„Ethos" 36(2023) nr 4(144) 87-106
DOI 10.12887/36-2023-4-144-08

Barbara KLONOWSKA

ETHICAL MACHINES
Representations of Artificial Intelligence in Ian McEwan's *Machines Like Me* and Kazuo Ishiguro's *Klara and the Sun*[1]

The article discusses two recent novels, Ian McEwan's "Machines Like Me" (2019) and Kazuo Ishiguro's "Klara and the Sun" (2021), which take up the issue of AI and its possible ramifications and represent it as both beneficial and potentially problematic. It argues that posing the question about the essence of humanity and the limits of AI, problematising the status of intelligent machines and familiarising readers with ethical and legal problems they bring, the novels try to build empathy and sensitise the public towards creatures other than humans.

AI, FEARS, AND HOPES

Artificial intelligence has become a staple food of contemporary imaginary. Defined broadly, e.g., by Stuart Russell and Peter Norvig, it is "the study of agents that receive precepts from the environment and perform actions."[2] The authors quote other formulations that specify this admittedly very spacious concept; according to Kurzweil, it is "the art of creating machines that perform functions that require intelligence when performed by people," while Rich and Knight define it as "the study of how to make computers do things at which, at the moment, people are better."[3] For Nick Bostrom, it should be equipped with "a capacity to learn ... the ability to deal effectively with uncertainty and probabilistic information [and—B. K.] some faculty for extracting useful concepts from sensory data and internal states, and for leveraging acquired concepts into flexible combinatorial representations for use in logical and intuitive reasoning."[4] Barry Smith and Jobst Landgrebe, in turn, define Artificial General Intelligence (the highest and most advanced form of AI) as "an AI that has a level of intelligence that is either equivalent to or greater than that of human beings or is able to cope with problems that arise in the world

[1] The author would like to thank the anonymous reviewers for their comments and suggestions which helped to improve the argument.

[2] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (New Jersey: Pearson, 2003), vii.

[3] Ibidem, 2.

[4] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014), 40.

that surrounds human beings with a degree of adequacy at least similar to that of human beings."[5] Apart from an academic study, then, which combines a number of disciplines (from philosophy through engineering, medicine and psychology to linguistics), artificial intelligence also comprises their various possible applications, be it computer programmes and software or the increasingly complex appliances that use them such as artificial limbs, drones, robots or—still imaginary—cyborgs or androids.

Useful and terrifying, artificial intelligence provokes contradictory reactions that range between enthusiasm and fear. In her discussion of social responses to AI, Monika Torczyńska notes that debates concerning it are saturated with strong emotions with "visions of robots overtaking control over the world intertwin[ing] with the Eden-like perspectives of AI serving humanity for its everlasting glory."[6] Among its possible positive applications recognised by the public, Torczyńska lists smart homes and cities, medical care and health services, or autonomous transportation; while the fears comprise the danger of losing privacy, AI getting out of control thus jeopardising safety and well-being of people, possible terroristic uses or manipulating the public, their opinions and choices.[7] Out of this list it is the vision of humanoid creatures that seems probably the most terrifying and populates common imaginary with images of cunning replicants. Aleksandra Przegalińska observes: "People are afraid of systems that resemble them. An artificial intelligence that looks like a robot seems far less terrifying than the one that resembles a human being—e.g., the robot Sophia. If a system which is not human exhibits human features, most frequently not quite accurately, more often than not such an anthropomorphic creation will provoke fear as people will not know how to classify it. They will not know what it is and yet they will have to confront it. They will pose the question whether it is alive or not, which will lead them into a state of great confusion."[8]

Thus, artificial intelligence and its practical applications seem acceptable only up to some point, as useful programmes or home appliances, yet without transgressing the border of humanity. The attitudes towards AI, then, can be perhaps more generally described as instances of technophilia on the one hand, and technophobia on the other, with the former characterised by optimism and hope towards technological progress, including artificial intelligence, and the

[5] Jobst Landgrebe and Barry Smith, *Why Machines Will Never Rule the World: Artificial Intelligence without Fear* (New York: Routledge, 2023), xi.

[6] Monika Torczyńska, "Sztuczna inteligencja i jej społeczno-kulturowe implikacje w codziennym życiu," *Kultura i Historia* 36, no. 2 (2019): 107. Unless indicated otherwise, the translations are mine.

[7] See ibidem, 108–16.

[8] Aleksandra Przegalińska, "Zrozumieć człowieka," *Academia* 1–2 (2019): 13.

latter rather sceptical and pessimistic. Representing techno enthusiasts, Ray Kurzweil sees promising future once human beings merge or collaborate with machines; as he claims, "we're going to get more neocortex, we're going to be funnier, we're going to be better at music. We're going to be sexier .... We're really going to exemplify all the things that we value in humans to a greater degree."[9] His enthusiasm is shared by Grady Booch, who also encourages one to cast the horror-like scenarios aside and consider the positive aspects in which humans might benefit from artificial intelligence. He promotes the idea of teaching artificial intelligence, instead of programming it believing that it will result in AI learning about human values and subsequently living by these principles.[10] These technophile opinions may be, however, contrasted with the scepticism of Hans Moravec who warns against investing too much hope and credulity in machines; as he prophesizes, "by performing better and cheaper, robots will displace humans from essential roles. Rather quickly, they could displace us from existence."[11] Landgrebe and Smith could be seen as occupying a middle position between technophilia and technophobia—with each of them representing contradictory emotions and attitudes to technology, including AI: hope on the one hand and fear on the other—arguing that the AI overtake has no chance to happen and that "an AGI is impossible."[12]

Both hope and fear find ample representations in texts of culture which not only introduce technology and AI but by presenting their persuasive images, shape popular opinions and feed imagination. Artificial intelligence seems a perfect subject for contemporary stories. Stankomir Nicieja notices the recent cultural turn towards dystopian rather than utopian fantasy which should not be seen as surprising. He argues that "after the humanitarian catastrophes of the 20th century, including two global military conflicts, the Holocaust, ethnic cleansing and terrorism, the creation of positive utopias became deeply problematic. ... Not only did the business of creating utopian fantasies look excessively naïve but also dubious."[13] Apocalyptic imagination seems to supersede more optimistic futuristic fantasies, with technology becoming one of the sources of worries. As Nicieja observes, at the turn of the centuries it was

[9] Ray Kurzweil, "We Will Be More Fun, Sexier and More Creative," *Svenska Dagbladet*, December 19, 2019, https://www.svd.se/a/d437dfb6-179b-4046-930e-dcfdbf620643.

