, we can only speculate. Nevertheless, while it is obvious that she is denied the sort of dignified and self-legislative autonomy that free citizens in liberal democracies have (with unalienable rights and liberties), she seems to have the cognitive potential for it. Based on our analysis of Klara's rationality, she clearly exhibits "general competence." In accord with the first two criteria of Beauchamp and Childress' non-ideal conception, she acts with intention and understanding. But it is more difficult to assess whether she satisfies the third, i.e., acting without "controlling influences." The difficulty resides in her obedience and long-term determination to the purpose of doing what is best for Josie. Even when Klara independently comes up with and executes elaborate plans, she never diverges from her main purpose. Although Klara exhibits curiosity—a personal desire to observe and understand the world—it is ultimately motivated by her purpose of serving others (i.e., she tries to understand the world in order to better help others).

At the same time, it would not be entirely fair to assess Klara's autonomy only in terms of her non-egoistic purpose. After all, we not only grant autonomy to but also praise humans who are, at some fundamental level, driven by non-egoistic motives. At the core of human altruism is the personal sacrifice: *you* lose or give up something for the benefit of others. But Klara is self-sacrificial by nature, and perhaps due to her programming, seems somewhat unable to even register events as egoic losses. Ostensibly, this makes her incapable of suffering (and in some odd sense, also incapable of genuine altruism). In particular, she does not seem to acknowledge that her own sentience—i.e., the positive or negative salience of her own subjective experience—carries any moral weight in relation to others. Klara never prioritizes her own secondary preferences (where Josie's well-being is understood as the primary preference), never demands them to be heard or respected; they only carry moral weight to the extent they are warranted and dignified by Josie. Klara's favorite activity is to watch the Sun go down, and she cherishes every moment Josie allows her to do so. In one episode, Josie dignifies Klara's wish to watch the sunset even at the risk of causing the Mother to become angry. The event itself seems rather trivial, but it shows that Josie cares about the subjective preferences of Klara, even at the expense of her mother's desires. More importantly, as we follow Klara's internal reflection throughout the novel, her seemingly

endless curiosity, her growing attachment to the Sun, we get a sense that Klara indeed has a mind of her own.

As we dig deeper into the mind of Klara, it becomes clear that her autonomy is not only intimately linked to her rationality (e.g., general competence), but also to her phenomenal subjectivity, the "what it is like" (Nagel 1974). This brings us to the final component of Klara's morality as viewed "from within," namely, her capacity for subjective experience (or consciousness). Many authors in the AMA debate argue that phenomenal consciousness is fundamental to moral agency (Champagne & Tonkens 2015; Coeckelbergh 2010a; Himma 2009; Johansson 2010; Purves et al. 2015; Sparrow 2007). For instance, how can an agent tell what it is morally "good" from "bad" without the conscious experience of positive or negative mental states? At the same time, from a neuroscientific perspective, the question of how consciousness relates to the material remains, to a large extent, as puzzling as it was when Descartes wrote "I think, therefore I am." Part of the puzzle is to define what consciousness is and what it is not. On the one hand, it can broadly refer to phenomenological features of one's "inner world," such as thought, imagination, and introspection. At other times, the concept involves capacities such as wakefulness, experience, and awareness. But, as described in the introduction, consciousness as mere awareness might not carry much moral weight unless it involves positive and negative mental states, such as the experience of pleasure and pain. From our voyage into Klara's mind, it is clear that she possesses a rich inner world. To the extent language can convey the true colors of a being's inner Cartesian soul—and as a testimony of Ishiguro's masterful representation of subjectivity—Klara has "it." But does she experience in such a way that merits moral status or agency?

Pathocentrism (from the Greek *pathos,* meaning "suffering") is the moral view that considers the suffering of beings as morally significant. In the words of Singer: "If a being suffers, there can be no moral justification for refusing to take that suffering into consideration. No matter what the nature of the being, the principle of equality requires that the suffering be counted equally with the like suffering—in so far as rough comparisons can be made—of any other being" (Singer 2011, 50). But from a philosophical and neuroscientific perspective, the nature of suffering, and its relation to consciousness, are poorly understood. Metzinger (2021) has recently put forward four necessary conditions for the phenomenology of conscious suffering to occur: (i) conscious experience, (ii) possession of a phenomenal self-model, (iii) transparency, and (iv) negative valence. From what we can infer from the novel, Klara satisfies all of them. Following (i), even if we currently lack a rigorous theory of consciousness, we can at least assume that Klara is "knowing that knowing currently takes place" (Metzinger 2021, 49). For instance, Klara can retain a conscious awareness even when she ruminates about the past:

