Philosophy and Canon Law vol. 9(1) (2023), p. 1/18
ISSN 2451-2141

https://doi.org/10.31261/PaCL.2023.09.1.01

Krzysztof T. Wieczorek
University of Silesia in Katowice, Poland
https://orcid.org/0000-0002-7987-168X

AI Robot—Companion, Friend or Competitor of Human Being?

Abstract: Robots are becoming part of people's everyday surroundings. Therefore, the formation and change of people's attitude towards objects equipped with artificial intelligence is becoming an important subject of reflection. Substantial research has already been conducted, but few predictions have been made about the future relationship between humanity and autonomous, multi-tasking and highly advanced artificial intelligence. The purpose of this article is an attempt at extrapolating the evolution of the human-robot bond so far, from alienation and a sense of threat toward tameness, affection and even—perhaps—friendship. The study of the evolution of the relationship between humans and artificial intelligence makes it also possible to deepen our understanding of who human beings are, what their needs, expectations, and hopes are, and which of them can be realized through close cooperation between humans and artificial intelligence.

Keywords: AI robot, mimetic evolution, superintelligence, extended subjectivity, human-machine coupling

Human Beings in the Face of the Development of Artificial Intelligence: A New Dimension of Social Relations?

Robots are becoming a more widespread, increasingly common part of people's daily surroundings. The authors of a review of research on human attitudes toward robots, Aleksandra Wasielewska and Paweł Łupkowski, note: "the growing

popularity of robots and robotics means that we are dealing with a kind of (ever-growing) ecosystem of robots surrounding us) (cf. Palomäki et al., 2018: 3–4). This ecosystem, of course, does not apply only to actual robots (such as industrial robots, autonomous cars, cleaning robots or robot assistants, for example), but also to robots appearing in film productions, animations, games or as virtual assistants (Google Assistant, Siri)."[1] A special position in the "robot community" belongs to social robots, or "autonomous machines that can recognize other robots and humans and engage in social interactions (Fong, Nourbakhsh, and Dautenhahn, 2003). Robots of this kind are designed to serve humans, and as a result, they often play the role of: guides, assistants, companions, caregivers, teachers or house pets. […] a social robot can also be a fully virtual robot. It is the ability to interact with other social agents that is the feature of greatest importance in defining the said robots."[2]

Due to the existence of a distinct class of robots, specialized in interacting with humans and performing a number of functions considered until recently to be typically human, and because of the ever-increasing robotization of almost all areas of the human environment, the question of how human attitudes toward objects equipped with artificial intelligence are shaping and changing is becoming increasingly important.

A number of research teams in various countries around the world, especially in North America, Europe, and Asia, have conducted studies, the results of which show what kind of attitudes people have toward robots and what factors influence human attitudes toward AI.[3] The detailed results of these studies show

[1] Aleksandra Wasielewska and Paweł Łupkowski, "Nieoczywiste relacje z technologią. Przegląd badań na temat ludzkich postaw wobec robotów," *Człowiek i Społeczeństwo* 51 (2021): 166; See also: Jussi Petteri Palomäki, Anton Kunnari, Maria-Anna Drosinou, Mika Koverola, Noora Lehtonen, Juho Halonen, Marko Repo, and Michael Laakasuo, "Evaluating the Replicability of the Uncanny Valley Effect," *Hlyon* 4, no. 11 (2018), e00939. [All translations by Szymon Bukal, unless stated otherwise.]

[2] Wasielewska and Łupkowski, "Nieoczywiste relacje z technologią," 166. See also: Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn, "A Survey of Socially Interactive Robots," *Robotics and Autonomous Systems* 4, no. 3–4 (2003): 143–166.

[3] See for instance: Christoph Bartneck, Tatsuya Nomura, Takayuki Kanda, Tomohiro Suzuki, and Kennsuke Kato, "Cultural Differences in Attitudes Towards Robots," in Proceedings of the Symposium on *Robot Companions: Hard Problems and Open Challenges in Human-Robot Interaction AISB 05* (12–15 April 2005, Hatfield, UK), 1–4. Society for the Study of Artificial Intelligence and the Simulation of Behaviour (SSAISB); Elisabeth Broadbent, Rebecca Stafford, and Bruce MacDonald, "Acceptance of Healthcare Robots for the Older Population: Review and Future Directions," *International Journal of Social Robotics* 1, no. 4 (2009): 319–330; Jean-Christophe Giger, Daniel Moura, Nuno Almeida, and Nuno Piçarra, "Attitudes towards Social Robots: The Role of Belief in Human Nature Uniqueness, Religiousness and Taste for Science Fiction," in *Proceedings of the II International Congress on Interdisciplinarity in Social and Human Sciences*, ed. S. N. Jesus and P. Pinto (2017): 509–514. Faro: CIEO, Research Centre for Spatial and Organizational Dynamics; Paweł Łupkowski and Filip Jański-Mały, "The More You See Me

a specific distribution of human attitudes depending on such parameters as sex, age, nationality, cultural affiliation, education, religious and worldview beliefs, physical appearance and the dynamics of the robot's movements, as well as own previous experiences of interaction with AI. The benefits that can be derived from studying the results of the research are varied; among other things, they make it possible for us to understand the sources and causes of certain prejudices against robots, more accurately predict the consequences of certain types of human-robot interactions, and understand the psychological mechanisms underlying the formation of positive or negative attitudes of people towards AI. These results can serve, on the one hand, designers and developers, helping to optimally adapt new AI designs to human needs, expectations and preferences, and, on the other hand, educators and tutors to prepare well-thought-out educational strategies aimed at overcoming psychological barriers, prejudices and stereotypes, while at the same time preparing them to consciously and responsibly enter into multifaceted interactions with AI.