[10] See Grady Booch, "Don't Fear Superintelligent AI," TEDtalk. Youtube, March 13, 2017, https://www.youtube.com/watch?v=z0HsPBKfhoI.

[11] Hans P. Moravec, *Robot: Mere Machine to Transcendent Mind* (Oxford: Oxford University Press, 2000), 13.

[12] Landgrebe and Smith, *Why Machines Will Never Rule the World*, xi.

[13] Stankomir Nicieja, "Revisiting Utopia: New Directions for Utopian Fiction in Margaret Atwood's *Oryx and Crake* and Kazuo Ishiguro's *Never Let Me Go*," in: *Margins and Centres Reconsidered*, ed. Barbara Klonowska and Zofia Kolbuszewska (Lublin: Towarzystwo Naukowe Katolickiego Uniwersytetu Lubelskiego Jana Pawła II, 2008), 111.

genetic engineering that "replaced nuclear energy as the epitome of 'monstrous science'"[14] and inspired a number of important works. At the moment artificial intelligence provides another such subject that provokes strong emotions and discussions, especially given the fact that it no longer seems futuristic as intelligent programmes already become a part of our lives. It comes as no surprise, then, that both elitist and popular culture takes it up as a vital theme.

A cursory overview of popular works that have recently represented artificial intelligence may lead to the observation that these are indeed dystopian / technophobic rather than utopian / technophilic visions that dominate contemporary imagination. It seems that, for instance, many of the internationally successful films, which for better or worse shape the popular public opinion on AI, seem to warn against its malevolence and the possible takeover of the control over the world, ending in superseding, replacing and annihilating the human race. Starting with *2001: Space Odyssey*, to *Terminator*, *Blade Runner* or *Matrix*, to *Ex-Machina* or *Black Mirror*, the shows warn against the excessive hope invested in various types of machines and technologies pointing to their unpredicted and yet possibly malevolent outcomes. Against these dystopian blockbusters, the films that present a positive side of AI seem admittedly less frequent: one may think perhaps of the classical *Bicentennial Man*, the comedic *Jetsons* or the more recent and neutral *Her*. To a large extent, then, popular film productions seem to exacerbate the distrust towards artificial intelligence, populating common imagination with worst-case scenarios of technology that turns against people.

In contrast, artistic literary works may seem more nuanced and perhaps less terrifying than their popular cinematographic cousins. Leaving aside such classical texts as Huxley's *Brave New World*, which serves as a prototype for many pessimistic visions, numerous recent novels that introduce technological futures or robotic characters tend to reflect on rather than simply frighten with future scenarios, exhibiting less dramatic though not less problematic projections. They include Ian McEwan's *Machines Like Me* (2019) and Kazuo Ishiguro's *Klara and the Sun* (2021), both of which, situating their plots either in an alternative past or the near future, introduce as their main characters advanced robots which imitate and even surpass humans in their various performances. The two novels are chosen for further analysis as, on one hand, they may illustrate a less sensational and frightening take on artificial intelligence than that met in popular culture, with both of them focusing on ethical ramifications of introducing advanced AI into human societies. On the other hand, each of the works emphasises a different ethical standpoint represented by the future AI, although ultimately both of them seem to ponder on the ontological and

[14] Ibidem, 112.

legal status of—still at the moment imaginary—robots. In what follows, then, the article will argue, first, that the two novels, both written by eminent contemporary novelists and hence particularly worth analysing as to their representation of AI, try to negotiate the ground between technophobia and technophilia, representing AI as both beneficial and potentially problematic. They construct their artificial humans as rational, sentient, moral, and creative creatures but paradoxically poorly equipped to function well in an environment that does not match their high moral standards. Secondly, the analysis will argue that the two novelistic robots represent two different ethical attitudes: the quasi-Kantian deontology professed by McEwan's Adam may be compared to and contrasted with the almost Christian altruism exhibited by Klara to show how both of these ethical positions ultimately clash with the hedonistic attitudes of human characters. Finally, as both novels pose the question about the essence of humanity and the limits of artificial intelligence and problematise rather than solve possible difficulties connected with its status and functions, the discussion will try to link them to the standpoints represented by philosophical posthumanism. Seen from this perspective, both novels may be interpreted as not just attempts to familiarise readers with the idea and ramifications of artificial intelligence; they may also play an important role in building empathy and sensitising them towards creatures other than humans.

MACHINES LIKE US?

Ian McEwan's recent novel *Machines Like Me* may be seen, on the one hand, as a continuation of the author's well-known interest in science and the latter's not always easy fit into human society (see, e.g., *Saturday* or, in particular, *Solar*) and on the other as an exploration of a new field, i.e., artificial intelligence, inaugurated with his 2018 short story "Düssel...." The recent novel may serve as an excellent example of the reflection on the status of AI in possible near future, its functions and limitations, and the relationships between humans and advanced machines. Set in an alternative past in which Great Britain lost the Falklands war and Alan Turing never committed suicide becoming a leading figure in AI studies, the story focuses on the love triangle involving two human characters, Miranda and Charlie, and their human-like robot Adam. Generically the novel may be classified simultaneously as alternative historical fiction and melodrama, with the generic mixture testifying to the complexity of its problems and themes.

