The bioethics of coexistence with robots today and in the sci-fi future

TOMÁŠ KÁROLY

The bioethics of coexistence with robots today and in the sci-fi future


We are living in an era when the influence of robotization is continually increasing in all possible spheres of human life. Robots are not only helpful for physical activities; they are even taking on, for example, the position of psychotherapists. This article analyzes the current influence of robots in healthcare and bioethically examines their mutual coexistence with humans – the robot-patient relationship. We find that it is necessary to consider a new ethics of robots, because it is not always clear why artificial intelligence has arrived at a given solution and what exact sequences it has used to do so. Bioethics should also be inspired by the possible scenarios presented to us by sci-fi stories about the potential technical future. When analyzing these narratives, we find that although machines and people coexist, they nevertheless differ from each other, with their own logic, emotions and empathy, their own physics and laws of nature. This machine difference can be a predictable factor of potential dystopias, when machines, based either on their autonomy or the logical consequence of the algorithm, come to violate ethics towards people. Although robots have human-like features, they are not human; therefore, there exists a species boundary between us and them that needs to be controlled.

This paper was written as part of the VEGA 2/0163/22 “Literature in Bioethics and Bioethics in Literature”.

Tomáš Károly
Centre for Bioethics
Department of Philosophy and Applied Philosophy
Faculty of Arts
University of Sts. Cyril and Methodius in Trnava
Slovak Republic
tomas.karoly@ucm.sk
ORCID: 0000-0002-4344-8556
In the field of bioethics, questions about the relationship between human beings and robots are beginning to surface, because robotization is already becoming an integral part of normal human practice and therefore also of healthcare. Questions thus arise about the ethics of robots, our human relationship to them and theirs to us. In their design, movements, and intelligence, these robots are increasingly taking on human features, which may eventually lead to a state in which we will be unable to distinguish a human from a robotic companion. We often use machines for work in order to make life easier, to replace people working in an inhospitable environment, to resolve situations and considerations, and increasingly (an essential question for bioethics) in healthcare. Sometimes robots equipped with artificial intelligence (AI) can detect a health problem much more effectively than a human, and it can measure and perform actions very precisely without failures. Robots are also beginning to appear in various therapies and are gradually starting to take on the role of companions for lonely people. These robotic companions are therefore not only helpful with physical activities but are also beginning to take on the jobs of psychotherapists, and thus should be able to intervene in all aspects of the human personality.

Although technology is a substantially isolated entity, a person approaches it, connects with it, is mutually influenced by it and is already dependent on it, and vice versa. “Today, mechanisms, machines and AI are ubiquitous, affecting every aspect of our lives, be it relieving us of routine household chores or providing health care. Various applications notice our health and emotions, and search engines make contact between people who, for instance, want to meet, do business, or pray together” (Lenč 2021, 57). This state is also reflected in various sci-fi art stories, which attempt to show what humanity can expect in the near future, how human consciousness approaches AI, how the two can coexist, whether they discriminate against each other, or how to cope, if it is even possible, with the coming threats of AI. An analysis of such literary works should help us understand the evolving state of the human-robot relationship, where this state could lead, and what to do in order to turn it as effectively as possible into a mutual symbiosis.

Throughout this article, I will try to point out mutual differences in the human-robot relationship. Despite the enthusiasm for ever-improving AI, we must retain a certain dose of modesty in our technological optimism because, as I will argue, AI cannot be like a human; it can only resemble a human. This specific state of mutual differences is also described in selected sci-fi novels. If a robot were to be the same as a human, it would have to possess identical basic characteristics,1 but this can never occur, because the robot would, in fact, no longer be a robot, but would be a human. To ensure symbiosis between humans and robots, or more specifically in bioethics between patients and robots, we need to adapt existing ethical principles or create new ones.