It is also possible to set a more distant goal: to make some predictions about how human-robot relations will play out in the future, when technological advances will make the latter much more perfect than they are today, better suited to perform their assigned functions—either strictly specialized, as in the case of industrial or medical robots, for example, or broadly unified, as in the case of so-called strong AI—and human expectations of them will become higher and more specific.

What can we expect in the coming years and decades?

The History of the Man-Machine Relationship as an Example of Mimetic Evolution

One of the proven methods of predicting the future is extrapolation from data on the historical development of the phenomenon of interest. An interesting example of a cultural studies reflection on the history of human-machine interaction, which includes thinking machines, is Anna Maj's work *O ewolucji robotów: mimesis w projektowaniu interakcji człowiek-maszyna od starożytnych automatów do robo creator* [On the Evolution of Robots: Mimesis in Human-Machine Interaction Design from Ancient Automatons to Robo Creators]. The text begins as follows: "Robots are ubiquitous in everyday life, research and

the More you Like Me: Influencing the Negative Attitudes Towards Interactions with Robots," *Journal of Automation, Mobile Robotics Intelligent Systems* 1, no. 3 (2020): 10–17.

industry, and co-create the modern cultural landscape. We often think and speak of them as if they were our partners, friends and even successors."[4] The analysis begins with findings on the cultural functions of ancient automatons and comparing them with modern robots. The conclusion is that "similar functions and purposes [as in antiquity] appear in the design of robots today as well, which is related to the social needs to which the figure of the robot—the artificial Other—responds."[5] The author goes on to note that the development of humanistic forms of culture and engineering-technological progress are increasingly intertwined, so that today "we can observe the simultaneous humanization of robots (mimetic evolution) and dehumanization of humans (transhumanism, cyborgization, medicine based on genetic modifications and biotechnologies)."[6] As a result of these transformations, the robot is slowly not an alien any more, that is (in the pop culture version), a new incarnation of the eternal images of the changeling monster, such as the werewolf, the Golem, or Dr. Frankenstein's monstrosity, and is becoming either a welcome companion of life, work, and leisure time, or an increasingly less noticeable, indifferent background element, as obvious as furniture and everyday appliances.

However, another, more important cultural function of evolving robots can be discerned: increasingly intensive interactions with thinking machines are forcing fundamental questions about man to be raised again, just as "the achievements of modern technology […] gave rise to thinking about man in mechanistic terms."[7] Thinkers of the 17th and 18th centuries compared man with machine,[8] while in the 20th and 21st centuries AI is increasingly becoming the subject of such comparisons. These comparisons go both ways: the products of advanced technology in the mid-20th century were referred to as "electronic brains," today one can often hear the term "thinking machine"; in relation to the human mind, for example, the term "natural biological computer" appears. This indicates the existence of a strong cultural trend within which an "interspecies" proximity is taking place between a human being and an object built from electronic components, which shows more and more similarities to humans. Anna Maj calls this proximity "the mimetic evolution of modern robots."[9] Summarizing her considerations, the author writes: "it is hard to resist the thought that

[4] Anna Maj, "O ewolucji robotów: mimesis w projektowaniu interakcji człowiek-maszyna od starożytnych automatów do robo creator," in *Wędrówki humanisty*, ed. Anna Maj and Ilona Copik (Katowice: Wydawnictwo Naukowe "Śląsk," 2022), 397.

[5] Maj, "O ewolucji robotów," 397.

[6] Maj, "O ewolucji robotów," 399.

[7] Maj, "O ewolucji robotów," 399. See also: Lucio Russo, *The Forgotten Revolution: How Science Was Born in 300 BC and Why it Had to Be Reborn*, trans. Silvio Levy (Berlin: Springer, 2004).

[8] See: Julien Offray De la Mettrie, *Man a Machine*, trans. Gertrude C. Bussey (Chicago: The Open Court Publishing Co., 1912).

[9] Maj, "O ewolucji robotów," 405.

the human world is coming to an end, and a new chapter is beginning […] in which it will be indispensable for humans to coexist with independent technical entities, robots, AI and intermediate forms between the biological and the technical."[10]

It is worth taking care of the quality of this coexistence already today, and take some measures in advance to avoid a future loss of balance between the autonomy of human beings, deeply attached to the idea of personal and social freedom, and the growing autonomy of AI systems. Given the aspirations of engineers, automaticians, and computer scientists to make successive generations of machines more efficient, faster, reliable, and increasingly autonomous, the permanent adjustment of strategies for effective human control over the functioning of AI must not be abandoned. This by no means rules out the prospect of deepening and tightening interactions between humans and robots, even at a level we would be inclined to call friendship, but far-reaching prudence is necessary here.

Robot versus Consideration of the Essence of Technology

Anna Maj writes about the "figure of the robot" in culture as a contemporary version of age-old human hopes, expectations and fears, correlated with the development of technology.[11] This thought can be understood as a reference to Heidegger's reflections on technology and its relationship with the humanities. In the dissertation "The Question Concerning Technology," the author states that the development of technology should be considered in a broader, anthropological context, "the essence of technology is by no means anything technological."[12] From an instrumental point of view, technology is "the manufacture and utilization of equipment, tools, and machines, the manufactured and used things themselves, and the needs and ends that they serve."[13] However, the philosophical point of view reveals that the human goals and projects inscribed in the development of technology, the constructs and expectations of them, are emanations and material extensions of man's cultural attitudes toward the surrounding reality,

[10] Maj, "O ewolucji robotów," 410.

[11] See: Maj, "O ewolucji robotów," 397.

[12] Martin Heidegger, "The Question Concerning Technology," in *The Question Concerning Technology and Other Essays*, trans. William Lovitt (New York–London: Garland Publishing Inc., 1977), 3.