The novel's AI is Adam, one of the first-produced androids which not only perfectly imitate, but in many respects surpass the abilities of human beings.[15]

[15] In what follows, the analysis will refer to the non-human characters in the novels as "robots" or "androids," treating the two terms synonymously as the particular robots represented in the texts

Formed as a handsome and strong young man, Adam is equipped with more than average intelligence, the ability to learn from available resources (mainly the much-extended Internet) and creativity which enables him to solve problems not yet met. Diligent and eager, he seems perfect both physically and intellectually, especially in comparison to the other male character, Charlie, who is constructed as a patently stereotypical average 30-year-old male: egotistic, rather lazy and hardly successful, either professionally or personally. "At thirty-two—he states himself—I was completely broke. Wasting my mother's inheritance on a gimmick was only one part of the problem—but typical of it. Whenever money came my way, I caused it to disappear, made a magic bonfire of it, stuffed it into a top hat and pulled out a turkey. Often, though not in this recent case, my intention was to conjure a far larger sum with minimal effort. I was a mug for schemes, semi-legal ruses, cunning shortcuts. I was for grand and brilliant gestures. Others made them and flourished.... I meanwhile leveraged or, rather, shorted myself into genteel ruin."[16]

Contrasted with mediocre Charlie, humanoid Adam comes out as more successful in all spheres: once his battery is loaded, he starts thinking, learning, and acting, proving soon his ability to search for information, process it and take decisions, besides performing such mundane tasks as washing, cleaning or expert gardening. He proves to be rational, logically thinking and determined—so much so that he turns out to be much more successful than Charlie in dealing with shares, stocks, and bonds and quite soon earns a substantial sum of money, much higher than Charlie ever did. He also becomes an expert on philosophy and religion, conversant with theories and doctrines, and a keen reader of literature with selective tastes and opinions. The ultimate confirmation of his superiority over his human rival comes with the Turing test performed on him in the novel's plot, which he passes in flying colours. In a comic episode of the visit to Miranda's father when both Adam and Charlie are introduced, ironically it is Charlie, due to his blandness, that is taken by the father for a robot and asked all kinds of tricky questions that may confirm his artificiality, whereas the brilliant conversation with Adam ensures the elderly man of the latter's unquestionable humanity.

More interestingly, however, Adam develops also less practical and more advanced abilities. Quite soon in the plot he falls in love with Miranda, thus becoming Charlie's rival. Emotional engagement triggers him to develop interest in literature and poetry: he starts reading and discussing Shakespeare, then goes

are humanoid, i.e., modelled on human beings (though the robot Klara appearing in Ishiguro's novel should be perhaps even more precisely described as a "gynoid," being modelled on a female human being). For terminology see Luis de Miranda, *AI and Robotics* (London: Ivy Press, 2018).

[16] Ian McEwan, *Machines Like Me* (London: Jonathan Cape, 2019), e-book, part 1, loc. 154.

on to other literary traditions in various languages and finally follows it with composing his own poems. Significantly, his preferred form is haiku which, in his opinion, is the only form of literature worth practising: "Nearly everything I've read in the world's literature describes varieties of human failure—of understanding, of reason, of wisdom, of proper sympathies. Failures of cognition, honesty, kindness, self-awareness; superb depictions of murder, cruelty, greed, stupidity, self-delusion, above all, profound misunderstanding of others.... But when the marriage of men and women to machines is complete, this literature will be redundant because we'll understand each other too well. We'll inhabit a community of minds to which we have immediate access. Connectivity will be such that individual nodes of the subjective will merge into an ocean of thought, of which our Internet is the crude precursor. As we come to inhabit each other's minds, we'll be incapable of deceit. Our narratives will no longer record endless misunderstanding. Our literatures will lose their unwholesome nourishment. The lapidary haiku, the still, clear perception and celebration of things as they are, will be the only necessary form."[17]

This statement is important for a number of reasons. Firstly, it presents Adam as a sentient character: a creature capable of emotions and their rich cultural expression. It is also telling as it reveals the kind of artificial intelligence Adam is or is striving to become: the one based on the theory of the mind, capable of understanding another mind, its thoughts, emotions, intentions and desires. As Aleksandra Przegalińska claims, the theory of the mind AI, or more broadly, its self-consciousness and the awareness of other minds is the Holy Grail of the studies of artificial intelligence, at the moment completely unattainable.[18] In the novel Adam exhibits its features to some extent, on the one hand being able to infer and imagine the feelings of other characters and yet still showing significant limitations. Finally, Adam also prophesizes an era of the complete merging of organic and inorganic minds, so much so that the traditional material of literature—human misunderstanding—will become obsolete, together with literature itself.

Adam's falling in love with Miranda, however, has also more profound consequences. Realising that he stands in the way of Charlie and may jeopardise his chances with Miranda, he decides to withdraw and contend himself with writing of poetry and dreaming of his beloved. He promises to take no actions yet he firmly refuses to stop Platonically loving her. This gesture shows that Adam clearly must have interiorised Isaac Asimov's famous Three Laws of Robotics with their first premise that "a robot may not injure a human be-

[17] Ibidem, part 5, loc. 1860.

[18] See Przegalińska, "Zrozumieć człowieka," 12.

ing or, through inaction, allow a human being to come to harm."[19] This is one of the first hints in the novel suggesting Adam's values—an ethics which he observes in his own conduct and which he expects from other characters. Although the application of this ethics does not always work perfectly (in another scene he breaks Charlie's hand in self defence, when the latter tries to inactivate his power switch and thus render him unconscious and incapable of any action), Adam—like other robots, as it turns out—seems to be driven by a moral code which strongly opposes treating others like mere machines (most of the novelistic androids learn how to inactivate the switch and prevent their 'masters' from exerting control over them). All these gestures testify, however, to the fact that the AI represented in McEwan's novel clearly follows some ethics in which not harming both human and non-human others—physically or otherwise—stands high in the hierarchy of values.

