Krzanowski, Roman; Polak, Pawel

The future of AI : Stanisław Lem's philosophical visions for AI and cyber-societies in Cyberiad

Pro-Fil. 2021, vol. 22, iss. Special issue, pp. 39-53

ISSN 1212-9097 (online)

Stable URL (DOI): https://doi.org/10.5817/pf21-3-2405
Stable URL (handle): https://hdl.handle.net/11222.digilib/144849

Pro-Fil (Special Issue) 2021: 39–53
https://doi.org/10.5817/pf21-3-2405

THE FUTURE OF AI: STANISŁAW LEM'S PHILOSOPHICAL VISIONS OF AI AND CYBER-SOCIETIES IN CYBERIAD

ROMAN KRZANOWSKI

Department of Philosophy, Pontifical University of John Paul II, Poland, rmkrzan@gmail.com

PAWEL POLAK

Department of Philosophy, Pontifical University of John Paul II, Poland, atpolakp@cyf-kr.edu.pl

RESEARCH PAPER ▪ SUBMITTED: 4/10/2021 ▪ ACCEPTED: 11/11/2021

Abstract: Looking into the future is always a risky endeavour, but one way to anticipate the possible future shape of AI-driven societies is to examine the visionary works of some sci-fi writers. Not all sci-fi works have such visionary quality, of course, but some of Stanisław Lem's works certainly do. We refer here to Lem's works that explore the frontiers of science and technology and those that describe imaginary societies of robots. We therefore examine Lem's prose, with a focus on the *Cyberiad* stories, to see what challenges our future technological societies may face when they entrust their lives to AI technology. For example, what questions should we ask, and what questions do we forget to ask, when developing AI systems and allowing these systems to control our lives. The technologically honed minds of our current technocrats are perhaps too limited to guide us into this future, because AI-based technology is relatively unchartered territory, as any new, complex technology is by nature. Lem's visions of future societies oriented around AI and robotics portray AI technology in a deeper and more nuanced way than the current technological visions offered by our leading technological prophets. Based on Lem's visions, what is to come may not turn out to be an AI-driven nirvana.

Keywords: AI Future; Robotic societies; Lem's visions of the future; cyber-societies; humanity's prospects

Introduction

Mainstream artificial intelligence (AI) research pays rather limited attention to the unpredictability of AI technology and its potential impacts on society, despite the fact that AI-based technology is relatively unchartered territory,[1] as any new, complex technology is by nature. For example, in John Brockman's 2020 collection of discussions about the state of AI

[1] We make this claim because it is a very new technology, and despite its successes, it is still in its early stages of development.

and its future development by leading AI researchers, little attention is paid to the potentially harmful effects of this technology. Clearly this should not be the case, and a glaring example of the harms that can result from such a *laissez-faire* attitude is provided by the social media platforms. Their impact on society, people's mental health, and democracy were originally presented as being beneficial (e.g., Zuckerberg 2017), but this was a gross miscalculation (e.g., Mineo 2017, Kaiser 2019, Wylie 2019, Zuboff 2019, Milmo 2021, Rimbert 2021). On the flip side, we could argue that the potential for such technology to wreak havoc in the fabric of society was always there, so it is debatable whether these nefarious capacities came about only after these systems were deployed or whether these systems were developed with such objectives from the very beginning.[2] The experience of social networks may be repeated with other AI-based technologies.

The true situation is probably that the creators of technology, even when they consider its potential harmful effects, never fully anticipate all the possible implications of what they are creating[3]. For many reasons, but let's call them "typically human" ones, they never publicly disclose their real intentions nor the recognized implications of their wares, stressing instead the benefits of their inventions. (After all, who would choose to publicly fund a doomsday device?)[4] In any case, we should be suspicious of any new technology that promises us a new nirvana or easy solutions to the eternal problems of humanity in some Promethean vision. We should question what these technologies will bring to us and how they will affect our societies and way of living. But how can we fathom such an unknown future?

Peering into the future is always a risky venture, and predicting the future of AI is no different. One way to anticipate the shape of yet-to-come AI-driven societies is to examine the visionary works of some sci-fi writers. Of course, not all sci-fi works have this visionary quality, but it is safe to say that some of Stanisław Lem's works do. (In reality, most sci-fi works are commercially oriented and deprived of any lasting value; see Swirski (1997) comments about sci-fi literature.) We suggest here that Lem's works explore the frontiers of science and technology, as well as the functioning of imaginary societies of robots, which are in fact humans that have had their nature transformed into machines.

---

[2] Researchers recently voiced strong concerns about the benefits of uncontrolled AI development: "Stephen Hawking, Elon Musk, Steve Wozniak, Bill Gates, and many other big names in science and technology have recently expressed concern in the media and via open letters about the risks posed by AI, joined by many leading AI researchers." Available at https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/?cn-reloaded=1. For more discussion about AI's benefits and threats, see, for example, the wider Future of Life Institute website linked above, as well as the discussion from 2018 Artificial Intelligence and the Future of Humans, which is available at https://www.pewresearch.org/internet/2018/ 12/10/artificial-intelligence-and-the-future-of-humans/. The point to make here is that what "*Stephen Hawking, Elon Musk, Steve Wozniak, Bill Gates, and many other big names in science and technology"* are saying now, Lem said 60 years ago.