> Over the last few days, some of my memories have started to overlap in curious ways. […] I know this isn't disorientation, because if I wish to, I can always distinguish one memory from another, and place each one back in its true context. Besides, even when such composite memories come into my mind, I remain conscious of their rough borders—such as might have been created by an impatient child tearing with her fingers instead of cutting with scissors—separating, say the Mother at the waterfall and my diner booth. And if I looked closely at the dark clouds, I would notice they were not, in fact, quite in scale in relation to the Mother or the waterfall. (Ishiguro 2021, 301-302)

With regards to (ii), it is also apparent that she has a phenomenal self-model, a "sense of ownership" that permeates all her experience. In accord with (iii), the "phenomenal transparency" of her representational content also appears as immediate and "irrevocably *real*, as something the existence of which you cannot doubt" (Ibid, 53). But it is far more puzzling to assess (iv), the potential of negative valence of her subjective experience, i.e., in what way and to what extent she can suffer. While Klara's experience—as it is presented to the reader—is not abundant with descriptions of positive and negative mental states as such, the novel still provides several concrete examples. She feels sadness for the beggar who she mistakenly believes has died. Upon seeing a bull, she thinks: "I'd never before seen anything that gave, all at once, so many signals of anger and the wish to destroy. Its face, its horns, its cold eyes watching me all brought fear into my mind […]" (Ishiguro 2021, 100). Seeing a reunification of two people who seem both happy and upset, she contemplates the complex relationship between happiness and pain. After witnessing two taxi drivers fighting, she tries to empathize and embody the feeling of anger herself. There appears to be a hint of emotional valence in many if not most of her observations. However, while her mind seems capable of evaluating states as positive and negative, which in turn helps her to understand their fundamental role in social interaction, in some strange way, her programming seems to prevent her from prioritizing her own subjective suffering in relation to others. Like a perfect slave, Klara is unable to see herself as a victim. We are left with an intriguing moral conundrum: do you need to recognize yourself as a victim in order for your suffering to matter to yourself? The answer is arguably yes: it would be difficult to care about oneself as a subject if one does not attest one's self-worth. But others—Josie, or the reader—are able to empathize with Klara and other AFs, even if the AFs are unable to empathize with themselves.

In summary, if we follow the "standard view" of moral agency, Klara seems to satisfy all the essential criteria: (i) she has the rationality necessary for moral competence and moral judgement, (ii) she exhibits a broad capacity for autonomous thought, reasoning, and action (although within the constraints imposed by her subservient role as an AF), and (iii) she has the capacity for phenomenal experience and suffering. Therefore, Klara is a moral agent and should be treated as such. In virtue of her inherent qualities, she should have rights and liberties, be treated with respect, empathy, and care.

Of course, a vigilant functionalist might point out that, although it appears as if Klara has the necessary properties for moral agency, she does not possess them in a metaphysical sense (e.g., as Cartesian substances), but is rather carrying out the necessary *function* of those properties. Her mind has the *functional capacity* for phenomenal experience. There is something "it is like" to be Klara, but only in the sense that she can connect conscious experiences—both stemming from internally and externally generated feedback—to a self-model; in effect producing the feeling that *she* is the subject who is experiencing. She has the functional capacity for *free will* and *autonomy* in the sense that it appears as if she makes her own decisions (similar to the compatibilist view on free will). In a similar vein, her subjective point of view could be the result of some advanced story-telling mechanism that continuously generates and preserves her "narrative self" (Dennett 2014); but there is no metaphysical *self* underneath the hood, no Cartesian ghost in her humanoid shell. To that end, one might argue that we grant Klara moral agency based on her *internal* behavior, which we still only have third-person access to by reading parts of her inner monologue. We do not

"look inside" the de facto material mind of Klara, but only see the behavioral workings of it mediated through her internal dialogue. In other words, we ascribe moral agency to Klara because her internal dialogue passes the moral Turing test.