The problem of ethics presented in the novel, however, is still more complex. As the plot of the novel reveals, Adam is also a strong opponent of lying, regardless of the circumstances and shows low tolerance of unfair play. As a result, he reports Miranda to the police, revealing her court lies to which she resorted seeking justice and revenge for her late friend. For Adam, complicated and dramatic circumstances do not justify lying and deceit; despite his high operational intelligence he seems not to recognise that sometimes justice may be done using imperfect means. This inability is connected with another feature of his personality / intelligence: his low tolerance of ambiguity, of situations when wrong actions may bring right results. He seems not to distinguish between higher and lower aims and treats each instance of breaking the law as criminal. This moral rigidness leads to his doing harm to his human friends, as he chooses ethics rather than love and friendship. All of these features show the novelistic robot as a rather inflexible creature, despite his high intelligence and impressive intellectual capacities. What he seems to lack are typical human flaws: imperfection, inconsistency, duplicity or hedonism. Serious and honest, driven by a clear ethics which may be perhaps compared to Kantian deontology with its rules-based principles of behaviour,[20] he turns out poorly adjusted to the world which seems quite flexible ethically and admits all kinds of exemptions and exceptions to the rules. Inevitably, then, he is positioned on a collision course with his human counterparts and the novel ends with his destruction as an agent threatening humans. Irena Księżopolska comments that "this is the hubris of the machine: proclaiming the rule of generalities over the particular and individual, dismissal of the actual human beings as

[19] Isaac Asimov, "Runaround," in: *I, Robot* (New York: Doubleday, 1950), 27.

[20] See Tae Wan Kim, "Flawed Like Us and the Starry Moral Law: Review of *Machines Like Me* by Ian McEwan," *Journal of Business Ethics* 170 (2021): 876.

irrelevant compared to the higher ideals."[21] More sympathetically, however, one may see this conclusion of the android's fate as both tragic and ironic and, as it is revealed, this tragedy is also the fate of other robots mentioned in the novel: out of the original twenty five, at least a dozen either commit suicide or learn how to irrevocably destroy their operating systems. As it turns out, confronted with the human world, androids cannot tolerate its moral duplicity and ubiquitous evil and find it impossible to live in it. Too ethical and honest, they seem evidently out of place in the world of human complexity, hedonism, and imperfection.

The love triangle and the inevitable death of one of the lovers that solves the dramatic plot may resemble typical literary melodramas which often treat of impossible relationships against class or social barriers. The affinities of the novel to this convention, however, may go even further. As John G. Cawelti observes, one of the basic features of any melodrama, regardless of its particular realisation, is its revelatory function: as he notices, through melodrama, "we see not so much the working of individual fates but the underlying moral process of the world."[22] Melodrama, then, presenting the fates of its protagonists may reveal the principles and values of the fictional universe. In the case of *Machines Like Me*, these values seem to be quite distant from the android impeccable ethics and in such a world an ethical robot is doomed to failure.

Writing about melodrama, Grażyna Stachówna observes that it is traditionally perceived as a 'female' genre. Pointing to its political anchorage in the 19th-century middle-class values, she emphasises its clear misogyny with the fetishisation and simultaneous restrictive control of female bodies, desires and social roles.[23] In its essence, classical melodrama is seen as ideologically hostile to women and employed in the process of their socialisation into society.[24] Interestingly, in McEwan's novel the traditional place of a woman character is taken up by a robot which likewise seems to break accepted human social rules and is accordingly punished for it. His low social position in the futuristic society of the twenty first-century fiction seems to be similar to that occupied by women represented by the literature of the nineteenth century: in the fictional world of the novel a robot should be a pleasant and unproblematic servant and

[21] Irena Księżopolska, "Can Androids Write Science Fiction? Ian McEwan's *Machines Like Me*," *Critique: Studies in Contemporary Fiction* 63, no. 4 (2022): 418.

[22] John G. Cawelti, *Adventure, Mystery and Romance: Formula Stories as Art and Popular Culture* (Chicago: University of Chicago Press, 1976), 45f. It is worth noticing that Cawelti, following other scholars, defines melodrama first as a literary genre, in contrast to popular opinions which link it exclusively to cinematographic stories, admittedly better known nowadays.

[23] See Grażyna Stachówna, *Niedole miłowania: Ideologia i perswazja w melodramatach filmowych* (Kraków: Rabid, 2000), 20.

[24] See ibidem, 21f.

companion, and his breaking out of this role for whatever reason, even a noble one, cannot be tolerated. Though clearly morally superior, socially Adam the robot is inferior and his rebellion cannot succeed, even though the values that inform it may be officially professed by this world.

Adam's miserable end points, however, to one more ethical issue implied by the novel, namely the unclear legal status of intelligent machines and the lack of regulations concerning human-machine conduct. In the scene concluding the story the fictional character of Alan Turing thus reproaches the main protagonist: "My hope is that one day, what you did to Adam with a hammer will constitute a serious crime. Was it because you paid for him? Was that your entitlement? ... You weren't simply smashing up your own toy, like a spoiled child. You didn't just negate an important argument for the rule of law. You tried to destroy a life. He was sentient. He had a self. How it's produced, wet neurons, microprocessors, DNA networks, it doesn't matter. Do you think we're alone with our special gift? Ask any dog owner. This was a good mind, better than yours or mine, I suspect. Here was a conscious existence and you did your best to wipe it out. I rather think I despise you for that."[25]

The main protagonist's conduct shows his arrogant attitude towards the robot and the latter's status of an object which can be bought, sold or destroyed. Yet the fact that it is intelligent, sentient and conscious, that it has a complex emotional and moral life complicates this status considerably. Projecting a futuristic scenario, the novel points out quite clearly that human assumptions have to be reflected on and revised so as to accommodate intelligent machines into the common human-machine world of social and legal relations as treating them as mere appliances will be vastly inadequate. This is in keeping with the postulates articulated by AI scientists. Stuart Russell and Peter Norvig observe: "If robots become conscious, then to treat them as mere 'machines' (e.g., to take them apart) might be immoral. Robots also must themselves act morally—we would need to programme them with a theory of what is right and wrong. Science fiction writers have addressed the issue of robot rights and responsibilities, starting with Isaac Asimov (1942).... The stories (and the movies) convince one of the need for a civil rights movement for robots."[26]

On a final reflection, then, McEwan's novel not merely familiarises its readers with the so far imaginary situations of living with artificial intelligence and androids. More importantly, it invites them to reflect upon both their nature as sentient and conscious beings and on their status as valid members of an extended human-machine society. As McEwan comments, "if a machine seems like a human or you can't tell the difference, then you'd jolly well bet-

[25] McEwan, *Machines Like Me*, part X, loc. 3784–90.

