



AI is not a Tool

The Impact of Growing AI Agency on the Future of Work

Alexander van Biezen¹  

¹Arcadia University, Belgium, alexander.vanbiezen@arcadiascholen.be 

Abstract

Research Question (RQ): What are the underlying philosophical assumptions shaping current perceptions of artificial intelligence (AI) as a mere tool, and how do these assumptions influence our understanding of AI's growing agency and its potential impact on the future of work?

Purpose: The paper aims to critically examine the widespread assumption that AI systems remain passive instruments entirely under human control. It explores how emerging forms of AI agency—understood as autonomous or semi-autonomous decision-making capacities—challenge this notion and what implications this shift entails for human labour, ethics, and social stability.

Methods: The study adopts a philosophical and conceptual methodology grounded in the philosophy of mind and the philosophy of science. It draws on classical thought experiments (Searle's Chinese Room, Jackson's Mary, Penrose's arguments on non-algorithmic consciousness) and integrates recent interdisciplinary debates on AI agency, autonomy, and consciousness. The analysis is based on a critical literature review combining philosophical, technological, and socio-political sources.

Results: Findings indicate that the assumption of AI as a "dumb tool" no longer holds. Evidence of growing AI autonomy demonstrates that decision-making processes once reserved for humans are increasingly being delegated to machines. This outsourcing of human agency risks creating social and ethical blind spots, potentially leading to unequal labour transformations and governance challenges. However, a managed transition toward human-AI cooperation could foster innovation and inclusion if grounded in ethical oversight and policy regulation.

Organization: For organizations, the study highlights the need to anticipate shifts in work structures and decision-making processes caused by AI systems with growing agency. It encourages managers and policymakers to design governance frameworks that maintain human oversight while enabling responsible collaboration with AI.

Society: At the societal level, the research underlines the urgency of open policy debates and ethical reflection on AI regulation. Addressing the implications of AI autonomy is essential to preserve human agency, democratic accountability, and social justice in the digital era.

Originality: The article contributes to bridging philosophical inquiry and socio-technical analysis by reframing AI not merely as a technological tool but as an emerging actor in human decision-making systems. It advances the concept of "AI agency" as a key lens for understanding the transformation of work.

Limitations / Further research: The study is conceptual and does not include empirical data. Future research should investigate how organizations and workers experience AI agency in practice, possibly through ethnographic or organizational case studies, and explore policy instruments capable of mitigating risks related to automation and technocratic governance.

Keywords: artificial intelligence, AI agency, AI consciousness, workforce, future, philosophy of science, philosophy of mind.

1 Introduction

Sometimes reality catches up with us faster than expected. The past few years, each time when I mentioned the possibility of AI systems developing ‘agency’ to my students, I was met with blank stares or frowning. Today, ‘AI agent’ has become a new buzzword overnight. At the beginning of March 2025, a report from *The Information* announced that OpenAI, one of the leading AI research organizations, may be planning to charge up to \$20,000 per month for specialized AI ‘agents’ (Palazzo & Weinberg, 2025). The prices vary from \$2,000 a month for an AI agent at the level of a ‘high-income knowledge worker’ (Wiggers, 2025), a software developer agent will cost about \$10,000 a month, and at the top we find a ‘Phd-level research’ (Edwards, 2025) agent which will cost you no less than \$20,000 a month.

The etymological origin of ‘agency’ stems from the Latin verb *agere* which means “to do”, “to act”. The word ‘agent’ stems from the present participle *agens, agentis*, “one who acts” or “one who does an act” (i.e. an agent). The topic of agency has a much longer history in the field of ethics, predating the advent of AI. Giving a precise definition of what characterizes ‘agency’ is quite a challenging task for philosophy, as it is related to intentional action and goal-directed behavior and, eventually, to the intricate philosophical question of *what it means to be a person*.