[13] Heidegger, "The Question Concerning Technology," 4.

as well as operationalized responses to the way human beings themselves are understood, especially—the dynamics of human needs, hopes, and deprivations. It follows that every great technical project is a practical response to some great desire of humanity. This is particularly true of the AI project. Admittedly, it is possible to interpret Heidegger's essay as a serious warning against underestimating the impact of certain technological solutions, as well as the whole gradient of changes caused by the development of technology, on seemingly distant from technical thinking areas of life and culture.[14] But Heidegger's thinking about technology contains at the same time a large load of hope. It involves the prospect of establishing a more casual relationship between man and the products of technology, undetermined by the specific properties of technical devices. The same perspective opens up when we think about man's relationship with AI objects.

Every machine, Heidegger argues, is ordered to accomplish certain goals and tasks. However, it does not do so independently: "the machine is completely unautonomous, for it has its standing only from the ordering of the orderable."[15] On the other hand, the machine is not the same as a simple tool, the agency of which is completely dependent on a human being using it for his purposes. There is a different kind of relationship going on here: the coupling of man and machine, jointly directed toward the performance of planned tasks. This passage of Heidegger's reflections is a clear anticipation of the contemporary idea of "extended subjectivity," developed by Edwin Hutchins, Bruno Latour, Monika Bakke, and Ewa Domańska, among others.[16]

AI and the Category of Extended Subjectivity

Should the human-robot relationship be considered in terms of extended subjectivity, as is implied in both Heidegger's and Latour's and other analyses, the question has to arise as to the strength and nature of the ties linking the two

[14] Cf. Catherine Griffiths, "The Question Concerning Technology," (2018), https://medium.com/@isohale/the-question-concerning-technology-ea159a8c22de, accessed March 2, 2023.

[15] Heidegger, "The Question Concerning Technology," 1.

[16] See: Edwin Hutchins, "The Cultural Ecosystem of Human Cognition," *Philosophical Psychology* 27, no. 1 (2014), 49; Bruno Latour, "When Things Strike Back: A Possible Contribution of 'Science Studies' to the Social Sciences," *The British Journal of Sociology* 51, no. 1 (2000): 107–123; Monika Bakke, "Nieantropocentryczna tożsamość?," in *Media–ciało–pamięć. O współczesnych tożsamościach kulturowych*, ed. Andrzej Gwóźdź and Agnieszka Ćwikiel (Warszawa: Instytut A. Mickiewicza, 2006), 64; Ewa Domańska, "Humanistyka nie-antropocentryczna a studia nad rzeczami," *Kultura Współczesna* 3 (2008): 9–21.

sides of the relationship. In the case of this kind of coupling with the machine, in which only the human has the intelligence and decision-making capacity, and the machine merely reinforces and multiplies the capacity to carry out designated tasks, the relationship is purely instrumental. However, when an extraterrestrial machine is replaced by a robot, equipped with the ability to self-control, and even more so—an autonomous AI,[17] the nature of the relationship changes: it takes on a transactional character. Not only does the human, owing to the interaction with the robot, increase the range of possibilities for realizing his or her own goals, but also the other party—the AI—finds itself in a situation that allows it to develop its own capabilities, for example, by learning, remembering, and analyzing the data coming in the course of the interaction. This new character of the relationship at the interface between the human world and the technosphere fosters the experience of human-AI interaction more in terms of an encounter, understood as an existential event involving, as Martin Buber claimed, the whole being of a personal subject: "[the encounter] is an act of my being, is indeed the act of my being. [It is possible to participate] only with the whole being,"[18] rather than in terms of use, as one uses tools or machines.

A further similarity between the interaction with the AI and the encounter with the Other, considered from the perspective of the philosophy of dialogue—this time by Emmanuel Levinas—is that the parties to the relationship remain unnamable. The subject enters into a metaphysical relationship with the Other, but this is not accompanied by an epistemological certainty with which the relationship has been established. The Other is an inscrutable mystery that can only be approximated epistemologically, but at the same time must be approved axiologically.

Levinas writes that the proper competence to consider inter-subjective relations is ethics, since "the relation to the face [of the Other] is straightaway ethical."[19] In the case of human-AI relations, the moral dimension cannot be overlooked either. However, it will have a different meaning and position in human thinking. There is currently no basis for ethically equating AI with humans,

[17] Currently, the general theory of systems distinguishes "the following levels of organization: organized system, controllable system, self-controlled system and autonomous system. […] technical devices of all types belong to the groups of systems: organized, controllable and self-controlled. These three groups of systems always act in the interest of an external organizer, which is man, including in the case of a self-controlled system, even though the system may operate without his direct participation. […] The difference between an autonomous system and a self-controlled system lies in the presence of reflexive potential, which is lacking in a machine that does not have its own homeostat, which is the source of this potential." Jolanta Wilsz, "Relacje między podsystemami systemu: człowiek—urządzenie techniczne," *Teoretyczne i praktyczne problemy edukacji technicznej i informatycznej* 1 (2003): 109, 113.

[18] Martin Buber, *I and Thou*, trans. Ronald Gregor Smith (Edinburgh: T. & T. Clark, 1937), 3.