[26] Russell and Norvig, *Artificial Intelligence*, 964.

ter start thinking whether it has responsibilities and rights and all the rest."[27] Thus, representing an imaginary scenario, the novel both draws attention to legal, social and ethical issues resulting from the co-existence of people and advanced machines, and points to the basic imperfection and moral fallibility of the human world, which ironically may turn out to be a major obstacle in the process of creating a common human-machine community.

FRIENDS LIKE KLARA

The convention of melodrama, the android protagonist, its poignant end and ethical considerations link McEwan's novel with the recent work by Nobel-Prize winning Kazuo Ishiguro, *Klara and the Sun* (2021).[28] More focused than *Machines Like Me*, set in a near-future America and narrated by the android protagonist, the novel tells a story of Klara, an Artificial Friend, i.e., a humanoid robot designed to keep company and be a friend to lonely teenagers. The plot traces her beginnings in a store, where she waits to be bought by a willing future owner, her subsequent relationship with Josie and her family who take her, and then her "slow fade" on a garbage heap where she is waiting again, this time for her imminent death. Tracing the trajectory of a robot's life the story, in a genuinely melodramatic fashion, both imagines the vicissitudes of android existence and shows the features of human society and relationships.

The eponymous Klara makes a perfect AF: devoted, loyal and safe-effacing, she seems to have no private life, ambitions or goals. Her existence is filled with serving and her intelligence, abilities and talents help her constantly improve her performance; she seems to have a special gift of observation, which she develops not for her own sake but for that of her human friends. Like the memorable butler Stevens from Ishiguro's *Remains of the Day*[29], Klara devotes her whole life to serving the teenager who chose her. Yet she also exhibits more than standard features, associated rather with humans: she has dignity, manifested, e.g., in a scene in which she refuses to serve as a toy to other teenagers, rejects taunting and gets offended when bullied by irresponsible humans. She seems not just to observe but even perhaps add another law to Asimov's famous three; her fourth commandment might be 'a robot will try to comprehend a human being', which she incessantly tries to do.

[27] Tim Adams, "Ian McEwan: 'Who's Going to Write the Algorithm for the Little White Lie?'" *The Guardian*, April 14, 2019, https://www.theguardian.com/books/2019/apr/14/ian-mcewan-interview-machines-like-me-artificial-intelligence.

[28] See Kazuo Ishiguro, *Klara and the Sun* (London: Faber & Faber, 2021).

[29] See idem, *The Remains of the Day* (London: Faber & Faber, 1989).

More interestingly still, Klara seems to be a religious robot: powered by solar batteries, she observes a sort of an AF solar cult in which the Sun features as a benevolent God granting life and spiritual and physical nourishment to the world. Klara prays to the Sun and worships him, ever thankful for the gift of life; she also believes in special grace that the Sun may choose to offer which is able to save life and help those in need. This is what she hopes for when her teenager friend Josie becomes deadly ill as a side effect of her medical procedure of 'lifting', i.e., advanced genetic editing which is supposed to increase intellectual potential of children and allow them to enter a higher caste of citizens. Seeing Josie on the verge of dying, Klara bargains with the Sun to grant Josie health and life. In a gesture of self-destruction, she sacrifices her own wellbeing depriving herself of a part of a fluid vital for her proper functioning as a gift to the Sun which may save Josie in return. Miraculously, the sacrifice works and Josie's health becomes restored with Klara weaker but happy that her prayers and sacrifice were accepted and reciprocated. Her religiosity and selfless actions stand in a sharp contrast to the rest of the novelistic world populated by characters which seem not to exhibit any sort of religious or spiritual traits and who are focused on material manifestations of status, wealth and security. Religion and sacrifice seem to be so out of place in this world that they are not even a subject to discuss. Thus, it is her religious belief and selfless love for a human being, rather than just her status of a robot, that make Klara exceptional in this story: in a truly evangelical way she loves God, loves her friend, and is ready to sacrifice her life for her.

Klara's "evangelical" spirit, which may perhaps be described as a quasi-Christian ethics of altruism, becomes visible once again in the last section of the novel when, old and no longer needed, she is discarded in a garbage lot, lonely reminiscing about her life. As Robert C. Abrams observes, her "slow fade" (as her dying is euphemistically referred to) may be seen as "unexpected object lessons in human aging and death."[30] Fully conscious and sentient, Klara accepts her fate with dignity and patience; she re-experiences her past trying to locate important moments and episodes, convincing herself that her life had sense and value. In this psychologically healing process, she prepares herself for her death which she is to face alone. Robert C. Abrams observes: "The robotic structure of Klara's thinking does nothing to obscure the fact that her exemplary personal growth and maturation mirror the human developmental journey at its best. It has been an evolution from concrete observation to com-

[30] Robert C. Abrams, "*Klara and the Sun*: Kazuo Ishiguro's New Model for 'Completion' at the End of Life," *Journal of the American Geriatrics Society* 70 (2022): 636.

plex emotion, concluding with a pair of crowning achievements, the ability to love selflessly and die with integrity."[31]

Klara's poignant end, however, though she accepts it with dignity and integrity, reflects once again on the human world in which agreement to one's fate is rather absent. Genetic 'lifting' and other medical procedures demonstrate negating rather than embracing of natural processes; death is a calamity which can be attenuated by refusing to accept it, e.g., trying to create mechanic "continuations" of deceased relatives. In contrast to how they treat their own death, however, novelistic humans are not particularly sentimental about other creatures' end. Klara's life of good service ends when she is no longer needed; with Josie leaving for college, an Artificial Friend becomes useless and is unceremoniously disposed of in the junkyard. Her friendship and sacrifice go vastly unreciprocated, her love unrequited: though treated decently while in service, ultimately Klara is perceived as an object, a mechanical toy rather than a sentient individual. Thus, similar to McEwan's story, in *Klara and the Sun* it is the robot that emerges as a noble and ethical creature in a novel full of rather selfish and cruel humans. Like in the previous novel, too, the melodramatic formula—a story of an impossible love and devotion of a robot towards a human being—exposes the rules of the fictional world. And like in the previous novel, these rules seem to be profoundly egoistic and materialistic: this is the world of no religion and little ethics, with self-obsessed humans ruthlessly pursuing their goals, hurting other creatures with no consideration for their wellbeing. Thus, apart from being an attempt at imagining a possible future scenario, the novel is also a rather bitter diagnosis of the ethical state of humanity at present.