[3] The predictions of future in general and future of technology specifically are mostly wrong (e.g., Pogue 2012, Pestov 2017, Larkin 2018, Bush 2021). Despite these obvious failures the research into the future is a serious business, that in some (rare) cases may be accurate, as The Rockefeller Foundation (2010) report on the future of technology Lockstep scenario shows.

[4] With the pandemic ravaging the world (we are currently in 2021), one may revise the claim that public research funding is always beneficial. See, for example, Maher (2012), Millett et al. (2016), and Selgelid (2016). As always, "should" is not "is."

While Lem's language and stories may seem bizarre now, when we match modern AI concepts to Lem's ideas, these stories represent penetrating philosophical explorations of General AI, Universal AI, supra-intelligence, artificial consciousness, post-humanism, trans-humanism, and autonomous systems well beyond level 6 (see, for example., *The Six Levels*…, 2021). Lem depicts societies of the future where software and hardware constructs saturate every aspect of private and public life, so they really are AI–cyber societies. These societies are entirely computerized and automated, and smart machines make decisions everywhere. Lem explores how these potential societies function, what their values are, and what drives them.

However, for all their fantastic and amusing aspects, as Lem himself indicates, the stories should not stop the reader from perceiving the deeper meaning. Lem is unbounded by local technological and rational considerations (even physical laws), at least in terms of his time and place, so his landscape of possible variations in the societal forces and mechanisms reaches beyond what the visionaries and prophets of technology, even today, can conceive or imagine.[5] He is also not bound by political correctness or the requirements of funding agencies at the time.

In this essay, we specifically explore Lem's *Cyberiad* stories and the visions within them of AI-Cyber societies and humanity.[6] This is only a small sample of Lem's oeuvre, but such a narrow focus gives us an opportunity to more closely look at Lem's ideas than would be possible with a more generalized or wider study, especially given that we are constrained by the limited length of this paper.[7] The *Cyberiad* subtitle is *Fables for the Cybernetic Age*, so it is not "of" but "for," signifying that these are not just stories but stories with a message, one to learn from and be forewarned by.

We look at five stories: *How the World Was Saved, Trurl's Machine, The First Sally (A) or Trurl's Electronic Bard, The Fifth Sally or The Mischief of King Balerion, and The Sixth Sally or How Trurl and Klapaucius Created a Demon of the Second Kind to Defeat the Pirate Pugg*. These stories were selected because they carry, in the authors view, clear warning messages for us and our technological age. They warn us of the possible nefarious consequences of new technologies if they are allowed to be deployed unchecked and uncontrolled. The protagonists in all the *Cyberiad* stories are two robot constructors, Trurl and Klapaucius, a fact that brings Lem's stories even closer to some of our current ideas (e.g., Lovelock's Novacene cyborgs). Thus, in this essay, we regard Lem primarily as a philosophical visionary of post-human societies.

---

[5] In this essay, we are not interested in the philosophical aspects of Lem's works or his views on philosophy. Of course, Lem's stories have philosophical dimensions, but they are not works of philosophy per se. As such, they are fascinating in their own right, despite having limited philosophical import. The philosophy in Lem works is reflected in the Kobiela – Gomluka (2021) collection of essays. Here, we focus on exploring Lem's visions of future societies where new technologies dominate in order to understand what these technologies may mean for us. Lem's visions were ahead of his time, at least at the time of their writing and arguably even now in late 2021. Only slowly do we begin to realize their significance, and only gradually do we begin to understand Lem's message about technology, humanity, and our future.

[6] More about the *Cyberiad* stories can be found at http://english.lem.pl/index.php/works/novels/the-cyberiad.

[7] For a review of Lem's corpus, a list of his works, and insightful commentaries, one may refer to *A Stanisław Lem Reader* (Swirski, 1997) or (Kobiela – Gomłuka, 2021) collection of papers.

Visions of AI Technology and Robotic Societies

The *Cyberiad* story *How the World Was Saved* begins with the construction of a machine with an innocent and almost mundane capability. As Lem writes, "One day, Trurl the constructor put together a machine that could create anything starting with n[8]" (Lem 2014, 3). This innocent function, however, very soon turns out to serve as a doomsday device. The machine, on receiving a request to produce nothing ("I will tell you do to nothing.") started to empty the universe of things (i.e., it began to produce nothingness). How come, though? Was it a programming error or an unforeseen consequence of the design, which is so often the case in our primitive AI systems? This machine is saying this: "Do not be deceived. I have begun, it is true, with everything in *n*, but only out of familiarity. To create however one thing, to destroy, another thing entirely. I can blot out the world for the simple reason that I am able to do anything and everything—and everything means everything—in *n* and consequently Nothingness is child's play for me. In less than a minute now you will cease to have existence, along with everything else" (Lem 2014, 6). The machine is fortunately stopped before it destroys everything.