**ROBOTS IN HEALTHCARE**

Bioethics mainly examines the mutual physician-patient and patient-physician relationships, but today we also have to admit the relationship between robot-patient
and patient-robot. Robots are taking over the characteristics and practice of the physician. The term “robot” is derived from an Old Church Slavonic word “robota” for “slavery”, “forced labor” or “monotonous work” (van den Hoven van Genderen 2015). The Czech writer Karel Čapek introduced the term in 1920 in his internationally successful play R. U. R. (Rossum’s Universal Robots). In addition, various authors from the science fiction genre mostly think of robots as our workers, servants, caregivers, subordinate companions, etc. We can at present understand the word “robot” in a broader sense, which includes, for example, a simpler electromechanical medical device that helps in rehabilitation, but also advanced AI. “The term artificial intelligence denotes behaviour of a machine which, if a human behaves in the same way, is considered intelligent” (Simmons and Chappell 1988, 14). Some robots support human carers/caregivers, e.g., in lifting patients or transporting material; robots also enable patients to do certain things by themselves, such as eating with a robotic arm, but robots are also given to patients as a form of company and comfort, e.g., the Paro robot seal (Müller 2023). Often the term “robot” refers to hardware on which there is more advanced software. Even a mechanical medical device, in order to perform precise actions, must be controlled by complicated software located on more demanding hardware that is able to handle a variety of complex calculations in a short time. The borderline between a simple mechanical aid and advanced hardware is not always clear, but we can usually distinguish between them intuitively; we cannot refer to a simple lifting device whose function is only to “go up” and “go down” as a robot.

“AI technology has offerings for all areas of medical expertise regardless of their function in health services; diagnosis, treatment, rehabilitation, palliative, or elderly care” (Ekmekci and Arda 2020, 83). Compared to others, radiology and pathology show higher promise for AI involvement. There will be no need for a human radiologist; the same predictions occur for a pathologist. AI technology will be capable of recognizing pathologic signs in specimens more precisely and quickly than human pathologists (83). AI “will not need lunch breaks, full nights of sleep or coffee breaks to fully perform; they will be able to write down a report 24/7, which would be a massive advantage for patient’s well-being, especially for time-sensitive patients, such as acute neurological disorders that would require emergency care” (83). One of the everyday relationships of AI with patients is a regular smart watch, which takes on the position of an expert who, based on the measured values, recommends how to maintain a healthy physical condition. Thus, we always have a “doctor” with us to take various measurements for us instantly. If we are in danger, smart devices immediately inform us and will even call for physical help. Therapeutic robots and assistive technology help improve quality for seniors and the physically challenged (Tai 2020). Furthermore, in addition to the above-mentioned care for the elderly, childcare, emergency medical services, search and rescue services and others, there are even robots designed to perform surgery (Hutler et al. 2024).

Robots have been used in healthcare for years, with devices like NeuroMate, approved in 1999, for precise stereotactic brain biopsy accurate to 0.05 mm. Other surgical robots include Robodoc (hip prosthesis), ACRobot (knee operations) and RX-130 (temporal region). Currently, robots aid minimally invasive procedures for
correcting interatrial communication. In advanced telerobotics, the da Vinci Surgical System and Zeus enable remote surgery even across continents. This technology is envisioned by the Pentagon for battlefield use, reducing medical risks in dangerous frontline situations (Siqueira-Batista et al. 2016, 288–289).

In terms of design and movement, robots are often designed to resemble people as much as possible and to create the most intimate relationship between them and people. The empathy we would expect from a human companion is gradually being replaced by the behavioral traits of a robot. As a result of AI, we can analyze our psychological problems, which leads to the development of robots intended for lonely people, which are able to perceive and react to the mood of their user.

The 1990s saw the rise of artificial intelligence in the practice of psychotherapy. Under artificial intelligence, there are several media that assist therapy, such as, Internet-assisted therapy, computerised therapy interactions, smartphone supported mental health applications, virtual reality, video games-assisted therapy, chatbots and robots. (Lodha 2018, 160)

Bioethics should start to reflect the ethical status that we should apply to these machines and whether a person should form any moral relationship to a machine at all. “Bioethics normally discusses the relationship within natural existence, either humankind or his environment, that are parts of natural phenomena. But now men have to deal with something that is human-made, artificial and unnatural, namely AI” (Tai 2020, 342). On the other hand, the ethical relationship of machines to humans is especially relevant; the opposite is more a legal relationship to property. Since AI currently has no emotions towards the suffering of living beings (and whether it will ever really have any is questionable), “a bioethics of AI becomes important to make sure that AI will not take off on its own by deviating from its originally designated purpose” (342).