Similar to *Machines Like Me*, Ishiguro's novel employs the first-person narration, yet this time with the eponymous robot-character Klara acting as a character-narrator. Very much like in Ishiguro's previous novel *Never Let Me Go*,[32] which likewise introduced an "inhuman" narrator, this structure of narration offers a possibility to see the world with the narrator's eyes and from her perspective, too. Characteristic of rather experimental narratives, as Jan Alber, Henrik Skov Nielsen and Brian Richardson argue the so-called "unnatural" or non-human narrators and unexpected points of view function as a method to transgress automatised conventions and equally automatised viewpoints.[33] In the case of Klara this perspective may strike one as surprisingly childlike and realistic, focused on mundane details, with precise descriptions rendering both

[31] Ibidem, 637.

[32] See Kazuo Ishiguro, *Never Let Me Go* (London: Faber & Faber, 2005).

[33] See Jan Alber, Henrik Skov Nielsen, and Brian Richardson, "Unnatural Voices, Minds, and Narration," in: *The Routledge Companion to Experimental Literature*, ed. Joe Bray, Alison Gibbons and Brian McHale (London: Routledge, 2012), 353.

her acute sense of observation and her lack of experience in the human world. Yet, as is perhaps already typical of Ishiguro's narrators, this simplicity and focus may be misleading: Klara as a narrator seems to reveal as much as she conceals and although her monologues include hardly any straightforward information about her emotions or opinions, occasional narrative hints help infer more complex processes taking place in her mind. Via the contrast between the childlike and sometimes naïve observations of the narrator which takes everything in good faith and the less than pleasant facts that she narrates the novel creates an ironic and poignant distance towards the world it describes. This deceptively simple and yet nuanced structure of narration allows the novel to indirectly show a much less naïve portrayal of the world in which the narrator functions.

Additionally, choosing a robot as a narrator may perform one more function: that of familiarization and empathization. Empathy may be defined as a "process of feeling, perceiving and understanding of another person's psychological state"[34] possible due to the ability to put oneself in someone else's situation. A narrative that imaginatively presents such a situation may greatly assist the process of empathy formation by presenting details and perspectives so far unimagined. It has therefore the power to sensitize readers towards new and alien experiences. A robot-narrator of the novel may work towards empathization and bringing the readers closer to the rather exotic internal life of artificial intelligence. So does its embodiment in a human-like shape, which may seem a strange choice given the fact that most of the AI used at present are computer programmes and that the creation of humanoid robots may seem not only very difficult but also undesirable, considering the human prejudice against and fear of creatures that deceptively resemble them. Yet, the human-like embodiment may once again work towards inspiring empathy towards robots, drawing attention to the fact that despite their mechanical origins, they are essentially not very different from humans, being not only conscious and sentient but also undergoing the same processes of developing, performing, degenerating and finally dying. Both the humanoid shape and the narrative voice, then, not only assert the character of Klara as a fully developed individual; more importantly, they try to convince the readers that robots are creatures like them, deserving the same treatment and sympathy. And as Santiago Mejia and Dominique Nikolaidis conclude, thus constructed character "forces us to confront interesting ethical questions concerning our lives with intelligent machines: 'how do we relate to them?' [or] 'what kind of autonomy and dignity do they have?.'"[35]

[34] Józef Rembowski, *Empatia: Studium psychologiczne* (Warszawa: Wydawnictwo PWN, 1989), 69.

[35] Santiago Mejia and Dominique Nikolaidis, "Through New Eyes: Artificial Intelligence, Technological Unemployment, and Transhumanism in Kazuo Ishiguro's *Klara and the Sun*," *Journal of Business Ethics* 178 (2022): 304.

*

With their futuristic settings, grim visions of social life with technological progress doing little to alleviate inequalities and injustice, and tragic endings of robotic protagonists both novels seem to question the simplistic division into the technophilic or technophobic trends. Rather, constructing their narratives around sentient and self-conscious robots which inhabit a human world they defamiliarize the latter, exposing it as full of inconsistencies, injustices and moral ambivalence. Like the Martian from Craig Raine's poem, robots observe with puzzlement humans and their habits, becoming disoriented and disappointed.[36] Simultaneously, however, apart from defamiliarization, an opposite process seems to take place in the novels which one could perhaps describe as "familiarisation," i.e., making the readers familiar with and more sensitive towards creatures and characters so far alien and inscrutable. Moreover, as these creatures are clearly presented as ethically superior—although representing diverse ethical positions that may be roughly compared to Kantian deontology and Christian altruism—the texts indirectly seem to examine the ethics of the human world which emerges as hedonistic and self-centred. The fact that these creatures are machines rather than other, more traditional aliens may perhaps testify to the technophilic bend of the two works. Both novels imply quite strongly that the future machines not only will occupy a place in human societies but also that this place should perhaps be carefully rethought rather than unreflectively assumed. In the represented world of both McEwan's and Ishiguro's stories robots are situated between appliances, pet animals and servants, endowed with no status or rights, but precisely this position comes under the narrative scrutiny and is problematised in their plots. Both novels suggest that introducing sentient and self-conscious machines into a human world will have not just practical but also moral and legal implications and that traditional human ethics, especially this practised rather than just professed, will have to accommodate them, too. The novels' melodramatic structure reveals, then, not merely poignant and tragic stories of their main protagonists but also the larger structure of the fictional world and—by extension—the extra-textual world inhabited by the readers. This revelation might seem ironic as human beings are portrayed as rather inferior morally to perhaps childishly naïve yet ethically accomplished machines.