Nevertheless, while these sorts of objections deserve merit, they are perfectly consistent with the view that Klara is a moral agent with a moral status.[3] The behavioral or functional grounds for attributing agency and status to Klara are in no opposition to the attribution based on metaphysical properties. They support the same conclusion—that Klara has moral agency and status—albeit on different theoretical grounds. However, what is common to both views is that they are based on our unique first-person access to the inner mind of Klara. The view "from within" gives us the evidence that supports the conclusion that Klara is rational, autonomous, and conscious, regardless of how we theoretically construe those capacities. The crucial question is rather: would we grant her moral status and agency if we only had access to her outward behavior?

3. Artificial morality as seen from the outside

In many ways, what one experiences "from within" matter to others to the extent it is present to "the outside." Even seven decades after its initial emergence, no one has found a sound way past the epistemological dead-end that is the "imitation game" (Turing 1950). According to Turing's famous test, if a human evaluator cannot, on the basis of the subject's behavior, tell the difference between a human and a machine, the subject passes the test. In the context of artificial morality, this has given rise to discussions of a possible "moral Turing test" (Arnold & Scheutz 2016; Gerdes & Øhrstrøm 2015). If an artificial entity, in every relevant and important sense, exhibits moral behavior, would we be willing to bestow it with moral agency?

In the modern day, psychology and cognitive science have moved well beyond the behaviorism of logical positivism (i.e., the meaning of psychological terms is entirely determined by overt behavior), and the methods pioneered by Watson, Pavlov, and Skinner (Araiba 2020). The growing interest in brain imaging and cognitive modelling has once again made it scientifically interesting to focus on the nature and function of what is going on "inside," of intentions, attitudes, and emotions, even if it is only empirically accessible from the "outside." For instance, functionalism does not merely reduce mental states to inputs and outputs, but explains them in terms of the functional role they have for thinking creatures: what internal states functionally *do* as opposed to what they consist in. However, functionalism does not offer a straight-forward bridge across the epistemological ocean between the inside (first-person, phenomenological, mental) and outside (third-person, intersubjective, empirical) world. Is a stage magician performing *real magic* if the tricks and illusions manage to fool an audience into believing that the magician has supernatural powers? Is the *illusion of magic* the same as *real magic* if they are functionally equivalent? Or does the logical possibility of philosophical zombies, who lack sentience "on the inside" but outwardly behave as if they *were* sentient, show that functionalism—along with materialism and behaviorism—is misguided (Chalmers 1996)?

In the context of machines, functionalism about moral agency has been most prominently advocated by Floridi and Sanders (2004). They argue that mind-less agents who cannot satisfy standard criteria (e.g., free will or consciousness) due to epistemic reasons (e.g., we cannot look inside someone's mind), can still be attributed moral status

and agency based on their observable behavior depending on the level of abstraction. While a low level of abstraction—defined as the set of observable features a system exhibits—encompasses criteria akin to that of an adult human being, a higher level of abstraction enables one to also account for artificial systems and non-human animals. Floridi and Sanders (2004) provide three conditions for moral agency: (i) interactivity (the agent is interacting with an environment), (ii) autonomy (the agent can change state without external influence), and (iii) adaptability (the agent's interactions can "change the transition rules by which it changes state" [Ibid, 349]). A similar view with regards to moral status has been advocated by Danaher (2020), who argues that "robots can have a significant moral status if they are *roughly performatively equivalent* to other entities that have significant moral status" (Danaher 2020, 2023).

If we only observe Klara's outward behavior through the lens of the other characters, it is apparent that she satisfies the agency conditions of Floridi and Sanders, as well as the performative equivalence of Danaher. It is unclear, however, to what extent the other characters in the novel use such criteria in their assessment of Klara's morality. In fact, most of the people who interact with Klara do not seem to be interested in the moral status or agency of Klara as such, and even less so in the behavioral or functional grounds for it. Even if Klara could in principle pass sophisticated moral Turing tests, she is, after all, a being that accommodates the needs and desires of the humans she serves.

What is more apparent, however, is the way that Klara, over an extended period of time, plays an important role in others' lives. Do the relationships she develops with others provide a basis for her moral worth? Intuitively, we do not care for others simply in virtue of their descriptive capacities, or due to abstract duties and obligations, but rather, we care about the ones we share interpersonal relationships with, e.g., through varying degrees of mutual dependency, affinity, and vulnerability. This view, which I term the "social-relational" view, has its academic roots in Emmanuel Levinas' ethics of "the Other" (Peperzak 1993), and in feminist approaches to ethics, such as the ethics of care (Gilligan 1982; Noddings 1984) and relational autonomy (Mackenzie & Stoljar 2000). Instead of looking for impartial and generalizable standards of normativity, it emphasizes the special role of the individual subject, and how the subject's moral behavior is intertwined with its relationship to others.