So, what are the lessons here? Quite a few in fact. Let's start with *lesson one*: Seemingly innocuous technologies, as AI may appear to some, can have an unforeseen potential for destruction. We just cannot foresee what may happen once these technologies are deployed and what havoc they could potentially wreak on our lives, societies, and planet. What is more, there is no returning to the previous state, whatever it was, as the machine is saying, "[…] of course I can restore nonsense, narrow-mindedness, nausea. As for other letters, however, I cannot help" (Lem 2014, 7). *Lesson two* is that once these constructs are deployed, it may be almost impossible to stop them, and even if we do succeed in this, our devices may leave behind an irreparable trail of destruction. In the story *How the World Was Saved,* once the machine was stopped in the process of producing nothingness, it says, "Take a good look at this world, how riddled it is with huge, gaping holes, how full of Nothingness, the Nothingness that fills the bottomless void between the stars, how everything about us has become lined with it…And I hardly think the future generations will bless you for it…" (Lem 2014, 7). *Lesson three* reflects how engineers and programmers cannot fully foresee what they produce, because the possible causal interactions are infinite. This conflicts with classical engineering, where artifacts have a narrowly defined scope of proper use. On introducing AI engineering with general purpose solutions like *general intelligence*, there is a seemingly exponential explosion of possible interactions, making foresight impossible. Consequently, engineers' visions are technology bound, so they cannot be left alone to their own devices.

In the story *Trurl's Machine*, Trurl, the constructor of the previous machine, builds a gigantic thinking machine. The machine has enormous physical proportions and massive internal complexities, such that even Trurl does not understand the real capacities of this machine. As a test, Trurl asks the machine what two plus two is, and the machine answers that it is seven. This starts a dialogue between Trurl and the machine: "Nonsense, my dear. The answer is four. Now be a good machine and adjust yourself. What is two and two?" (Lem 2014, 9). The machine

[8] All quotations and references to *The Cyberiad* relate to the 2014 edition of Lem's *The Cyberiad* published by Penguin Books.

answers again that it is seven. Despite several attempts to adjust the machine, these attempts seem rather haphazard, and the machine continues to insist on two plus two being seven. Trurl begins, in exasperation, to insult the machine. The machine finally responds, "You have insulted me for fourth, fifth, sixth and eight times. Therefore, I refuse to answer all further questions of a mathematical nature" (Lem, 2014, 11). At that point, the machine dislodges itself from its foundations and begins to pursue Trurl with the obvious intention of punishing him for all his insults. Trurl runs away with the machine in hot pursuit, with it wreaking havoc and destruction in its path. The machine finally succumbs to natural forces and dies in a sense, thus saving Trurl.

Now, what are the lessons in *Trurl's Machine*? Again, there are quite a few. *Lesson one* is that a sufficiently complex system will always be beyond our understanding and control. AI systems are already like this and will continue to become even more so in future. *Lesson two* is that super-intelligent AI systems, such as what we think of as General AI systems, could have their own logic that we cannot understand. These systems may pose a threat to societies, because their logic is not our logic, and their objectives, being autonomously created, will not necessarily align with ours. Indeed, some form of the paper clip AI factory is a highly possible scenario (Bostrom 2014, Rogners 2017, Gans 2018)! *Lesson three* is that engineers are generally seen in a positive light as individuals who are improving human living standards, but this is generally an illusion that derives from the uncritical Promethean view of technology.

The reality of the engineering profession is different, because engineers and other technologists rarely foresee, nor are they interested in foreseeing, the effects of their constructions beyond the purely technical perspective. Rarely are they aware, or wish to be aware, of the huge moral responsibility their work may carry. Due to this shortcoming, engineers can potentially play very negative roles in society. Lem warns us that technology in the hands of short-sighted individuals can be dangerous when left without supervision and societal control, and by no means should we perceive engineering as a kind of Nietzschean *Übermensch* beyond reproach. *Lesson four* is that building autonomy without control into AI-driven systems will result in systems that set their own goals and *telos*, and some possible scenarios that may not necessarily benefit the human race are presented by Lovelock (2020). Lovelock envisions a new era on the Earth called Novacene in which synthetic machines take over the Earth. In Novacene, the Earth is populated by cyborgs, which are self-replicating, self-improving mechanical systems that will eventually dominate and rule the Earth. These cyborgs will possess intelligence and knowledge beyond our understanding. To quote Lovelock, "cyborgs […] will design and build themselves from the artificial intelligence systems we have already [sic!] constructed. These will soon become thousands then millions of times more intelligent than us" (Lovelock, 2020, 29). Despite the vast difference in intellectual power, the relationship between cyborgs and humans will be peaceful, at least in the early stages of Novacene, as Gaia (i.e., the Earth) will need to maintain biological life (including us) to maintain the thermal balance of its (her?) ecosphere. In other words, cyborgs and we will need each other, but only during the early phase of Novacene (Lovelock, 2020, 30). Eventually, however, we will be "no more masters of our creations than our much-loved pet is in charge of us" (Lovelock, 2020, 119). It seems Lem was not too far removed from the current thinking about the prospects of General AI and synthetic life, because Lovelock's Novacene cyborgs and Lem's Trurl-made thinking machines are frighteningly similar. An extra twist is added to Lem stories, though, when we realize that the two engineer–constructor protagonists of the *Cyberiad*, Trurl and Klapaucius, are robots themselves. They are creating new

living systems by developing new inventions, just like Lovelock imagined in Novacene. We may wonder if the *Cyberiad* world is a reflection of Lovelock's Novacene, or vice versa?