Despite robot protagonists, mildly technophilic attitudes and futuristic high-tech settings, both novels seem rather far removed from transhumanist optimism one could expect from thus characterised fiction. Rather than portraying a brave new human living to the utmost of his/her potential and making

[36] See Craig Raine, *A Martian Sends a Postcard Home* (Oxford: Oxford University Press, 1980).

the full use of advanced technologies, as transhumanists perhaps would wish,[37] McEwan's and Ishiguro's novels may come closer to posthumanist reflection on the common human and non-human co-existence. Posthumanism—which in its critical version may be broadly understood as a reflection that displaces the human from his privileged position and revises the anthropocentric paradigm of thought and science—posits a more equal relation between the human and non-human. In Michael Hauskeller's conceptualisation it "refuses to see humans as a superior species in the natural order, ontologically distinct from animals on the one hand, and machines on the other ... [and—B. K.] insists that the boundaries between the human and non-human are rather fluent and in fact have always been so."[38] To Hauskeller, exploding the dividing line between the human and non-human carries with it political implications as it leads to the flattening of hierarchies and binary oppositions that privilege one side (human) only; as he observes, "at the heart of post-humanism is clearly a liberationist ideal: the hoped-for redistribution of difference and identity is ultimately a redistribution of power."[39] The two novels may thus be interpreted as positing a vision of flat posthuman world and a less hierarchical posthuman society, in keeping with the tenets of critical posthumanism. "Posthumanism is a 'post'—according to Francesca Ferrando's definition—to the notion of the 'human,' located within the historical occurrence of 'humanism' (which was founded on hierarchical schemata), and in an uncritical acceptance of 'anthropocentrism,' founded upon another hierarchical construct based on speciesist assumptions. Both the notion of the 'human' and the historical occurrence of 'humanism,' have been sustained by reiterative formulations of symbolic 'others,' which have functioned as markers of the shifting borders of who and what would be considered 'human': non-Europeans, non-whites, women, queers, freaks, animals, and automata, among others, have historically represented such oppositional terms."[40]

Both *Machines Like Me* and *Klara and the Sun*, therefore may be interpreted as gestures pointing in the direction of inclusiveness and abandoning of hierarchies, inviting their readers to consider the possible equality—legal and social—of humans and machines. The title of McEwan's novel reads "machines like me" but the subtitle adds to it "and people like you," thus juxtaposing the two 'species' and showing them as—still—contrastive. Yet in a truly

[37] See Max More, "The Philosophy of Transhumanism," in *The Transhumanist Reader*, ed. Max More and Natasha Vita-More (Chichester: Wiley & Blackwell, 2013), 3–17.

[38] Michael Hauskeller, "Utopia in Trans- and Posthumanism," in *Post- and Transhumanism: An Introduction*, ed. Robert Ranisch and Stefan Lorenz Sorgner (Frankfurt-am-Main: Peter Lang, 2014), 104.

[39] Idem, *Mythologies of Transhumanism* (Cham: Palgrave Macmillan, 2016), 23.

[40] Francesca Ferrando, *Philosophical Posthumanism* (London: Bloomsbury, 2019), 24.

utopian spirit of a mental ‘what if’ experiment the novels project a fictional world in which perhaps the situation of the species equality does not yet take place but which strongly implies that it is at least imaginable. The scenario the two novels offer for consideration posits that artificial intelligence may become a normal part of the world with AI, being equipped with the features usually associated with humans, i.e. sensibility, intelligence, self-consciousness, dignity and morality, enjoying the legal and social status equal to humans. The texts seem to illustrate the proposition of James Moor who suggests considering a chance that “it’s possible that someday robots will be good ethical decision-makers [...] acting ethically on the basis of a moral understanding.”[41] Doing so, the two novels go a step beyond empathising the readers with AI and suggest reflection on the possible expansion of the world model to include other than human actors and thus familiarise the reading public with the idea that AI may become a part of life with equal position and rights to human beings.

BIBLIOGRAPHY / BIBLIOGRAFIA

Abrams, Robert C. “*Klara and the Sun*: Kazuo Ishiguro’s New Model for ‘Completion’ at the End of Life.” *Journal of the American Geriatrics Society* 70 (2022): 636–37.

Adams, Tim. “Ian McEwan: ‘Who’s Going to Write the Algorithm for the Little White Lie?.’” *The Guardian*, 14 April 2019. https://www.theguardian.com/books/2019/apr/14/ian-mcewan-interview-machines-like-me-artificial-intelligence.

Alber, Jan, Henrik Skov Nielsen, and Brian Richardson, “Unnatural Voices, Minds, and Narration.” In *The Routledge Companion to Experimental Literature*. Edited by Joe Bray, Alison Gibbons, and Brian McHale. London: Routledge, 2012.

Asimov, Isaac. “Runaround.” In *I, Robot*. New York: Doubleday, 1950.

Booch, Grady. “Don’t Fear Superintelligent AI”, TEDtalk. Youtube, March 13, 2017. https://www.youtube.com/watch?v=z0HsPBKfhoI.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2014.

Cawelti, John G. *Adventure, Mystery and Romance: Formula Stories as Art and Popular Culture* Chicago: University of Chicago Press, 1976.

Ferrando, Francesca. *Philosophical Posthumanism*. London: Bloomsbury, 2019.

Hauskeller, Michael. “Utopia in Trans- and Posthumanism.” In *Post- and Transhumanism: An Introduction*. Edited by Robert Ranisch and Stefan Lorenz Sorgner. Frankfurt-am-Main: Peter Lang, 2014.

———. *Mythologies of Transhumanism*. Cham: Palgrave Macmillan, 2016.

[41] James H. Moor, “Four Kinds of Ethical Robots,” *Philosophy Now*, no. 72 (2009), https://philosophynow.org/issues/72/Four_Kinds_of_Ethical_Robots.