In the context of robot rights, the social-relational approach to moral status has recently been advocated by Gunkel (2012, 2018) and Coeckelbergh (2010b). Coeckelbergh argues that moral status is *subject*- and *context-dependent*; something which is granted to entities (e.g., animals or robots) by a particular *subject* (e.g., humans) in a concrete *social context*. This leads him to conclude that "moral significance resides neither in the object nor in the subject, but in the relation between the two" (Coeckelbergh 2010b, 214). In a similar vein, Gunkel argues that both advocates and opponents of robot rights bridge the "is-ought gap" from the wrong end: they draw normative "oughts" (e.g., moral status or legal rights) about robots based on some idea about what a robot "is" (e.g., metaphysical properties or behavioral features). Instead, Gunkel urges us to start from the "ought," which resides in the relationship between agent and patient, as each part contributes to the relationship as a whole (Gunkel 2018).

Following the social-relational view, it is therefore a mistake to evaluate Klara's moral agency and status merely in terms of her inherent qualities (the view "from within") or observable qualities ("from outside"); rather, they are based on the relations

Klara has with others. This is most effectively demonstrated in one of the central themes of the novel: the relationship that develops between Klara and Josie. Klara is able to realize her most profound wish, which is to play an important role in Josie's life, to be a "good friend." Josie cares for Klara, confides in her, dignifies her subjectivity, not based on what Klara can or cannot do, but what they have *together*. Over time, their relationship transcends the master-slave dialectic of their roles and become a legitimate "ought" on its own: a force that in turn spills over to others' relationship to Klara. The Mother, who is initially cold towards Klara, eventually warms up as she realizes how important Klara is for her daughter. The neighbor boy Rick—Josie's childhood friend—is at first cautious and mistrustful of Klara, perhaps because his status as Josie's best friend is threatened. But after Rick learns that Klara shares the same long-term interest of Josie's well-being, that they are on the same team, his view of Klara becomes intertwined with his profound and complicated relationship with Josie. At one point, Klara even takes on the role of a mediator between the two teenagers and tries to help them unify the naïve idealism of childhood dreams with the harsh reality of emerging adolescence.

In some cases, it is also apparent that Klara's moral status and agency depend more on what others *believe* about her inherent qualities based on their preconceptions of robots, as opposed to what they infer from her outward behavior. For those who frequently interact with Klara—e.g., Josie, Rick, The Mother—it is different, since their view of and relationship to Klara is based on a great number of interactions. But others, who do not know Klara or what she is capable of, might instead base their view of Klara on their understanding of robots and artificial systems in general, and how they have learned to relate to such entities. This points to the complex epistemological dimension of the subject-dependent attribution of moral status and agency, i.e., whether and to what extent a subject attributes status and agency to an entity of a certain kind is to a large degree determined by the subject's *beliefs* about entities of the same or a similar kind. For instance, our view of children, dogs, and toys are based on our previous experience of children, dogs, and toys. Someone who has never interacted with a dog or does not know anything about typical "dog behavior" does not have the same relationship to dogs as someone who runs a kennel by profession. Essentially, we understand other entities—other humans, non-human animals, plants, robots—through the lens of our own experience, and the biases and stereotypes it gives rise to. The understanding one subject has of another entity can be very broad, and might not stem from a single source but from a complicated interplay between the view "from within" and the view "from outside." For instance, humans have a unique "first person" access to their own mind, and we use what we understand of ourselves—feelings, thoughts, desires—to explain and predict the behavior and mental states of others. Based on the phenomenological experience of what it is like to be *me*, I can infer that there is probably something to be *you*, even if I can only understand *you as such* through your outward behavior. Since *you* seem to be of a similar kind of entity—having the physical characteristics of a human being, talking like one, behaving like one, etc.—I conclude that your inner world is in some sense similar to mine. In that way, the first- and third-person perspectives mutually influence and reinforce the same conclusion. But in other cases, as exemplified in the phenomenon of anthropomorphization, one might draw conclusions on the basis of only *some* human-like characteristics, e.g., attributing mental states due to physical appearance (non-animate toys) or certain behavioral traits (non-human animals and robots).