In the story *The First Sally (A) or Trurl's Electronic Bard,* we encounter a newly constructed machine that can write poetry. The basis of the concept behind the software for this electronic bard is that all poets are products of civilization, one where the mind of a poet is a sort of program working with accumulated information. Thus, it seems obvious that if we feed the entire history of human civilization into the machine, it should be able to generate poetry, having been given some understanding of what poetry is. After generating some incomprehensible garbage, Trurl's machine begins to create complex poetry and eventually exceeds in quality and quantity the work of any human poet. The generated verses take any desired form, and the machine floods the inhabitants of the universe with masterpiece after masterpiece, eventually converting stars and galaxies into forms of poetic work. Ultimately, nobody wants these masterpieces anymore, nobody wants this synthetic art, but the machine cannot be stopped. The disastrous works of this electronic poet were eventually curtailed not by turning the machine off but by hauling it to a dark corner of the universe where its products cannot be seen or heard by anyone.

So, where is the lesson in this? There is quite a substantial one: We are exploring the creativeness of AI systems, something we are bound to see more of in future. We have paintings generated by AI, and we have poetry and novels (*Art that…* 2021, Hart 2020, Lau et al. 2020, Tang 2020). All these efforts are based on Lem's assumption in the story, namely that human creativity is simply a (programmable) function that requires suitable software and a knowledge base. As such, a specific work of art is simply a specific combinatorial construct selected based on some optimization function that follows some rules. It uses a knowledge base representing the human experience, or at least some digital version of it. Human creativity can therefore be programmed, because the human experience is "digitizable," and there is nothing special about it. There is therefore nothing special about us humans. Thus, Shakespeare was just the lucky monkey (*Parable of Monkies*, 2021)? With such creative AI, we could potentially flood the universe with unheard of masterpieces in vast numbers, because machines will produce more and more, faster and faster, like an army of Beatos de Lieban, Leonardos, Matisses, Turners, Dicksons, Blakes, Owens, Sassons, Kafkas, Lems, and so on. But is this really what we want? Is an AI–generated Blake or Dickson really Blake or Dickson? We suggest not, which begs the question of what the point of AI-created art is? Is there one? News about computer-generated poetry and art is more aimed at the lay audience or funding agencies to generate some excitement in the media rather than report some significant technical progress. Lem's message, however, is that in future societies, human art will still have a unique place, whatever Google and other technocratic institutions may claim.[9]

In the story *The Fifth Sally or The Mischief of King Balerion,* we witness the havoc resulting from a device that facilitates the transfer of the mind and its complete personality from one body

[9] Art is perceived as the search for the expression of human spiritual values like beauty, goodness, and truth. If society were flooded with artificial art, it would inevitably lead to devaluing the meaning of art and the role it plays in society. This would in all probability lead to the "mechanization" of man and the loss of spirituality, as suggested by Jay David Bolter in (Bolter, 1984).

to another. This device has of course been constructed and created by Trurl. It is acquired by the king of a distant world as a sort of toy, so he can play hide and seek with his subjects. A problem arises after several exchanges because nobody knows where the king is, or more specifically, *who* the king is, because the king has taken on the physical embodiments of various characters. Thus, for example, a sailor claiming to be the king looks to all outside observers to be a sailor, despite his claims to the contrary. A cascade of similar cases follows. Trurl, in the king's body, is under deep sedation, because Trurl cannot grasp the reality of the situation, and he claims that he is not the king despite observers clearly seeing him as the king. Klapaucius, after enumerable mishaps, explains to the king-Trurl the gravity of the situation, and it is only then that Trurl restores order by putting himself into his own body again. In turn, the sailor's mind is moved into the body of the king, the mind of the king is passed into a cuckoo clock, and the cuckoo clock's mind is placed in the body of the chief policeman. This way, order in the kingdom is restored. With some irony, Lem adds that the cuckoo clock with the mind of the king and the chief policeman with the mind of the cuckoo clock were performing just right.

Is there a lesson for us in *The Fifth Sally or The Mischief of King Balerion*, even if the story seems completely improbable? What is this story about? In the story, Lem talks about synthetic minds that have an ontological presence and can therefore be thought of as manipulatable objects, just like how AI parlance refers to synthetic minds and whole brain emulation (WBE). A synthetic mind is an artificially created system that seeks to emulate some aspects of human mental function, while WBE is a theoretical framework for exploring the possibility of creating a synthetic mind that accurately emulates all aspects of the human mind. In this case, Lem is talking about complete mind transfer (i.e., WBE).

A WBE is by definition indistinguishable from the human mind (Shanahan 2014, Sandberg – Bostrom 2008, Eckersley – Sandberg 2013, Koene 2006, Hayworth 2010). Sandberg and Bostrom position the brain and the mind as being essentially the same, so they claim that an exact emulation of the brain will also be an exact emulation of the mind. In other words, the functions of the brain are the functions of the mind. We are of course operating on the assumption that WBE is technically feasible, even if our current technology is not adequate for it (e.g., Shanahan 2014, Sandberg – Bostrom 2008, Hopkins 2012). Indeed, we are merely claiming that there are no technical obstacles[10] prohibiting such a construct in principle. Such an emulated brain would be denoted as an artifact, where by "artifact" we refer to a construction that has been purposely created by a human. (For a more complete definition of this term, see, for example, Margolis – Laurence 2007, Houkes 2009, Thomasson 2009, Borgo – Vieu 2009). Our definition therefore follows the dictionary usage.[11] We do not identify any specific physical medium for such an artifact. It could be made from biological material or be silicon-based much like current computer hardware, or it could be something else entirely (Shanahan 2014). What is important is that it is a human construct rather than a naturally occurring object. As WBE is defined as an "exact working copy of a particular brain in nonbiological substrate" (Shanahan 2014, 15, Sandberg – Bostrom 2008), it replicates all the capacities of the human brain.