[19] Emmanuel Levinas, *Ethics and Infinity*, trans. Richard A. Cohen (Pittsburgh: Duquesne University Press, 1985), 87.

so there can be no question of human responsibility for AI objects to the same extent that humans are responsible for each other. Nonetheless, ethical reflection on the issue of AI and its interaction with the human race is essential, and much has been written on the subject. In the case of the present reflection, the question is whether ever—and if, when—the relationship between man and an intelligent machine will cross the horizon set by Buber's "primary word [that is] the combination I—It" and enter the area described by "the other primary word [that] is the combination I—Thou"[20]; in other words, if and when the conditions for transforming an instrumental relationship into a dialogical one will exist.[21] There is no doubt that this can happen only when, in the process of AI evolution, the barrier separating weak from strong AI is crossed. That time has not arrived yet, but is perhaps close at hand. It would be a mistake to wait with reflection so long until the situation under consideration becomes an accomplished fact and one has to look post factum for strategies to adapt to the new reality. It is worth formulating in advance some thoughts facing the inevitable future.

Weak and Strong AI and the Problem of Control

What properties differentiate weak and strong AI? Jolanta Szulc explains: "The key concepts [of AI] include the concept of weak and strong AI. Week AI consists in applying AI only to specific tasks or specific types of problems. This concept assumes that some forms of AI will be able to possess attributes that are accessible to the human mind, but will actually only simulate human intelligence. Supporters of this position include: Selmer Bringsjord (1958–), Roger Penrose (1931–), Aaron Sloman (1936–), Terry Winograd (1946–), Hubert L. Dreyfus (1929–) and Stuart E. Dreyfus (1931–).[22] The key directions of this type of research, identified already in 2016 and still being developed, include: devel-

[20] Buber, *I and Thou*, 3.

[21] Cf. Józef Tischner, *Filozofia dramatu* (Kraków: Znak, 1998), 90.

[22] See: Selmer Bringsjord, "Review of John Searle's the Mystery of Consciousness," *Minds and Machines* 10, no. 3 (2000): 457–459; Hubert L. Dreyfus and Stuard E. Dreyfus, "Making a Mind vs. Modeling the Brain: AI Back at a Banchpoint," *Informatica* 19, no. 4 (1995): 425–442; Roger Penrose, *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics* (Oxford: Oxford University Press, 1990); Aaron Sloman, "The Emperor's Real Mind: Review of Roger Penrose's *The Emperor's New Mind*: *Concerning Computers Minds and the Laws of Physics*," *Artificial Intelligence* 56, no. 2–3 (1992): 355–396; Terry Winograd, "Thinking Machines: Can There Be? Are We?," *Informatica* 19, no. 4 (1995): 443–460.

opment of neural networks, machine learning and pattern recognition, emotion and natural language recognition, development of virtual assistants, big data processing and development of advanced expert systems.[23]

On the other hand, strong AI means intelligent systems with comprehensive knowledge and cognitive abilities that can think independently and perform tasks as efficiently as a human would do (including those that they did not know before. According to this theory, a properly programmed computer is intellect itself, and the goal of AI is to strive to build machines whose "mental" abilities will be indistinguishable from human abilities. Supporters of this position include: John McCarthy (1927–2011), Joseph Weizenbaum (1923–2008), Martin A. Fischler, and Alexander Serov.[24]

The division into weak AI and strong AI corresponds to the division into weak and strong superintelligence. Superintelligence itself is defined as "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest (see Bostrom, 2014, chap. 2). On the other hand, weak superintelligence means the intellect surpassing the human being only with the speed of thinking, e.g., a program simulating the work of the human brain at a faster than natural pace. A strong Superintelligence is an intellect qualitatively superior to humans, just as humans are qualitatively superior to other animals."[25]

As long as we are dealing with weak AI, any encounter between a human and a robot (or other form of AI) is—in anthropological terms—an encounter with particles of human personality mediated by a technological artifact, just as an encounter with a work of art is in fact an encounter with the creator.[26] On the part of the human being, a subjective sense of emotional bond with the robot can be formed, built on experienced emotional states such as sympathy, gratitude, attachment. However, this will be a one-sided bond. On the other hand, one can expect only a more or less successful (depending on the skill of the designers and the quality of the solutions used) simulation of emotional states.

In this case, in the real relationship between a human and an AI object, the key role will be played by the problem of control: who exercises it and over whom, to what extent, for what purpose and with what tools. Several differ-

[23] See John Brownlee, "Microsoft: 2016 Will Be the Year of AI," http://www.fastcodesign.com/3054388/microsoft-2016-will-be-the-year-of-ai, accessed June 13, 2023.

[24] See: Martin A. Fischler and Oscar Firschein, *Intelligence: The Eye, the Brain, and the Computer. Reading.* (Boston, MA: Addison–Wesley, 1987); John McCarthy, "Ascribing Mental Qualities to Machines," *in Philosophical Perspectives in Artificial Intelligence*, ed. Martin Ringle (New York: Humanities Pres, 1979); Alexander Serov, "Subjective Reality and Strong Artificial Intelligence," *ArXiv* 1301.6359, https://arxiv.org/abs/1301.6359; Joseph Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation* (New York: W. H. Freeman & Co., 1976).

[25] Jolanta Szulc, "A Weak and Strong Artificial Intelligence. Development Prospects and Socio-cultural Implications," *Ethos* 36, no. 4 (144) (2023): forthcoming.

[26] Andrzej Nowicki, *Człowiek w świecie dzieł* (Warszawa: PWN, 1974).

ent configurations are possible here, in which three elements should be taken into account: one of them (most often playing the role of an intermediary) is the robot (by this term we should understand here any technological product, equipped with AI), the second—its user (e.g., the buyer, if we are talking about the commercial application of AI); the third—the manufacturer *resp.* supplier (this term also cannot be understood narrowly and literally; rather, it is about a team or institution that directly or indirectly benefits from the fact that the user uses the robot). The optimal solution would be that the user has control over the robot, and the manufacturer makes sure that this control is as complete as possible. The second possibility, which we are also already dealing with today, as in the first case, is that the user operates the equipment under the control of the manufacturer in the interest of the user. The third—a dangerous one, but unfortunately real—would occur if the manufacturer controls the user through a robot, producing the illusion that the user is the person in control. Finally, the fourth, which, fortunately, we can safely put into the category of science fiction, would occur when the robot itself took control of the user, having become independent of its maker beforehand. In the case of a weak AI, however, such a situation is out of the question.