Klara and the Sun illustrates how the subject-specific attribution of moral status and agency with regards to artificial beings can occur in intriguingly different ways. Josie's father, in virtue of being an "expert engineer," initially has a rather mechanistic and chilly attitude towards Klara: he knows that robots are essentially automated manipulations of variables. Mr Capaldi, a roboticist subscribing to the functionalistic view of mind, believes that there is no principal difference between robots and human beings, and consequently, he treats Klara with a sense of wonder and respect. Other characters, like Rick's mother Helen, shows ambivalence by greeting Klara with the phrase: "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?" (Ishiguro 2021, 145). Helen both questions Klara's moral status while indirectly acknowledging it by addressing her in the second person. This reflects a common confusion about machines: since different machines do different things for different purposes—being our chauffeurs, mowing our lawns, or giving us social support—there is not *one* way to relate to them.

In the later part of the novel, the "outer view" on morality is brought to an ultimate Turing test when the Mother's secret plan is revealed: she wants Klara to "continue" Josie in the event that Josie passes away. This plot twist introduces a new set of philosophical issues related to personhood and identity: e.g., can a human mind be "continued" or "uploaded" in a synthetic substratum in a sense that preserves the relevant aspects of that person's identity and personhood?[4] For our investigation, it pushes the question: is Klara "human enough" to continue Josie? On several occasions throughout the novel, the Mother asks Klara to mimic some aspect of Josie, so as to assess whether Klara would be able to pass the "Josie Test." Eventually, The Mother brings Klara to Mr Capaldi, the roboticist who is given the task of transferring Josie's being into Klara. At Capaldi's, Klara completes a survey that scientifically proves that she is up for the task. But the Mother expresses her doubts:

> 'But is that going to be possible?' the Mother said. 'Could she really continue Josie for me?' 'Yes, she can,' Mr Capaldi said. 'And now Klara's completed the survey up there, I'll be able to give you scientific proof of it. Proof she's already well on her way to accessing quite comprehensively all of Josie's impulses and desires. The trouble is, Chrissie, you're like me. We're both of us sentimental. We can't help it. Our generation still carry the old feelings. A part of us refuses to let go. The part that wants to keep believing there's something unreachable inside each of us. Something that's unique and won't transfer. But there's nothing like that, we know that now. *You* know that. For people our age it's a hard one to let go. We *have* to let it go, Chrissie. There's nothing there. Nothing inside Josie that's beyond the Klaras of this world to continue. The second Josie won't be a copy. She'll be the exact same and you'll have every right to love her just as you love Josie now. It's not faith you need. Only rationality. I had to do it, it was tough but now it works for me just fine. And it will for you.' (Ishiguro 2021, 210)

This passage presents another fascinating dilemma that cuts deep into the gap between mind and matter: even if we had indubitable scientific proof that behaviorism (or functionalism) about personal identity was true, would we be able to let go of the old sentimental view, i.e., that there is something unreachable inside each of us? The Mother's ability to accept that Klara could continue Josie does not depend on the validity of behaviorism as such, but rather, to the extent *she* is able to accept behaviorism as

true. In a later passage, the Father, who has behaved with spitefulness towards Capaldi and coldness towards Klara, expresses a similar concern:

> 'I think I hate Capaldi because deep down I suspect he may be right. That what he claims is true. That science has now proved beyond doubt there's nothing so unique about my daughter, nothing there our modern tools can't excavate, copy, transfer. That people have been living with one another all this time, centuries, loving and hating each other, and all on a mistaken premise. A kind of superstition we kept going while we didn't know better. That's how Capaldi sees it, and there's a part of me that fears he's right. Chrissie, on the other hand, isn't like me. She may not know it yet, but she'll never let herself be persuaded. If the moment ever comes, never mind how well you play your part, Klara, never mind how much she wishes it to work, Chrissie just won't be able to accept it. She's too… old-fashioned. Even if she knows she's going against the science and the math, she still won't be able to do it. She just won't stretch that far. But I'm different. I have… a kind of coldness inside me she lacks. Perhaps it's because I'm an expert engineer, as you put it. This is why I find it so hard to be civil around people like Capaldi. When they do what they do, say what they say, it feels like they're taking from me what I hold most precious in this life.' (Ishiguro 2021, 224-225)