[10] For example, such emulation would not break any physical laws like a *perpetuum mobile* system would.

[11] "[A] usually simple object (such as a tool or ornament) showing human workmanship or modification as distinguished from a natural object" (Merriam-Webster, 2021).

Thus, the research that would to some extent justify that the concept behind *The Fifth Sally or The Mischief of King Balerion* is already being realized, at least in a conceptual manner. However, nobody in the research community questions what the consequences would be were WBE capacities actually available to us. Lem shows us what may possibly happen once we master WBE technology. We can only hope that the WBE of *The Fifth Sally* will never be fully realized.

The story of *The Sixth Sally or How Trurl and Klapaucius Created a Demon of the Second Kind to Defeat the Pirate Pugg* is about a machine that collects information. In the Pirate Pugg's own words, "My name is Pugg, I am thirty arshins in every direction, and it is true I rob, but in a manner that is modern and scientific, for I collect precious facts, genuine truths, priceless knowledge, and in general all information of value" (Lem 2014, 148). The Pirate Pugg hunts for cosmic caravans passing his way and releases his captives only after they give him all the information they have, and the information must be novel, true, and relevant. As the Pirate Pugg has been collecting information for eons, hardly anyone can ever satisfy his thirst for knowledge, so most encounters with this machine end up with the destruction of the captives. Trurl and Klapaucius only escape the Pirate Pugg after providing him with a device capable of extracting all the relevant information from the atomic level of cosmic matter. Trurl and Klapaucius's machine for extracting information swamps the Pirate Pugg with relevant and true information at a speed and volume that swiftly overwhelm his capacity to comprehend, making him practically inoperable. In Lem's words, "thus was the Pirate Pugg severely punished for his inordinate thirst for knowledge" (Lem 2014, 159).

It may be that *The Sixth Sally* comes closer to our reality than the other stories. Indeed, we are living in an information age where we are flooded with information, so we no longer know (if we ever did) what information is relevant and what is not. What will this lead to? Will we be made "practically inoperable" by the excess information? What should we know? And what is the "right" amount of information we need to live? Also, what would it mean for us to become "practically inoperable"? Does this mean that we would become passive epistemic agents, just consuming whatever comes our way?

As implied in *The Sixth Sally*, it seems that a contemporary, comprehensive, and complete description or conceptualization of the world, or indeed the universe, is impossible without some notion of form, organization, or structure (Krzanowski 2020, Krzanowski 2021). The constructors' machine extracts information (i.e., knowledge) from the atomic structures of the Universe. (We should add that when Lem was writing *The Cyberiad*, the dominant concept of information was mostly related to Shannon's theory of communication and related domains.) A reductive description of the world in purely mechanical terms based on groups of elements is incomplete, so some form must be added to it. This means that the elements making up the world, whatever they may be, must have some organization or form to become something. It is then quite easy to claim that information (in some way or other) is everywhere and in everything, so it is a foundation of the universe—this is what Lem is saying.

The modern realization of the Pirate Pug is the Internet. It is a relentless collector and disseminator of information, any information. While the Pirate Pug ambushes cosmic caravans, the Internet also ambushes us. The Internet and the Pirate Pug are pro-active, dynamic epistemic agents. The term epistemic agency (EA) is most commonly used to denote the ability to choose,

reflect upon, and freely form beliefs (e.g., Elgin 2013, Olson 2015, Puzzo 2015). Epistemic agency may also refer to a passive or active capacity of a system, organization, artifact, or person to impact or influence someone else's doxastic position (e.g., Schlosser 2019). The Internet therefore has epistemic agency in the sense of influencing our beliefs, views, and choices (e.g., Wylie 2019, Zuboff 2019).

Our own epistemic agency is founded on free access to information and a set of critical reasoning skills, namely our reasoning and judgment faculties. These act as knowledge filters separating useless information from the relevant information in the knowledge machine. While the existence of our epistemic agency is often denied and trivialized (e.g., Kornblith 2012, Ahlstrom-Vij 2013, Puzzo 2015), it serves a critical function for us by providing the foundation for our knowledge of the world, because poor knowledge leads to poor decisions and life choices. While it is true that our beliefs are largely shaped by our schools, parents, society, culture, the media, and religions and their clerics, we have developed critical faculties to evaluate these influences and gained some understanding of their role and *modus operandi* as epistemic agents. Indeed, the precise function of epistemic agency is to reflectively engage with the world. The extent to which we care to do it, rather than just passively internalize external messages, is a separate question.