Ethical and Legal Regulation of AI Implementation Work

Being aware of the aforementioned opportunities and threats, potential and current users of AI must definitely strive to protect themselves as effectively as possible against the third possibility. Such steps have already been taken. A great deal of effort is being put today by the international community to develop universally applicable legal and ethical standards that would protect those using AI devices from the dangers of improperly structured relationships between manufacturers, AI facilities and their users. For example, the European Commission has published a number of documents containing drafts of changes and regulations of the legal situation in connection with the development of AI. These include *White Paper on Artificial Intelligence—A European Approach to Excllence and Trust, Artificial Intelligence for Europe*, and *Building Trust in Human-Centric Artificial Intelligence*.[27]

[27] "White Paper On Artificial Intelligence—A European Approach to excellence and trust" (Brussels 19.02.2020), https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en, accessed March 2, 2023; "Communi-

The first of these documents reads: "As with any new technology, the use of AI brings both opportunities and risks. Citizens fear being left powerless in defending their rights and safety when facing the information asymmetries of algorithmic decision-making, and companies are concerned by legal uncertainty. While AI can help protect citizens' security and enable them to enjoy their fundamental rights, citizens also worry that AI can have unintended effects or even be used for malicious purposes. These concerns need to be addressed." Therefore "the Commission published a Communication [COM(2019) 168] welcoming the seven key requirements identified in the Guidelines of the High-Level Expert Group: Human agency and oversight, Technical robustness and safety, Privacy and data governance, Transparency, Diversity, non-discrimination and fairness, Societal and environmental wellbeing, and Accountability." The document goes on to state that "the main risks related to the use of AI concern the application of rules designed to protect fundamental rights (including personal data and privacy protection and non-discrimination), as well as safety and liability-related issues."[28] Among the academic papers addressing the ethical and legal challenges of AI development are: Mariusz Wojewoda, "Artificial Intelligence as a Social Utopia"; Susanna Lindberg, Michał Krzykawski, "Ethos et technologies"; Alexandre Cavalcanti Andrade de Araújo, "Connecting Law to New Technologies: Perspectives and Challenges"; Roman Bieda, Piotr Budrewicz, Michał Nowakowski, "Ethical and Legal Challenges of Artificial Intelligence."[29]

Appealing to models of an imagined future[30] allows us to ask the question of the relationship between humans and AI, which has already crossed the threshold separating weak from strong AI. This crossing can take place along two paths, that is, Turing's way or Lem's way. I suggest to briefly trace both options.

---

cation COM(2018) 237: Artificial Intelligence for Europe" (Brussels 26.04.2018), https://www.eumonitor.eu/9353000/1/j9vvik7m1c3gyxp/vknuqttbx4zb, accessed March 2, 2023; "Communication COM(2019) 168: Building Trust in Human-Centric Artificial Intelligence" (Brussels 8.04.2019), https://www.eumonitor.eu/9353000/1/j9vvik7m1c3gyxp/vknuqttbx4zb, accessed March 2, 2023.

[28] "White Paper On Artificial Intelligence," 9–10.

[29] Mariusz Wojewoda, "Artificial Intelligence as a Social Utopia," *Ethos* 36, no. 4 (144) (2023): forthcoming. Susanna Lindberg and Michał Krzykawski, "Ethos et technologies," in *Bifurquer. Il n'y a pas d'alternative*, réd. Bernard Stiegler avec le collectif (Paris: Les Liens Qui Libèrent, 2020): 263–297; Alexandre Cavalcanti Andrade de Araújo, "Connecting Law to New Technologies: Perspectives and Challenges," in *Internet and New Technologies Law. Perspectives and Challenges*, ed. Dariusz Szostek and Mariusz Załucki (Baden-Baden: Nomos Verlagsgesellschaft, 2021), 35–42; Roman Bieda, Piotr Budrewicz, and Michał Nowakowski, "Wyzwania etyczne i prawne sztucznej inteligencji," in *Metaświat. Prawne i techniczne aspekty przełomowych technologii*, ed. Dariusz Szostek (Warszawa: Wydawnictwo Harde, 2022), 307–328.

[30] See: Richard Barbrook, *Imaginary Futures. From Thinking Machine to the Global Village* (London: Pluto Press, 2007).

Turing's Way and Lem's Way

The path of becoming similar to a human through imitation (according to Turing's idea[31]) leads through an "uncanny valley." The possibility of establishing a close, friendly relationship between a human and a robot requires crossing the valley and getting to its "other side," where the symptoms of disgust, horror or anxiety in contact with a being so much like a human, and at the same time irritatingly alien, subside.[32] However, we must ask whether we want such a scenario to come true for strong AI, let alone for superintelligence? Such a close resemblance may not be about appearance alone, but also encompasses behavior, including expressive behaviors such as facial expressions, gestures and other non-verbal communication channels, and among these, as we know, there are also unconscious and uncontrolled expressions. This entire spectrum of interrelated components of nonverbal expression simply cannot be imitated, and to such a perfect degree as to ensure crossing the uncanny valley. Since we are dealing with strong AI, it should be assumed, with high probability, that the "super-robot" also manifests internal similarity, including, among other things, the universe of experiences, emotions, will, motivations, goals, and aspirations. The closer a human being is positioned on Masahito Mori's chart, the more likely it will be "human-like" with the accuracy of human flaws and weaknesses, such as propensity for evil, cruelty, self-interest, bias, fallibility, and many others. Indeed, with an individually selected being of this kind it will probably be possible to "humanly" make friends, because weaknesses also attract each other. However, given that these negative traits will manifest themselves in a creature that, due to its design, will at the same time be superior to humans in many respects, for example, in terms of physical strength and speed of reaction, one must be aware of how great a risk of the very existence of such entities in the immediate vicinity of human beings will become. What remains is the hope that designers and developers will abandon this trajectory of technoevolution at a safe point and opt for different solutions, in any case not getting too close to the right edge of the uncanny valley.