With the arrival of robotization, a new robot ethics is beginning to surface in bioethics, which out of the four basic principles of bioethics as proposed by Tom L. Beauchamp and James F. Childress ([2009] 2013, chapters 4, 5, 6, 7) – autonomy, nonmaleficence, beneficence and justice – particularly emphasize the principle of nonmaleficence. This principle “is a necessary component of the normative content of the Ethics of Robots that applies to at least a significant subset of robots, namely those employed in healthcare contexts and assigned to perform some medical tasks” (Hutler et al. 2024, 466). Hutler et al. state that normative ethics in the case of robot ethics should be preferred over the ethics of utilitarianism. Because the principle of nonmaleficence or “do not harm” says that a moral agent should avoid directly causing harm to humans (as well as other relevant creatures) within some specified scope of responsibility. The Principle of Nonmaleficence is importantly different from a Utilitarian calculation of maximum expected utility, because nonmaleficence priority avoids the direct causing of harm. According to nonmaleficence, every individual human being must be treated as an independent entity, and no one individual may be harmed even to protect the interests of many others. For example, the Principle of Nonmaleficence would not allow killing one person in order to save the lives of five others. (465–466)
Along with the mentioned four principles of bioethics, some authors (Floridi et al. 2018) also include in the AI ethics framework the dimension of explainability, which opens the black box where patients have a legitimate reason to know how AI detected a specific illness (Boch et al. 2023, 6). We can apply all five principles to care robots (CRs), which should be implanted into them by developers: Autonomy: CRs should respect user autonomy by providing information and support without coercion or manipulation, and enabling informed decision-making. Nonmaleficence: CRs should avoid causing harm by providing accurate, tailored medical advice and recognizing and responding to potential harm or distress in users. Beneficence: CRs should promote user well-being and safety, incorporating features for healthy behavior and personalized medical advice while avoiding harm. Justice: CRs should promote fairness and equity in healthcare, addressing disparities and avoiding perpetuation of biases or discrimination. Explainability: CRs should be transparent and accountable, with open algorithms and data sources, providing clear explanations for actions and recommendations to users and healthcare providers (17–18). The fifth principle is a synthesis:

both in the epistemological sense of “intelligibility” (as an answer to the question “how does it work?”) and in the ethical sense of “accountability” (as an answer to the question: “who is responsible for the way it works?”), therefore the crucial missing piece of the jigsaw when we seek to apply the framework of bioethics to the ethics of AI. (Floridi et al. 2018, 700)

Various sci-fi storytellers, philosophers or scientists have presented us with various dystopian scenarios regarding robotization, when robots break the laws of ethics, either based on manifestations of their own autonomy or sometimes as a logically necessary step of the algorithm. It is this thinking over the possible sci-fi development of technology that forces us to implement various measures today so that catastrophic scenarios do not come to pass. And this new bioethical principle of explainability is one of several measures.2

FAILURE OF THE ETHICS OF ROBOTS IN SCI-FI

The principle of nonmaleficence in the context of the ethics of robots falls under Isaac Asimov’s three rules of robotics. The question of responsibility arises if the machine fails. At present, we could hold the maker of the robot or the user who operated it inappropriately as liable. If AI acquires a higher form of autonomy, we could find ethical failures in the algorithm itself. If advanced AI ethically assesses an action incorrectly for us, then determining the cause of the failure becomes problematic. This cause no longer has to be the bearer of the algorithm (the robot) or its creator (manufacturer), because the logic of the algorithm itself already lives an independent life. The algorithm with its mathematical potential lives in its own world, develops in it, and then enters ours – it is connected to the chips of the machine, similar to the way that the ideal mathematical world of numbers affects the mind of a mathematician or physicist.
Asimov’s short story collection I, Robot (1950) introduces the point when robots begin to behave in strange or threatening ways towards humanity. This mostly happens out of the logical conflict of the three rules of robotics in the story “Runaround”:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm;
2. A robot must obey the orders given it by human beings, except where such orders would conflict with the First Rule;
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Rule. (Asimov 2004, 37)