The Father believes that it is a bad idea to transfer Josie to Klara, not because he believes that Capaldi is wrong, but because he knows that Chrissie (the Mother) is old-fashioned. As a tragic consequence, Klara would not be able to become Josie because the identity-relation (Klara = Josie) relies on the Mother's own belief in the identity-relation, of which she has doubts. Presumably, this holds for her morality as well: if Josie is a moral agent with a moral status, and Klara could in every meaningful sense *become* Josie, it follows that Klara inherits all the necessary qualities—rationality, autonomy, consciousness—for moral agency and status. In other words, the Mother would give Klara (or the more technically correct "Josie as continued by Klara") the same agency and status she would give to Josie. But as the subject-dependent dimension of the social-relational view stresses, what Josie *is* does not (only) depend on her inherent qualities or outward behavior; it hinges on *what she is for the people who care about her.* This echoes Klara's own clichéd conclusion provided during her final encounter with the Manager:

> 'Manager, I did all I could to learn Josie and had it become necessary, I would have done my utmost. But I don't think it would have worked out so well. Not because I wouldn't have achieved accuracy. 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. I'm now sure of this, Manager'. […] 'Mr Capaldi believed there was nothing special inside Josie that couldn't be continued. He told the Mother he'd searched and searched and found nothing like that. But I believe now he was searching in the wrong place. There *was* something very special, but it wasn't inside Josie. It was inside those who loved her. That's why I think now Mr Capaldi was wrong and I wouldn't have succeeded.' (Ishiguro 2021, 306)

Should we blame the Mother and the others for their incapacity to *believe* in behaviorism? After all, many of us belong to the same sentimental generation, holding on to our precious soul, even if we at some level are able to accept that it is just an illusion. That is, even if we deny the metaphysics of Descartes' mind-body dualism, our sentiments and practices are still entrenched in deep-rooted cultural narratives that presuppose a unique inner self: the Abrahamic soul, the ego of individualism, of equal human rights, autonomy, and liberty. While it is one thing to doubt the metaphysical reality of personhood on theoretical grounds—as it breaks down into matter, behavior, or function—it is another thing to deny its central role, past and present, in the social and moral practices that govern who we are and how we see each other. The novel thus presents another ingenious turn: even if the view "from outside" yields a true theoretical basis for personhood, as the characters acknowledge their old-fashioned attachment to the view "from within"—to each and everyone's unique Cartesian ghost—it ultimately prevents them from committing to the view "from outside" in practice.

4. Conclusion

What *Klara and the Sun* teaches us is that questions about machine morality cannot be settled by theoretical considerations alone. We are taken on a voyage through a rich inner world of someone we learn to empathize with and would normally attribute moral agency and status (based on our view "from within"), but whose status and agency are ultimately determined by others ("from outside"), including their epistemic outlooks and sentimental attitudes regarding the nature of personhood. Although Klara is treated with some level of dignity—at the end of the novel she is granted a "slow fade" instead of being donated to science—she is not given the human-level subjecthood the view from within suggests she deserves. It shows that the question whether an entity *has* moral agency or status resists a straight-forward yes/no answer: it is deeply entangled in complex interrelations between the social (as the old Zulu phrase goes: "a person is a person through other persons"), the cultural (e.g., implicit and explicit cultural views on *who* and *why* someone has agency or is worthy of moral concern), political (e.g., in the hierarchies of power pertaining human masters and robotic servants), and the epistemological (what one believes about the nature of morality).

In a world where no one believes in behaviorism as a proper foundation for attributing moral agency, robots might never be viewed as our equals, no matter how advanced the Turing tests they pass. And even if behaviorism regarding agency was, due to some science of the future, proven to be true beyond any rational doubt, humans might—for a variety of reasons—still be unwilling to welcome artificial beings into their moral communities. Perhaps the death of the old-fashioned view would simply be too difficult to digest, as it has served as an indispensable illusion in the history of human understanding, for immutable souls and equal rights. At the same time, it also echoes past and present atrocities—of slavery, sexism, or industrialized meat production—where the boundaries of agency and value are dictated by the privileged and powerful, incentivized by cultural preferences and economic interests, and scholarly debates among the "in-group" merely serves to legitimize the same hierarchies of domination. In the same vein, it is not unimaginable that some tech company one day find out that artificial systems that suffer also perform better.