The Internet can be understood as a technological complex for transferring, harvesting, analysing, and manipulating data and its users' experiences, and this is a new form of epistemic agency. With it permeating even the most intimate aspects of our lives (e.g., Wylie 2019, Zuboff 2019), we urgently need to affirm its epistemic role. Indeed, the Internet's epistemic purpose is not to ensure our well-being but rather instill in us someone else's values and beliefs (e.g., Kaiser 2019, Wylie 2019), all with the sole purpose of making us less critical, more obedient, less reflective, and more susceptible to external persuasion. In fact, the Internet, as an epistemic agent, seeks to dissolve our doxastic attitudes and substitute them with artificially created ones. Once we are deprived of our own epistemic capacities, the Internet can control our choices, decisions, views, and beliefs in ways we do not realize. There are ample examples of this (e.g., Mineo 2017, Kaiser 2019, Wylie 2019, Zuboff 2019). The mechanisms that the Internet uses for its epistemic agency fall nothing short of brainwashing, with it employing devices like perspecticide, echo chambers, filtering, personalization, astroturfing, fake news, deep fakes, and cognitive hacking, to name but a few. The scale of these activities and their destructive effects on society are difficult to fathom for a non-technical person (e.g., Gibbs 2014, King et al. 2017, Kaiser 2019, Wylie 2019), which includes most of the public. Lem's lesson from this story is that too much information, even relevant information, does not lead to knowledge but rather stupor and apathy.

What Do We Learn from Lem?

The questions about AI technologies and the future of humanity, ones we are just now beginning to realize, were foreseen by Lem more than half a century ago when computers were a little-known construct and post-humanity was a non-existent concept. What is more, AI technology was nothing but a dream that was reflected somewhat in Turing's and Wiener's works. Lem did not have the insight, experience, or expertise of Stephen Hawking, Elon Musk, Steve Wozniak, or Bill Gates, yet his visions were much more specific and precise than the

prophecies of these individuals. Camouflaged in sci-fi stories and couched in the fantastic language of fairy tales, Lem's visions of a technological post-human, yet so human, civilization driven by automated systems is only now, and by very few, being slowly comprehended as events unravel before our eyes.

The stories discussed in this essay illustrate the dark scenarios resulting from uncritically accepting AI technology, something that has been alluded to in more vague ways by Hawking, Musk, Wozniak, and Gates. The story of *How the World Was Saved* teaches us that complex technology always has some unforeseen consequences that may have disastrous, irreversible effects. *Trurl's Machine* conveys the warning that General AI or super-intelligent systems will have their own logic and goals, and we may be victims of their designs. *The First Sally (A) or Trurl's Electronic Bard* represents a critique of synthetic art, because art is the expression of a deeply human experience, spiritual life, and values. Any form of synthetic art will therefore be empty and vacuous. *The Mischief of King Balerion*, meanwhile, carries the warning that any attempt to emulate the human mind through WBE will, in all probability, wreak havoc through society, something that is never mentioned by WBE research teams. *The Sixth Sally* or *How Trurl and Klapaucius Created a Demon of the Second Kind to Defeat the Pirate Pugg* is a clear warning against the unlimited flow of unfiltered information, which will result in mental stupor rather than the enlightenment that the prophets of the Information Highway have promised.

Lem's visions of the dangers of AI and cyber societies are much more specific than any warnings coming from our own technology gurus.[12] Indeed, the technologically honed minds of our technocrats are simply unwilling or unable to guide us into the future. (Imagine autonomous robots with the power to decide our future running on MS Windows or an Android-like OS.) The future they foresee is more of the same but with a different shade, a definite improvement without any glitches. (The usual message is that any glitches will be fixed in a subsequent release, but the problem Lem indicates is that there may be no time for such a release.) Our technology leaders always offer us progress, whatever this may mean in practice. To really explore what may be coming, and what new technologies may bring to us, we need the mind to become unhinged, like that of Lem. Lem's visions of future societies oriented around AI and robotics explore AI technology in a deeper and more nuanced way than the current technological visions from our leading technological prophets. What is coming, based on Lem's visions, may not be the AI-driven nirvana in the dreams of Schwab (2016) or Kurzweil (2015).

While Lem's fantastic worlds are societies of AI automata, at their foundations, they are deeply human with human suffering, desire, pleasure, values, sins, problems, and challenges. This is why we may regard Lem's robotic societies as potential human societies in a future digital world following the technological singularity (e.g., Kurzweil 2005, Chalmers 2010, Talaga 2021, and

[12] The Future of Life website claims that "Most researchers agree that a superintelligent AI is unlikely to exhibit human emotions like love or hate, and that there is no reason to expect AI to become intentionally benevolent or malevolent. Instead, when considering how AI might become a risk, experts think two scenarios most likely: The AI is programmed to do something devastating…The AI is programmed to do something beneficial, but it develops a destructive method for achieving its goal." Available at https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/> The Future of Life institute is obviously missing the nefarious effects of AI systems that are already deployed (e.g., social networks), proposing some "maybe" and "possibly" scenarios at some unspecified future time. This all treats AI as being distant, unspecific, and less real.

many others). Lem points out that technology is always unpredictable and fallible, and it always has some effects and applications that we cannot foresee. There are also always errors and malfunctions in such systems. Technology, as our creation, is never perfect, so it is bound to fail sooner or later in some unpredictable way at some unknown place and time, or it may do things that were not planned (i.e., unforeseen). What is more, the more complex and all-encompassing technology becomes, the more complex and significant its failures will be. Technology has a hidden dark side, and it will always enslave us in some way.[13] We design technology to help us in life, but technology itself changes us, and with each technological enhancement, we lose a part of ourselves. Now, Lem asks this: What do we gain in exchange? What would happen if we lost agency, and when would we even realize that we had lost it? To avoid completely irreversible disasters, whether foreseen or unforeseen, Lem says we need a red button and someone to control it. Technological eschatology does not offer any kind of paradise but rather promises the dehumanization of reality. In Lem's *Cyberiad* stories, the red button was Trurl and Klapaucius, who are god-like constructors and engineers, but their ilk are not among us today. If we do not have them, we may be heading for a disaster on a cosmic or at least planetary scale. In the best case, we could dehumanize ourselves, raiding our essential humanity in return for unclear promises of technological progress.