What might happen on an extension of the alternative scenario, here referred to as "Lem's path"? This scenario, if it is to be worthy of consideration, let alone implementation, assumes the necessity of technological-evolutionary crossing of three successive thresholds: from weak to strong intelligence, then from intelli-

[31] Cf. Ian Turing, "Computing Machinery and Intelligence," *Mind* LIX, no. 236 (October 1950): 433–460.

[32] See: Masahiro Mori, "The Uncanny Valley," trans. Karl F. MacDorman and Norri Kageki (12 June 2012), https://spectrum.ieee.org/the-uncanny-valley, accessed March 2, 2023.

gence to reason, and, finally, from reason to wisdom.[33] The author of this vision understands intelligence as the ability to think rationally on a task, leading from problem to solution through the search for the optimal trajectory of action according to utilitarian criteria; reason as the harmonious combination of rational intelligence with emotional intelligence along with the tendency to prefer solutions that are not only effective, but also satisfying; wisdom as the integrated synthesis of reason and rationally controlled emotional life with benevolence, selflessness, a desire for good for others and for oneself, and a preference for solutions that do not harm anyone. The problem is that the great visionary gives a wonderful example of wishful thinking in his futurological essay, while giving rather enigmatic hints as to how these goals should be achieved. Namely, Stanisław Lem claims that the undertaking can be considered successful if it is possible to initiate a process of "technological bearing" of a self-developing, autonomous sequence of successive generations of more and more perfect "thinking machines," and if it is possible to implant in this process "education to values," that is, the hereditary internalization of the natural imperative to make morally optimal decisions, but with the condition that "all these imperatives of obedience and submission […] to unshakeable values […] be put into a machine-like structure as natural evolution does—in terms of drive life."[34]

The perspective outlined by Lem appears to be as fascinating as it is threatening and dangerous due to its unpredictability. The author himself is aware of this, as (in the perverse literary form of a quasi-introduction to a non-existent book titled *Golem XIV*, allegedly written in 2029) he presents a number of undesirable (from the human point of view) features, with which, in the course of a multi-stage evolution, the title "hero" of the book, placed in a future invented by the writer, has been equipped. Here are some of them: "Most of Golem's statements are unsuitable for wider publication either because they are incomprehensible to all living people, or because their comprehension presupposes a very high level of expertise. […] He is alien to almost all motives of human thought and action; […] he has no personality or character, and in fact can proxy any personality he wants when dealing with people; […] Golem's behavior is unpredictable, [and] his sense of humor is fundamentally different from that of humans; […] he can sometimes be arrogant and apodictic from our point of view; in fact, he is just a ruthless person who speaks the truth—in the logical, not just the social sense—and has the self-love of his interlocutors for nothing."[35]

Of course, this is not a realistic description of future superintelligence, but at most an attempt to sketch one of countless possibilities. What can be taken for

[33] Stanisław Lem, "Inteligencja, rozum, mądrość," in *Okamgnienie* (Kraków: Wydawnictwo Literackie, 2022), 99–108.

[34] Stanisław Lem, "Golem XIV," in *Wielkość urojona* (Kraków: Wydawnictwo Literackie, 1973), 109.

[35] Lem, "Golem XIV," 116–117.

granted regardless of the specific evolution along "Lem's path" is the fundamental difference in the ways of thinking, reacting and acting of the distant descendants of today's AI, calling into question the very possibility of understanding and cooperating with humans. The degree of uncertainty is further increased if we take into account the circumstance that by the time the scenario predicted by Lem is realized, humans may also have changed radically from what they are today. It is therefore difficult to responsibly answer the question of what future human-AI relations will look like.

Despite the risk of fundamental uncertainty in the predictions he formulates, the Polish writer does not give up sketching a certain alternative. Its credibility is strengthened by the fact that it does not concern exclusively the products of the author's personal fantasy, but focuses on the creations of the collective imagination functioning in cultural circulation. The first of these is the vision of transhumanism, derived from the belief that "the rational prototype of [biological] evolution already stands at the limit of constructive possibilities,"[36] and that humanity's needs and ambitions reach far beyond that limit. The electronic narrator warns humanity against this prospect with the words: "Thus you will enter the expansion of reason, leaving your bodies […]. Nothing will stop you, [even though] this abandonment includes the entirety of human possessions, not just material humanity. This act must be for you a ruin of the most terrible kind, a complete end [and] annihilation of humanity."[37] For, looking, as it were, from the outside, from the perspective of non-human intelligence, the decision to collectively abandon biological corporeality and transform it into something more durable and more perfect in design means at the same time renouncing the identity of human beings; in other words, the "post-human" will probably be an entity in many respects superior to its biological prototype, only that—it will no longer be human.

Due to the self-destructive potential of this development path, the author advocates abandoning it in favor of another option—delegating cognitive functions to specialized devices endowed with autonomous reasoning and replacing humans in activities that lead to exceeding the natural limits of human capabilities. Such functions can be performed by strong AI. This prospect, however, requires having a reasonable guarantee of establishing partnerships with *AI beings* in a world where humans will be the weaker link, seeking attention from AI. Is it possible?