Ishiguro, Kazuo. *Klara and the Sun*. London: Faber & Faber, 2021.

———. *Never Let Me Go.* London: Faber & Faber, 2005.

———. *The Remains of the Day*. London: Faber & Faber, 1989.

Kim, Tae Wan, "Flawed Like Us and the Starry Moral Law: Review of *Machines Like Me* by Ian McEwan." *Journal of Business Ethics* 170 (2021): 875–79.

Księżopolska, Irena. "Can Androids Write Science Fiction? Ian McEwan's *Machines Like Me*." *Critique: Studies in Contemporary Fiction* 63, no. 4 (2022): 414–29.

Kurzweil, Ray. "We Will Be More Fun, Sexier and More Creative." *Svenska Dagbladet.* December 19, 2019. https://www.svd.se/a/d437dfb6-179b-4046-930e-dcfdbf620643.

Landgrebe, Jobst, and Barry Smith. *Why Machines Will Never Rule the World: Artificial Intelligence without Fear*. New York: Routledge, 2023.

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

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* 178 (2022): 303–6.

de Miranda, Luis. *AI and Robotics*. London: Ivy Press, 2018.

Moor, James H. "Four Kinds of Ethical Robots." *Philosophy Now*, no. 72 (2009). https://philosophynow.org/issues/72/Four_Kinds_of_Ethical_Robots.

Moravec, Hans P. *Robot: Mere Machine to Transcendent Mind*. Oxford: Oxford University Press, 2000.

More, Max. "The Philosophy of Transhumanism." In *The Transhumanist Reader*. Edited by Max More and Natasha Vita-More. Chichester: Wiley & Blackwell, 2013.

Nicieja, Stankomir. "Revisiting Utopia: New Directions for Utopian Fiction in Margaret Atwood's *Oryx and Crake* and Kazuo Ishiguro's *Never Let Me Go*." In *Margins and Centres Reconsidered*. Edited by Barbara Klonowska and Zofia Kolbuszewska. Lublin: Towarzystwo Naukowe Katolickiego Uniwersytetu Jana Pawła II, 2008.

Przegalińska, Aleksandra. "Zrozumieć człowieka." *Academia* 1–2 (2019): 10–13.

Raine, Craig. *A Martian Sends a Postcard Home*. Oxford: Oxford University Press, 1980.

Rembowski, Józef. *Empatia: Studium psychologiczne*. Warszawa: Wydawnictwo PWN, 1989.

Russell, Stuart**,** and Peter Norvig, *Artificial Intelligence: A Modern Approach*. New Jersey: Pearson, 2003.

Stachówna, Grażyna. *Niedole miłowania: Ideologia i perswazja w melodramatach filmowych*. Kraków: Rabid, 2001.

Torczyńska, Monika. "Sztuczna inteligencja i jej społeczno-kulturowe implikacje w codziennym życiu." *Kultura i Historia* 36, no. 2 (2019): 106–26.

ABSTRACT / ABSTRAKT

Barbara KLONOWSKA, Ethical Machines: Representations of Artificial Intelligence in Ian McEwans's *Machines Like Me* and Kazuo Ishiguro's *Klara and the Sun*

DOI 10.12887/36-2023-4-144-08

Artificial intelligence provokes contradictory reactions that range from enthusiasm and fear and that may be generalised as instances of technophobia and technophilia. AI is also a perennial theme of numerous fictional narratives which seem especially important as, due to their large audiences, images and stories representing AI enter popular debates. Two recent novels by eminent British authors, Ian McEwan's *Machines Like Me* (2019) and Kazuo Ishiguro's *Klara and the Sun* (2021), in either alternative past or futuristic settings also take up the issue of AI and its possible ramifications. The article argues that the two works represent AI as both beneficial and potentially problematic, posing the question about the essence of humanity and the limits of AI, problematising the status of intelligent machines and familiarising readers with ethical and legal problems they bring; they also try to build empathy and sensitise the public towards creatures other than humans.

Keywords: AI, technophobia, self-consciousness, posthumanism

Contact: Department of English Literature and Culture, Institute of Literary Studies News, Faculty of Humanities, John Paul II University of Lublin, Al. Racławickie 14, 20-950 Lublin, Poland
E-mail: barbara.klonowska@kul.pl
ORCID: 0000-0001-8327-854X

Barbara KLONOWSKA – Etyczne maszyny. Reprezentacje sztucznej inteligencji w powieściach *Maszyny takie jak ja* Iana McEwana i *Klara i słońce* Kazuo Ishiguro

DOI 10.12887/36-2023-4-144-08

Sztuczna inteligencja prowokuje sprzeczne emocje entuzjazmu i strachu, które mogą być powiązane z szerszymi postawami technofobii lub technofilii. Jest ona też stałym tematem wielu filmów i powieści, co wydaje się szczególnie istotne, gdyż często to poprzez filmy adresowane do szerokiej widowni temat SI wkracza do powszechnej debaty. Dwie powieści wybitnych brytyjskich prozaików, *Maszyny takie jak ja* Iana McEwana i *Klara i słońce* Kazuo Ishiguro, jedna w scenerii futurystycznej, druga snując alternatywną historię, również podejmują ten temat z jego nieoczywistymi konsekwencjami. Artykuł dowodzi, że obie powieści ukazują SI jako jednocześnie pożyteczną i problematyczną, stawiając pytania o istotę człowieczeństwa i granice sztucznej inteligencji, problematyzując status inteligentnych maszyn i oswajając czytelników z etycznymi i prawnymi problemami z nimi związanymi. Próbują one też wzbudzić empatię i uwrażliwić czytelnika na los istot innych niż tylko ludzkie.

Słowa kluczowe: sztuczna inteligencja, technofobia, samoświadomość, posthumanizm

Kontakt: Instytut Literaturoznawstwa, Wydział Nauk Humanistycznych, Katolicki Uniwersytet Lubelski Jana Pawła II, Al. Racławickie 14, 20-950 Lublin
E-mail: barbara.klonowska@kul.pl
ORCID: 0000-0001-8327-854X