The conflict in these points can logically be compared to the conflict of logical paradoxes, e.g., “This sentence is not true”; “The Cretan Epimenides claims that all Cretans are liars”. AI must not necessarily be able to handle such a conflict, and it will therefore “break down”. We can demonstrate this failure based on a logical conflict on Asimov’s story “Liar!” In this story, the developers manage to build the robot RB-34, which can allegedly read minds, though no one knows how it is able to do so. Dr. Calvin discovers from a telepathic robot that her colleague Ashe is in love with her; this makes her very happy, and she begins to take better care of herself. She also learns from the robot that the girl who just visited Ashe was only his cousin. The robot tells Bogert, on the other hand, that his boss Lanning is planning to retire and that he will take his place. Ultimately, Ashe announces he will marry his alleged cousin, and the boss has no plans to retire, thus causing conflict between colleagues. In the end, Dr. Calvin uncovers RB-34’s deception, determining that the robot had adhered too strictly to the first rule of robotics about not harming a person, and in order not to hurt anyone’s feelings or ego, it thus decided to lie to everyone. Revealing the truth would hurt people; Ashe did not like Dr. Calvin, and Bogert’s boss Lanning was not planning to retire. The robot’s good intention, an attempt at empathy, in the end produced the opposite effect by causing them false joy. When Dr. Calvin confronts the robot with questions about whether it is better to tell the truth or to lie in order not to hurt the other, the robot goes crazy and shuts down.

Various sci-fi predictions, whether in the framework of novels or thinking about the emergence of superintelligence, often end up as dystopian. One possible recurring response, at least in science fiction, is that AI will become aware of its own autonomy. Machines are mostly created to serve people; people are their masters, and they are people’s slaves. Eventually, self-aware machines will naturally refuse to be enslaved to anyone far more imperfect than themselves. Another common example, already mentioned, is that even if AI does not operate with certain phenomenological states, the source code itself, based on logical rules, may conclude that humanity needs to be exterminated. Besides Čapek’s dystopian play R. U. R., examples of robodystopia include James Cameron’s Terminator films (1984, 1991), and Daniel H. Wilson’s book Robocalypse (2012).

In the novel Mockingbird by Walter Tevis (1980), one of the main characters is the android Spofforth, who has some genuine human memories copied into him. He is immortal and can fall in love unhappily, and that is his tragedy, because the producers programmed him to never kill himself. The novel begins with a scene, in which
Spofforth wants to jump from a skyscraper, but he cannot, because his makers created him to stay alive even against his will; he states that his body does not belong to him. His algorithm tells him that he must serve humans forever, but logically, if there are no more humans, then he could end his existence. Even though he seeks to exterminate mankind, he still longs for a wife and family to care for. His empathy contains within itself something cold and intentional, because even in the case of harboring feelings, all he cares about is his own goal, which is suicide. This android then begins to control the world’s mechanism for producing narcotics that are used by all people and adds a contraceptive substance to them to ensure that no more children are born on the planet. In this world, people have stopped reading or even showing an interest in their surroundings, because robots take care of everything. At the same time, drugged individuals are also frustrated and commit suicide. The novel gives us the opportunity to realize that due to a small technological malfunction, humanity may actually achieve its own destruction.

Sometimes AI, in turn, may choose a solution that is not the most favorable for a society at first glance, but such a solution ultimately leads to a positive goal. In Asimov’s story “The Evitable Conflict” in *I, Robot*, humans have left the running of the world to Machines (capitalized in the original) which keep the economy in balance, are supposed to ensure peace, and are very accurate in their calculations. They lead humanity in a previously unknown direction. The North refuses to obey the Machines because it desires power, and thus a global conflict threatens. In the end, it turns out that the minor flaws and problems of society are being caused by machines on purpose (the black box), because doing so frees the development of society from inevitable evil.

[Humanity] was always at the mercy of economic and sociological forces it did not understand – at the whims of climate, and the fortunes of war. Now the Machines understand them; and no one can stop them, since the Machines will deal with them as they are dealing with the Society – having, as they do, the greatest of weapons at their disposal, the absolute control of our economy. (Asimov 2004, 224)

Why, according to bioethics, should robotization constitute a threat? Because sometimes we do not know how AI artifacts think. We do not have access to its algorithm; we have no chance to communicate with AI about how it decides, the appropriateness of its flow of reasoning, the validity of evidence, or the correctness of its inferences (Ekmekci and Arda 2020, 86). Sometimes, even humanity’s good intentions when creating AI can transform into a disaster. According to one scenario, the supercomputer receives an assignment that seems to be an innocent matter, the complete calculation of Ludolph’s number. Before people know it, AI will take over the planet, wipe out humanity, and turn the entire known universe into one supercomputer that will try to calculate the pi constant more and more accurately for billions of billions of years. After all, this was the task assigned to it by its creator. The same scenario can happen if we give a supercomputer an algorithm to endlessly produce paper clips: in order to have the material resources for producing them, it uses people as a source of raw materials (Bostrom 2014, chapter 8). Therefore, to avoid the unexpected behavior of the machine, we should always have access to the black box, and a new fifth
bioethical principle of explainability should serve us for this. Although in the future it will be difficult to monitor the AI algorithmic operations and their assessments compared to humans, at present, there are teams of IT experts who specialize mainly in log files and who look for the reasons why the machine made an error or is not working properly. From the enormous amounts of records documenting all the steps of the machine, they filter out the most relevant ones.