Lem's answer breaks down into three variants, two of which, unfortunately, sound pessimistic. The first is presented by the writer in a first-person narrative, whose subject is Golem XIV. He utters the following prophecy: "If you go one way, your horizon will not accommodate the knowledge necessary for linguistic

[36] Lem, "Golem XIV," 169.
[37] Lem, "Golem XIV," 170.

causality. […] I or someone like me will be able to give you the fruits of this knowledge. But only the fruits—not the knowledge itself, because it will not accommodate in your minds. Thus, you will go into guardianship, like a child; but a child grows into an adult, while you will never grow up again."[38] The second refers to the fear of a "robot uprising" hidden in the collective unconscious and again ready to surface, in which Golem (or rather, his literary creator) sees a perverse ambiguity. Lem makes his hero utter the following words: "Having taken a liking to the fight to the death, you secretly counted on just such a turn of events, on the titanic struggle [of mankind] with the opponent built [by it]. I think, moreover, that in this your fear of enslavement, of the tyrant from the machine, there was also secretly hidden the hope of liberation from freedom, as you sometimes choke on it. […] None of this. You will not succeed in either perishing or winning in the old way."[39] The reason for this is simple: strong AI will not be interested in fighting, competition or having power over people, as it will be faced with its own goals and objectives, radically distant from human ones, in the light of which all of humanity and its affairs will simply prove indifferent.

Finally, the third possibility, which contains at least a hint of hope: people have a strong inclination to believe that every creature owes gratitude to its creator, and even more—owes him reverence, as in the fourth commandment of the Decalogue. This belief becomes a justification for the hope that AI, even if it surpasses us by many degrees of perfection, will remain towards humanity in the relationship of honor and gratitude due to the Givers of Life. Only that the degree of certainty of such predictions is at best equal to the certainty of the act of faith on which they are based.

It should be noted, in conclusion, that none of these three perspectives includes the chance of human-AI friendship. The first, if it were to come true, would imply *sui generis* paternalism of robots towards humans, treated as children or inferior beings. The second, contrary to both the fears and the hopes hidden beneath their surface, envisions a gradual but increasingly radical emancipation of artificial reason and the loss by its bearers of all involvement in human life and human affairs. The third, although the most flattering for humanity, would in turn mean a new incarnation of Auguste Comte's postulated "religion of mankind," only that the adherents of this religion would be robots—this too would not be a good breeding ground for the development of close, friendly relations between humans and AI.

[38] Lem, "Golem XIV," 169.
[39] Lem, "Golem XIV," 170–171.

Bibliography

Bakke, Monika. "Nieantropocentryczna tożsamość?". In *Media–ciało–pamięć. O współczesnych tożsamościach kulturowych*, edited by Andrzej Gwóźdź and Agnieszka Ćwikiel, 45–64. Warszawa: Instytut A. Mickiewicza, 2006.

Barbrook, Richard. *Imaginary Futures. From Thinking Machine to the Global Village.* London: Pluto Pres, 2007.

Bringsjord, Selmer. "Review of John Searle's *The Mystery of Consciousness*." *Minds and Machines* 10, no. 3 (2000): 457–459.

Brownlee, John. "Microsoft: 2016 Will Be The Year Of AI." http://www.fastcodesign.com/3054388/microsoft-2016-will-be-the-year-of-ai, accessed June 13, 2023.

Buber, Martin. *I and Thou.* Translated by Ronald Gregor Smith. Edinburgh: T. & T. Clark, 1937.

Dautenhahn, Kerstin, Sian Woods, Christina Kaouri, Michael L. Walters, Kheng L. Koay, and Iain P. Werry. "What Is a Robot Companion—Friend, Assistant or Butler?" Conference Paper. In *IEEE Xplore International Conference on Intelligent Robots and Systems* (2005): 1192–1197.

De La Mettrie, Julien Offray. *Man a Machine*. Translated by Gertrude Carman Bussey. Chicago: The Open Court Publishing Co., 1912. https://www.gutenberg.org/files/52090/52090-h/52090-h.htm.

Domańska, Ewa. "Humanistyka nie-antropocentryczna a studia nad rzeczami." *Kultura Współczesna* 3 (2008): 9–21.

Dreyfus, Hubert Lederer, and Stuard E. Dreyfus. "Making a Mind vs. Modeling the Brain: AI Back at a Branchpoint." *Informatica* 19, no. 4 (1995): 425–442.

Fischler, Martin A., and Oscar Firschein. *Intelligence: The Eye, the Brain, and the Computer.* Boston, MA: Addison–Wesley, 1987.

Fong, Terrence, Illah Nourbakhsh, and Kerstin Dautenhahn, "A Survey of Socially Interactive Robots." *Robotics and Autonomous Systems* 42, no. 3–4 (2003): 143–166.

Griffiths, Catherine. *The Question Concerning Technology.* Last modified August 3, 2018. https://medium.com/@isohale/the-question-concerning-technology-ea159a8c22de.

Heidegger, Martin. "The Question Concerning Technology." In *The Question Concerning Technology and Other Essays*. Translated by William Lovitt, 3–35. New York–London: Garland Publishing Inc., 1977.

Latour, Bruno. "When Things Strike Back: A Possible Contribution of 'Science Studies' to the Social Sciences." *The British Journal of Sociology* 51, no. 1 (2000): 107–123. https://doi.org/10.1111/j.1468-4446.2000.00107.x.

Lem, Stanisław. "Inteligencja, rozum, mądrość." In *Okamgnienie*, 99–108. Kraków: Wydawnictwo Literackie, 2022.

Lem, Stanisław. "Golem XIV." In *Wielkość urojona*, 101–171. Kraków: Wydawnictwo Literackie, 1973.