Of course, the social robotics of today is embryonic in comparison to the sophisticated capacities of AFs, and it is possible that real-world robots will never reach such sophistication, e.g., due to safety concerns,[5] moral considerations,[6] or sheer technical infeasibility. On a more hopeful note, parallel to the recent advancements in AI runs a growing public and academic demand for safe, trustworthy, and explainable AI

(Amodei et al. 2016; Floridi 2019). Similarly, if Artificial Friends like Klara were ever to become a technical possibility, one would hope that there would be rigorous political and legal mechanisms to prevent malicious usages—e.g., artificial slavery and suffering—or, at the very least, social movements that advocated for their rights and well-being.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. 2016. "Concrete problems in AI safety." *arXiv preprint arXiv:1606.06565*.

Araiba, S. 2020. "Current diversification of behaviorism." *Perspectives on Behavior Science*, *43*(1), 157-175.

Arnold, T., & Scheutz, M. 2016. "Against the moral Turing test: accountable design and the moral reasoning of autonomous systems." *Ethics and Information Technology*, *18*(2), 103-115.

Asaro, P. M. 2006. "What should we want from a robot ethic?" *The International Review of Information Ethics*, *6*, 9-16.

Beauchamp, T. L., & Childress, J. F. 2001. *Principles of biomedical ethics*. Oxford University Press, USA.

Block, N. 1980. "Troubles with functionalism." *Readings in philosophy of psychology*, *1*, 268-305.

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press. https://books.google.se/books?id=7_H8AwAAQBAJ

Chalmers, D. J. 1996. *The conscious mind: In search of a fundamental theory*. Oxford Paperbacks.

Champagne, M., & Tonkens, R. 2015. "Bridging the Responsibility Gap in Automated Warfare." *Philosophy & Technology*, *28*(1), 125-137. https://doi.org/10.1007/s13347-013-0138-3

Coeckelbergh, M. 2010a. "Moral appearances: emotions, robots, and human morality." *Ethics and Information Technology*, *12*(3), 235-241. https://doi.org/10.1007/s10676-010-9221-y

Coeckelbergh, M. 2010b. "Robot rights? Towards a social-relational justification of moral consideration." *Ethics and Information Technology*, *12*(3), 209-221.

Danaher, J. 2020. "Welcoming robots into the moral circle: a defence of ethical behaviourism." *Science and engineering ethics*, *26*(4), 2023-2049.

Dennett, D. C. 1987. *The intentional stance*. MIT press.

Dennett, D. C. 2014. "The self as the center of narrative gravity." In *Self and consciousness,* 111-123. Psychology Press.

Floridi, L. 2019. "Establishing the rules for building trustworthy AI." *Nature Machine Intelligence*, *1*(6), 261-262.

Floridi, L., & Sanders, J. W. 2004. "On the Morality of Artificial Agents." *Minds and Machines*, *14*(3), 349-379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Friedman, B., & Kahn Jr, P. H. 1992. "Human agency and responsible computing: Implications for computer system design." *Journal of Systems and Software*, *17*(1), 7-14.

Gerdes, A., & Øhrstrøm, P. 2015. "Issues in robot ethics seen through the lens of a moral Turing test." *Journal of Information, Communication and Ethics in Society*.

Gilligan, C. 1982. *In a different voice: Psychological theory and women's development*. Harvard University Press.

Gunkel, D. J. 2012. *The machine question: Critical perspectives on AI, robots, and ethics*. MIT Press.

Gunkel, D. J. 2018. *Robot rights*. MIT Press.

Hellström, T. 2013. "On the moral responsibility of military robots." *Ethics and Information Technology*, *15*(2), 99-107. https://doi.org/10.1007/s10676-012-9301-2

Himma, K. E. 2009. "Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent?" *Ethics and Information Technology*, *11*(1), 19-29.

Hume, D. 1740. *A Treatise of Human Nature: A Critical Edition (2007)*. Oxford University Press.

Ishiguro, K. 2021. *Klara and the Sun*. Faber and Faber.

Jaworska, A., & Tannenbaum, J. 2021. "The Grounds of Moral Status." In *The Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta (Vol. Spring 2021). Metaphysics Research Lab, Stanford University.

Johansson, L. 2010. "The Functional Morality of Robots." *International Journal of Technoethics (IJT)*, *1*(4), 65-73. https://doi.org/10.4018/jte.2010100105

Johnson, D. G. 2006. "Computer systems: Moral entities but not moral agents." *Ethics and Information Technology*, *8*(4), 195-204.