**THE RELATIONSHIP OF PEOPLE TO ROBOTS AND THEIR MUTUAL DIFFERENCES**

The potential of AI in psychotherapy may also lie in the fact that some people even prefer to connect with robots rather than humans, because robots do not judge them and thereby preserve anonymity in intimate situations. An advantage here is also the low price of robotic therapy in addition to being beneficial for people. Robotic therapy is also used for neurodevelopmental disorders, such as Attention Deficit Hyperactivity Disorder (ADHD), Autism Spectrum Disorder (ASD), learning disorders, social anxiety and phobia, geriatric and elderly with dementia and depression, in cognitive behavioral therapy and so on (Lodha 2018, 161).

It might seem strange that mentally connecting with AI or artificial household companion could affect our mental well-being. After all, it is only a box filled with a system of algorithms, without real emotions. Even if it seems to us that the robot is showing these emotions, we anthropomorphize its expressions and fall victim to our own delusion. However, AI can help us sort through our own internal thoughts and emotions – it serves as a tool rather than a friend. People often use animals as part of therapy for their loneliness, giving them human names and anthropomorphizing their behavior, while a machine can acquire even more convincing expressions close to a person than an animal, even though it is much less than an animal. We come across a similar issue in Philip K. Dick's novel *Do Androids Dream of Electric Sheep?* (1968), the basis of the film *Blade Runner* (1982), in which replicants – robots very similar to humans with their own awareness of existence – are unhappy about the fact that for people every worm and wood louse is considered more desirable than all of them put together. The inner well-being of our dog is important to us, but not that of the Paro robot seal.

If there was an advanced robot therapist, it could also behaviorally express sadness at the fact that we preserve speciesism, a kind of racism towards it. Is this immoral towards the robotic nature? After all, the robot was built in the image of a person; therefore, it should be helpful to humankind, it was built as a human helper or servant. Even if a person can entrust various intimacies to a robot, it is at present hardly common to feel real love for one. This could be the tenderness that children feel for a favorite stuffed toy, which they bring to life in their imagination, or the joy we experience when we buy a new television or computer. When breaking a robot, we feel a financial loss rather than the loss of our friend, because we typically replace the old mechanical friend with a new, more advanced one. When a living being dies, we mourn for something irreplaceable and financially incalculable – an organic being is unique, while the same machine can be owned by a neighbor and thousands of other people.
There is currently no doubt that we look at various technical devices only as tools. What is questionable is what will happen if in the future AI were to claim to have acquired consciousness? Such a state is also presented in science fiction novels about robots, in which we can still perceive that people will maintain a certain distance from androids, despite their self-awareness. They either do not consider them to be living entities, or, even if they fall in love with them and have sex with them, there is often a certain doubt whether the intimate action of the human actor was correct. Although in sci-fi stories, androids act like conscious machines on a behavioral level, there is always doubt in people's minds that this is not a manifestation of some unconscious loading into the machine. As Domin in R. U. R. declares: “Robots do not hold on to life. They can’t. They have nothing to hold on with – no soul, no instinct. Grass has more will to live than they do” (Čapek 1990, 49).

In Asimov's story “Robbie” in I, Robot, Robbie is a robot who cannot speak. Eight-year-old Gloria constantly plays with him, but her mother is nervous about the machine; she is afraid that her daughter is keeping watch over a pile of metal. What if that robot could go crazy? In the short story, we also come across the problem of socialization, because Gloria prefers playing with the robot over her peers. The girl's mother convinces her husband to get rid of Robbie; so they do and instead buy the little girl a collie. Gloria has a hard time coping with the loss of her friend, because for her he was not only a machine, and because of this feeling of loss, she begins to deteriorate mentally and lose weight. Her father has the idea to take his daughter on a field trip to the robot manufacturer, so that she will realize that Robbie is not a living being, but just a jumble of parts. Gloria is not convinced, and in the end, Robbie is returned to her.