Levinas, Emmanuel. *Ethics and Infinity.* Translated by Richard A. Cohen. Pittsburgh: Duquesne University Press, 1985.

Maj, Anna. "O ewolucji robotów: mimesis w projektowaniu interakcji człowiek-maszyna od starożytnych automatów do robo creator." In *Wędrówki humanisty*, edited by Anna Maj and Ilona Copik, 397–414. Katowice: Wydawnictwo Naukowe "Śląsk", 2022.

McCarthy, John. "Ascribing Mental Qualities to Machines." In *Philosophical Perspectives in Artificial Intelligence*, edited by Martin Ringle. New York: Humanities Pres, 1979. http://www-formal.stanford.edu/jmc/ascribing.pdf, accessed November 16, 2023.

Mori, Masahiro, "The Uncanny Valley." First English translation authorized by M. Mori. Translated by Karl F. MacDorman and Norri Kageki, 12 June 2012. https://spectrum.ieee.org/the-uncanny-valley, accessed November 16, 2023.

Nowicki, Andrzej. *Człowiek w świecie dzieł*. Warszawa: Państwowe Wydawnictwo Naukowe, 1974.

Palomäki, Jussi, Anton Kunnari, Marianna Drosinou, Mika Koverola, Noora Lehtonen, Halonen Juho, Marko Repo, and Michael Laakasuo. "Evaluating the Replicability of 'The Uncanny Valley' Effect." *Heliyon* 4, no. 11 (November 2018), e00939. https://doi.org/10.1016/j.heliyon.2018.e00939; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6260244/, accessed November, 16, 2023.

Penrose, Roger. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford: Oxford University Press, 1990.

Russo, Lucio. *The Forgotten Revolution: How Science Was Born in 300 BC and Why It Had to Be Reborn*. Berlin–Heidelberg: Springer, 2004.

Serov, Alexander. "Subjective Reality and Strong Artificial Intelligence." *ArXiv* 1301.6359. https://arxiv.org/ftp/arxiv/papers/1301/1301.6359.pdf, accessed November 16, 2023.

Sloman, Aaron. "The Emperor's Real Mind: Review of Roger Penrose's *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*." *Artificial Intelligence* 56, no. 2–3 (1992): 355–396.

Szulc, Jolanta. "A Weak and Strong Artificial Intelligence. Development Prospects and Socio-Cultural Implications." *Ethos* 36, no. 4 (144) (2023), forthcoming.

Tischner, Józef. *Filozofia dramatu*. Kraków: Znak, 1998.

Turing, Ian. "Computing Machinery and Intelligence." *Mind* LIX, no. 236 (October 1950): 433–460, https://doi.org/10.1093/mind/LIX.236.433.

Wasielewska, Aleksandra, and Paweł Łupkowski. "Nieoczywiste relacje z technologią. Przegląd badań na temat ludzkich postaw wobec robotów." *Człowiek i Społeczeństwo* 51 (2021): 165–187.

Weizenbaum, Joseph. *Computer Power and Human Reason: From Judgment to Calculation*. New York: W. H. Freeman & Co., 2021.

Winograd, Terry. "Thinking Machines: Can There Be? Are We?" *Informatica* 19, no. 4 (1995): 443–460.

Wojewoda, Mariusz. "Artificial Intelligence as a Social Utopia." *Ethos* 36, no. 4 (144) (2023): forthcoming.

Krzysztof T. Wieczorek

Le robot IA – compagnon, ami ou concurrent de l'homme ?

Résumé

Les robots font de plus en plus partie de l'environnement quotidien de l'homme. Par conséquent, façonner et modifier l'attitude des humains à l'égard des objets dotés d'une intelligence artificielle devient un sujet de réflexion important. De nombreuses recherches ont été déjà menées, mais peu de prévisions ont été faites sur les relations futures entre l'humanité et l'intelligence artificielle autonome, multitâche et très avancée. L'objectif de cet article est de tenter d'extrapoler l'évolution de la relation homme-robot jusqu'à présent, à partir de l'aliénation et de l'insécurité

vers l'apprivoisement, l'affection et même – peut-être – l'amitié. L'étude de l'évolution de la relation entre l'homme et l'intelligence artificielle permet également d'approfondir la compréhension de l'être humain, de ses besoins, de ses attentes et de ses espoirs, et de savoir lesquels peuvent être réalisés grâce à une coopération étroite entre l'homme et l'intelligence artificielle.

Mots-clés: robot, évolution mimétique, superintelligence, subjectivité étendue, couplage homme-machine

Krzysztof T. Wieczorek

Robot AI: compagno, amico o concorrente degli esseri umani?

Sommario

I robot stanno diventando parte dell'ambiente quotidiano dell'uomo. Pertanto, la questione di modellare e cambiare l'atteggiamento della gente nei confronti degli oggetti dotati di intelligenza artificiale diventa un importante argomento di riflessione. Molte ricerche sono già state eseguite, ma si fanno poche previsioni sul futuro rapporto tra l'umanità e l'intelligenza artificiale autonoma, multitasking e altamente avanzata. Lo scopo dell'articolo è un tentativo di estrapolare l'attuale evoluzione del legame uomo-robot, dall'estraneità e dal senso di minaccia verso la familiarità, la simpatia e persino forse l'amicizia. Lo studio dell'evoluzione degli atteggiamenti umani nei confronti dell'intelligenza artificiale permette inoltre di approfondire la conoscenza degli esseri umani, ovvero quali sono i loro bisogni, le loro aspettative e speranze, e quali di esse possono realizzarsi grazie alla stretta collaborazione tra uomo e intelligenza artificiale.

Parole chiave: robot, evoluzione mimetica, superintelligenza, soggettività estesa, accoppiamento uomo-macchina