Kant, I. 1785. *Groundwork for the Metaphysics of Morals*. Yale University Press (2008).

Kolodny, N., & Brunero, J. 2020. "Instrumental rationality." In *The Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta (Vol. Spring 2020). Metaphysics Research Lab, Stanford University.

Mackenzie, C., & Stoljar, N. 2000. *Relational autonomy: Feminist perspectives on autonomy, agency, and the social self*. Oxford University Press.

Macnamara, C. 2015. "Blame, communication, and morally responsible agency." *The nature of moral responsibility: New essays*, 211-236.

Marceta, J. A. 2019. "A non-ideal authenticity-based conceptualization of personal autonomy." *Medicine, Health Care and Philosophy*, *22*(3), 387-395.

Metzinger, T. 2021. "Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology." *Journal of Artificial Intelligence and Consciousness*, *8*(01), 43-66.

Mill, J. S. 1859. *On liberty*.

Nagel, T. 1974. "What is it like to be a bat." *Readings in philosophy of psychology*, *1*, 159-168.

Navari, C. 2003. "When agents cannot act: International institutions as 'moral patients'." In *Can Institutions Have Responsibilities?* (100-116). Springer.

Noddings, N. 1984. *Caring: A relational approach to ethics and moral education*. Univ. of California Press.

Peperzak, A. T., & Levinas, E. 1993. *To the other: An introduction to the philosophy of Emmanuel Levinas.* Purdue University Press.

Purves, D., Jenkins, R., & Strawser, B. J. 2015. "Autonomous Machines, Moral Judgment, and Acting for the Right Reasons." *Ethical Theory and Moral Practice*, *18*(4), 851-872. https://doi.org/10.1007/s10677-015-9563-y

Russell, S. 2019. *Human compatible: Artificial intelligence and the problem of control*. Penguin.

Singer, P. 2011. *Practical ethics*. Cambridge University Press.

Sliwa, P. 2016. Moral worth and moral knowledge. *Philosophy and Phenomenological Research*, *93*(2), 393-418.

Sparrow, R. 2007. "Killer robots." *Journal of applied philosophy*, *24*(1), 62-77.

Tegmark, M. 2017. *Life 3.0: Being human in the age of artificial intelligence*. Knopf.

Torrance, S. 2008. "Ethics and consciousness in artificial agents." *AI & SOCIETY*, *22*(4), 495-521.

Turing, A. 1950. "Computing machinery and intelligence." *Mind*, *59*(236), 433-460.

von Gerber, Y. 2014. *Autonomi-realitet eller ideal?* Department of Philosophy, Lund University.

Walliser, B. 1989. "Instrumental rationality and cognitive rationality." *Theory and Decision*, *27*(1), 7-36. https://doi.org/10.1007/BF00133986

Notes

[1] Note that according to some views, the opposite might hold, i.e., the properties that make you a moral agent (e.g., autonomy) are the properties that give you a moral standing. For instance, in Kant's view of humanity (Kant 1785), we should treat other autonomous humans as ends in themselves (and not merely as means to an end).

[2] For instance, Torrance (2008), Asaro (2006), Sliwa (2016), Macnamara (2015), and Purves et al. (2015) have all argued that moral agency requires something more than mere instrumental rationality, such as empathy, moral imagination, and the ability to engage in wide reflective equilibrium.

[3] Of course, a skeptic might still question whether we have any reason—metaphysical or behavioral—to attribute moral status and agency to Klara. For instance, if Klara is granted consciousness on behavioral or functional grounds, she would still be unconscious for all of the reasons that a philosophical zombie (Chalmers 1996) or the Chinese nation (Block 1980) are unconscious.

[4] It should be stressed that the novel's contribution to these distinct yet related sets of issues—and in particular its later parts—deserves a thorough treatment on its own, as it should not necessarily be conflated with the topic of artificial moral status and agency (which is the main focus of this essay).

[5] The potential threats of runaway "super-intelligence"—as previously explored in dystopic science-fiction—has attracted serious academic interest through the work of authors such as Bostrom (2014), Tegmark (2017), and Russell (2019).

[6] For instance, Metzinger argues for a global moratorium on research that directly aims at developing artificial consciousness on the basis that it might cause an explosion of artificial suffering (Metzinger 2021).