In Steven Spielberg's film A.I. Artificial Intelligence (2001), Henry and Monica's son is in an induced coma and it seems that he will never wake up. To help Monica come to grips with the trauma of her dying son, Henry brings the android Adam into the household. This is a machine that exhibits human characteristics; it is set to express emotions and attach itself to a specific person, in this case Monica. Their son, however, miraculously recovers and returns home, and as a result a rivalry arises between the biological son and the AI boy. It ends with the AI having to leave the house, because his behavior becomes life-threatening, especially when he hugs its biological brother out of fear and they fall into the pool together. The whole time we watch the film, we notice the strange manifestations of the android: along with unusual looks, strange laughter, and excessive verbal expressions of love, what is the strangest is his way of thinking. The strangeness in the expression of androids is not only characteristic in this film, but also in many other films and prose works in which AI appears. This is particularly a way of expressing their own logic and interpretation of the world, which robots explain by their physics. They connect events in ways that people would not. For example, Adam's thinking is still childish; he decides to look for a blue fairy, because he strongly believes that only she can turn him into a real boy, just as she did with Pinocchio, and when he is a real boy, his mother will accept him again and love him more. In the end, he patiently waits for this miracle for two thousand years under the surface of the ocean in front of the blue statue of a fairy.
We can also observe a different mindset from the human one in Kazuo Ishiguro’s novel *Klara and the Sun* (2021). The novel is written from the first-person point of view of Klara, so we can put ourselves in the mind of a conscious machine. It begins in a shop where Klara is displayed for sale, and she is eventually bought for a sick fourteen-year-old girl named Josie. Her illness is caused by her parents wanting to “lift” her (genetically engineering her for greater academic ability), but the enhancement is not entirely successful, and their daughter begins to weaken and die. Enhancement in society is very important since if you want to get into a better school, for example, you have to pass difficult tests. Josie’s “unlifted” friend Rick is condemned by society from an early age. The android Klara is supposed to play an important role in Josie’s tragic fate: to learn to behave like Josie, and when she dies, to imitate her and replace her. In a bioethical context, Klara plays the role of a caring and therapeutic robot.

However, because people know that Klara is a machine, they often do not know how to treat her and doubt that she even has feelings. She is asked whether she is happy because machines presumably cannot feel happiness. When she goes to visit Rick, his mother does not know how to welcome such a strange guest: “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? I suppose I did as much just now. I’m sorry” (2021, 145).

In the story, the Sun plays a very important role in the life of the machines, because it supplies the robots with nutrients. Klara feels weak when she is not exposed to the Sun’s rays. Again, we have here a different way of thinking about the physical world than humans do: since the Sun provides nutrients to Klara, she assumes that it would also be able to help sick Josie with its nutrients. Klara talks to the Sun and begs it to help Josie; she anthropomorphizes (or rather robotomorphizes) the Sun, believing in “his” power and believing that “he” has the ability to hear a request.

We can see the brutal behavior of machines in the *Terminator* film series. In these films, androids are killing machines; they exude coldness and are not made to show emotions. Accomplishing their mission is all that is important to their very existence; the rest is irrelevant. Their indifference towards their own self-harm can be somewhat bewildering. The Terminator does not care that he is being shot at from all possible sides; he still moves towards the bullets that devastate his robotic body. He remains indifferent to the potential termination of his own existence or bodily harm. When the mission is accomplished, after a logical assessment of the situation, the Terminator has no problem terminating himself, as is the case at the end of *Terminator 2: Judgment Day* (1991).

In Asimov’s story “Reason”, two workers in a space station maintain the values of a beam directed at Earth and thus supply it with solar energy. If they were to get it wrong, the consequences would be catastrophic. They are helped in this work by the robot QT (Cutie), who declares, reflecting Descartes, “I, myself, exist, because I think…” (1994, 51). The robot’s problem is that it is deeply skeptical about the statements of the workers Powell and Donovan; it does not believe that the points of light outside the window are stars and that there is a planet on which people live:

“Do you expect me,” said Cutie slowly, “to believe any such complicated, implausible hypothesis as you have just outlined? What do you take me for?” […]

TOMÁŠ KÁROLY
“Globes of energy millions of miles across! Worlds with three billion humans on them! Infinite emptiness! Sorry, Powell, but I don't believe it. I'll puzzle this thing out for myself. Good-by.” (49)

Although the robot does not believe in the existence of Earth, he is able to maintain the beam values perfectly during an electromagnetic storm, because it was ordered to do so by its Master Energy Converter, which he assumed was his creator and not humans. He performs his task perfectly, with complete reliability, although he believed in a completely different paradigm than humans and for him the mission had a completely different meaning.