Bibliography

Art that capture the eye and the mind (2021), [online], [accessed 2021-09-02], available at: <https://www.artaigallery.com/ >.

Parable of Monkey (2021), [online], [accessed 2021-09-02], available at: <https://www.angelfire.com/in/hypnosonic/Parable_of_the_Monkeys.html >.

The 6 Levels of Vehicle Autonomy Explained (2021), [online], [accessed 2021-09-02], available at: <https://www.synopsys.com/automotive/autonomous-driving-levels.html>.

Ahlstrom-Vij, K. (2013): Why We Cannot Rely on Ourselves for Epistemic Improvement, in *Epistemic Paternalism: A Defence*. Palgrave Macmillan, 6–38.

Bolter, J. D. (1984): *Turing's Man: Western Culture in the Computer Age*, University of North Carolina Press.

Borgo, S. – Vieu L. (2009): Artefacts in Formal Ontology, in Gabbay, D. et al. (eds.) *Philosophy of Technology and Engineering Sciences*, North Holland, 273–307.

Bostrom, N. (2014): *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press.

Brockman, J. (2020): *Possible minds. 25 ways of looking at AI*, Penguin Books.

[13] As much as we possess technology, technology also possesses us. On the complex relations between technology, society, and individuals one may read, to start with, Heidegger's *The Question Concerning Technology* (Heidegger 1954), which is a rather dated but still relevant work, as well as more modern studies like, for example, Lazzarato's *Capital Hates Everyone* (Lazzarato 2021). In the *Cyberiad*, Lem makes the same point.

Bush, J. (2021): *Predictions in technology are almost always wrong*, [online], [accessed 2021-09-02], available at: <https://www.electronicspecifier.com/industries/industrial/predictions-in-technology-are-almost-always-wrong>.

Chalmers, D. (2010): The Singularity: A philosophical analysis, *Journal of Consciousness Studies* 17 (9(10)), 7–65.

Eckersley, P. – Sandberg, A. (2013): Is Brain Emulation Dangerous?, *Journal of Artificial General Intelligence* 4(3) 170–194.

Elgin, C.Z. (2013): Epistemic Agency, *Theory and Research in Education* 11(2), 135–152.

Gans, J. (2018): *AI and the paperclip problem*, [online], [accessed 2021-09-02], available at: <https://voxeu.org/article/ai-and-paperclip-problem>.

Gibbs, S. (2014): Facebook apologises for psychological experiments on users, *The Guardian*, 2014-7-2, [online], [accessed 2021-09-02], available at: <https://www.theguardian.com/technology/2014/jul/02/facebook-apologises-psychological-experiments-on-users>.

Hart, M. (2020): Google's New AI Helps You Write Poetry Like Poe. *Nerdist* [online], [accessed 2021-09-02], available at: <https://nerdist.com/article/google-ai-writes-poetry-like-legendary-poets/>.

Hayworth, K. (2010): *Killed by bad philosophy: Why brain preservation followed by mind uploading is a cure for death*, [online], [accessed 2021-09-02], available at: <http://www.brainpreservation.org>.

Heidegger, M. (1977): The Question Concerning Technology, in Heidegger, M., *Basic Writings*, Harper & Row.

Houkes, W. (2009): Introduction. In Gabbay D. et al. (eds.) *Philosophy of Technology and Engineering Sciences*, North Holland.

Kaiser. B. (2019): *Targeted*, Harper Collins Publishers.

King, G. – Pan, J. – Roberts, M. E. (2017): How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument. *American Political Science Review* 111(3), 484–501.

Kobiela, F. – Gomułka (eds.) (2021): *Filozoficzny Lem*, Wydawnictwo Alethea.

Koene, R. A. (2006): Scope and resolution in neural prosthetics and special concerns for the emulation of a whole brain, *The Journal of Geoethical Nanotechnology* 1, 21–29, [online], [accessed 2021-09-02], available at:

<https://www.terasemjournals.org/GNJournal/GN0104/koene_01a.html>.

Kornblith, H. (2012): *On Reflection*, Oxford University Press.

Krzanowski, R. (2020): Why can information not be defined as being purely epistemic? *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)*, 68, 37–62.

Krzanowski, R. (2021): *Ontological Information: Investigation into the properties of ontological information*, PhD thesis, Pontifical University of John Paul II in Kraków, [online], [accessed 2021-09-02], available at: <http://bc.upjp2.edu.pl/dlibra/docmetadata?id=5024>.

Kurzweil, R. (2005): *The Singularity is Near*, Viking Books.

Larkin, B. (2018): *30 Craziest Predictions About the Future Experts Say Are Going to Happen*, [online], [accessed 2021-09-02], available at: <https://bestlifeonline.com/crazy-future-predictions/>.