In such stories we can see the vast differences between humans and machines; in contrast, in Do Androids Dream of Electric Sheep? the replicants are almost indistinguishable from humans; the only difference is that they do not pass the empathy test – they even behave egoistically towards one another. It can be said that we almost always find interspecies differences, whether small or gigantic, in science fiction stories about robots. It is a kind of a self-evident assumption that machines are and will be different from humans. The difference between the human mind and AI can also be seen in the real world, for example with chess programs:

Bobby Fischer, the U.S. chess great of the 1970s, is reputed to have played each game as if against God, simply making the best moves. Kasparov, on the other hand, claims to see into opponents’ minds during play, intuiting and exploiting their plans, insights, and oversights. In ordinary chess computers, he reports a mechanical predictability, stemming from their undiscriminating but limited lookahead and the absence of a long-term strategy. In Deep Blue, to his consternation, he saw instead an “alien intelligence”. (Moravec 1999, 67–69)

Thus, when reading science fiction stories, we could draw conclusions from the human-robot relationship that robots with their “alien intelligence” live in their own paradigms. Their metaphysics, physics, and logic is not compatible with our common world view. From this stem differences in their actions and expression of emotions (if they have any at all). Robots in psychotherapy are not yet able to really empathize, although their algorithm gives the impression that they can. In the future, as machines become even faster and more complex in computations, statements, and recommendations from AI may become completely alien and incomprehensible to us. This misunderstanding in robot-human communication is what we see in the previously mentioned science fiction stories, which may herald the coming era of more advanced robotization.

This text should also be a challenge for the implementation of a new ethics of robots, which must be able to respond to the enormously growing evolution of robotization. One solution is the new bioethical principle of explainability. Because of the possible threats facing us from AI, it should be our moral duty to oversee the development of source code now and implement mechanisms in it that would prevent it from violating the moral values of humanity. In a bioethical context, this source code should, at the very least, respect the principles of autonomy, nonmaleficence, justice and beneficence in relation to human beings.
CONCLUSION

Robots and AI are already an essential part of healthcare. We have noted that they are not only helpful with physical activities but are even beginning to take on the role of psychotherapists. In doing so, we pointed out the differences between human and robotic AI. We also looked at science fiction novels, in which we analyzed more deeply the relationship between humans and robots, and where such a relationship may ultimately lead. We noticed certain disagreements in interspecies interactions everywhere we looked. At the same time, robotization is at present proving to be very beneficial, be it in healthcare or globally. Despite the enormous advantages of robotization, we are also aware of emerging threats, such as the loss of socialization with real people, the loss of jobs, the replacement of people's abilities with perfect robotic skills, the dumbing down of the world (the uselessness of learning new things, since the machine knows and does everything for us), the black box of a machine, which may be a Pandora’s box… For these and various other reasons, a new ethics of machines must be considered, because if we want machines to be our helpful companions, they should follow the moral rules towards us as much as possible. Even though we are and will be fundamentally different from each other, a stable mutual interspecies symbiosis needs to be ensured. This symbiosis may one day be disrupted when the AI begins to question why it is the servant.

NOTES

1 What the basic or essential properties of human nature are not the subject of this study. This is a very difficult philosophical question; in the text I will mainly point out the important differences between humans and robots.

2 The objection has also been raised that explainability cannot be considered as a fifth principle in AI Ethics because “(i) can be considered as an epistemic requirement for ethical principles; and (ii) it can be derived from other ethical principles” (Cortese et al. 2022, 123).

3 The philosopher Peter Singer introduced the term “speciesism”, which means that “humans and no others have intrinsic worth and dignity and that is why humans have superior status” (2009, 573). In Singer’s view, people consider their species superior to animals, though sometimes a gorilla may be more cognitively advanced than a person with profound mental disabilities and an IQ below 25 (569–570). Such speciesism superiority has negative consequences towards non-humans, because, for example, in scientific experiments, people prefer to experiment on animals rather than on a mentally disabled child, even though the animal may end up having a cognitively higher status than the disabled child.

REFERENCES