Lau J. H. – Cohn, T. – Baldwin, T. – Hammond, A. (2020): *This AI Poet Mastered Rhythm, Rhyme, and Natural Language to Write Like Shakespeare*, [online], [accessed 2021-09-02], available at: <https://spectrum.ieee.org/this-ai-poet-mastered-rhythm-rhyme-and-natural-language-to-write-like-shakespeare#toggle-gdpr>.

Lazzarato, M. (2021): *Capital Hates Everyone*, The MIT Press.

Lem, S. (2014) [1974]: *The Cyberiad*, Penguin Books.

Lovelock, J. (2020): *Novacene: The Coming Age of Hyperintelligence*, Penguin Books.

Maher, B. (2012): Controversial H5N1 influenza work likely to resume, *Nature* [online], [accessed 2021-09-02], available at: <https://doi.org/10.1038/nature.2012.12089 >.

Millett, P. et al. (eds.) (2016): *Gain-of-function research: Summary of the second symposium, March 10-11, 2016*, The National Academies Press.

Milmo, D. (2021): Facebook pauses work on Instagram Kids after teen mental health concern. *The Guardian*, 2021-9-27, [online], [accessed 2021-09-02], available at: <https://www.theguardian.com/technology/2021/sep/27/facebook-pauses-instagram-kids-teen-mental-health-concerns>.

Mineo, L. (2017): When it comes to internet privacy, be very afraid, analyst suggests: Surveillance is the business model of the internet,' Berkman and Belfer fellow says, *The Harward Gazette*, 2017-8-24, [online], [accessed 2021-09-02], available at: <https://news.harvard.edu/gazette/story/2017/08/when-it-comes-to-internet-privacy-be-very-afraid-analyst-suggests/>.

Olson, D. (2015): A Case for Epistemic Agency, *Logos and Episteme* VI (4), 449–474.

Pestov, I. (2017): *The absolute worst technology predictions of the past 150 years,* [online], [accessed 2021-09-02], available at: <https://www.freecodecamp.org/news/worst-tech-predictions-of-the-past-100-years-c18654211375/ >.

Pogue, D. (2012): Use It Better: The Worst Tech Predictions of All Time, *Scientific American,* [online], [accessed 2021-09-02], available at: <https://www.scientificamerican.com/article/pogue-all-time-worst-tech-predictions/>.

Puzzo, A. (2015): *Against Epistemic Agency*, MA Thesis, Queen's University, Kingston, Ontario, [online], [accessed 2021-09-02], available at: https://qspace.library.queensu.ca/bitstream/handle/1974/13727/Puzzo_Andrew_201509_MA.pdf; sequence=1.

Robert. P. (2021): La société des asociaux, *Le Monde diplomatique*, 68(810), 28, [online], [accessed 2021-09-02], available at: https://www.monde-diplomatique.fr/2021/09/RIMBERT/63484.

Rogners, A. (2017): The Way the World Ends: Not with a Bang But a Paperclip, *The Wired*, [online], [accessed 2021-09-02], available at: <https://www.wired.com/story/the-way-the-world-ends-not-with-a-bang-but-a-paperclip/>.

Sandberg, A. – Bostrom, N. (2008): *Whole Brain Emulation: A Roadmap,* Technical Report #2008-3, Future of Humanity Institute, Oxford University, [online], [accessed 2021-09-02], available at: <https://www.fhi.ox.ac.uk/brain-emulation-roadmap-report.pdf>.

Schab, K. (2016): *The Fourth Industrial Revolution*, Portfolio Penguin.

Schlosser, M. (2019): Agency, in Zalta, E.N. (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition), [online], [accessed 2021-09-02], available at: <https://plato.stanford.edu/archives/win2019/entries/agency/>.

Selgelid, M. (2016): Gain-of-Function Research: Ethical Analysis, *Science and Engineering Ethics* 22 (4), 923–964, [online], [accessed 2021-09-02], available at: <https://dx.doi.org/10.1007%2Fs11948-016-9810-1>.

Shanahan, M. (2015): *The Technological Singularity*, The MIT Press.

Swirski, P. (ed.) (1997): *A Stanisław Lem Reader: Rethinking Theory*, Northwestern University Press.

Talaga, N. (2021): Don't Worry About The AI Singularity: The Tipping Point Is Already Here, *Forbes*, 2021-6-21, [online], [accessed 2021-09-02], available at: <https://www.forbes.com/sites/nishatalagala/2021/06/21/dont-worry-about-the-ai-singularity-the-tipping-point-is-already-here/?sh=27a78a2c1cd4>.

Tand, D. (2020): *The Machines Are Coming, and They Write Really Bad Poetry*, [online], [accessed 2021-09-02], available at: <https://lithub.com/the-machines-are-coming-and-they-write-really-bad-poetry/>.

The Rockefeller Foundation (2010): *Scenarios for the Future of Technology and International Development*. The Rockefeller Foundation and Global Business Network.

Thomasson, A. (2009): Artefacts in Metaphysics, in Gabbay D. et al. (eds.) *Philosophy of Technology and Engineering Sciences*, North Holland.

Wylie, C. (2019): *Mindf*ck*, Profile Books.

Zuboff, S. (2019): *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, Profile Books.

Zuckerberg, M. (2017): Building global community, [online], [accessed 2021-09-02], available at: *https://www.teachthought.com/education/mark-zuckerbergs-manifesto/*.