According to a recent study from PULSE (Program on Understanding Law, Science, and Evidence) from UCLA School of Law, there is still a significant lack of agreement on the definition of agency and, as a result, there is still no consensus whether or not it is even possible to consider AI systems as agents (Newman et al., 2025). “(...) Depending on the perspective and definition (...) the agency of AI could be controversial, unimaginable, or an unquestionable truth (...)” (Newman et al., 2025). Especially the question whether or not an AI system can be considered as a goal-oriented entity remains controversial (Newman et al., 2025).

For our discussion at hand, we take the option of slipping through the horns of this dilemma. As a working definition of ‘agency’ with regard to AI in this article, we simply mean that some AI systems are getting to a whole new level of decision-making capabilities and autonomous action. Whether or not these current AI systems can really be said to have their own goals and are truly goal-oriented is not the key question in this respect. What matters here for our discussion at hand is that these AI systems are acquiring a level of autonomous decision-making which makes it possible and tempting for us, humans, *to transfer a growing part of human decision-making to these systems*.

Be that as it may, one thing’s for sure: developments in AI are making giant leaps at an ever-increasing pace. In September 2024, the well-known historian Yuval Noah Harari published *Nexus* on information networks from early history to AI today. In *Nexus*, Harari warns us: “AI

isn't a tool – it's an agent" (Harari, 2024, p. XXII), meaning that AI is capable of processing information all by itself, and thereby has the capacity to replace humans in the making of decisions (Harari, 2024, p. XXII). Harari was heavily criticized for this warning, being put away by some as a doomsday prophet. For instance, Don Lim, a seasoned IT specialist and formerly Chief AI Developer at Vindex, was very quick to reproach Harari in his article *Why Yuval Noah Harari's AI Doomsday Prophecies Are Misleading* (Lim, 2024) that he is being both alarmist and misinformed:

“He also argues “AI is not a tool, but an agent.” The current AI systems are far from autonomous entities; they are tools created, monitored, and controlled by humans. The idea that AI will become a self-governing force beyond human control is closer to science fiction than reality.” (Lim, 2024)

Some reproached him that his book *Nexus* is “based on shallow scholarship” (Ferguson, 2025), or even bluntly suggested that the real lesson we can learn from Harari is that “there's an incredible amount of money to be made with doomsday predictions” (Foreman, 2024).

But even Geoffrey Hinton, the so-called godfather of AI, who shared the 2024 Nobel Prize in Physics with John Hopfield for their foundational discoveries and inventions that enable machine learning with artificial networks, warns us that “AI systems may be more intelligent than we know and there's a chance the machines could take over” and “we're moving into a period when for the first time ever we may have things more intelligent than us” (Pelley, 2024). He even left Google a year before, in May 2023, precisely because of his concerns about the many risks of AI (Douglas Heaven, 2023).

If even the very brightest and the most well-versed among us in the field of AI yield warnings about the rapid developments in AI and its possible unforeseen devastating consequences, maybe it is time to listen to what they have to say and, at least, to postpone our judgement, even if only for a moment.

Vincent Ginnis, professor of mathematics, physics and artificial intelligence at my alma mater, the Free University of Brussels (Vrije Universiteit Brussel), and at Harvard University, has become very concerned about the disconcerting lack of concern about the possible dangers of AI. In an eye-opening opinion piece (Ginnis, 2025) he mentions that at a recent AI safety conference in Paris, no one seemed to be concerned any longer with the dangers of AI. Instead, *it was all about PR and power struggles*. Once, Ginnis states, AI safety was about risks for humanity. The possible threat that millions of people might lose their jobs, the danger that misinformation and manipulation would spread at a scale undermining democracy, the possibility that AI systems might one day take decisions we no longer comprehend, let alone control. Instead, at the conference in Paris it was all about power. Who will get AI? Who

controls it? Who is running ahead in the race? The focus, according to Ginnis, shifted from risks to geopolitics. He issues a firm, unequivocal warning:

“Humanity is creating a technology that surpasses its own knowledge and is completely unprepared for it. The first step is simple: acknowledge what is happening. The threat is real, the acceleration is dangerous, and the priorities are misplaced. There is still a long way to go, but we have to start somewhere.” (Ginnis, 2025, *my translation*)

Likewise, Koen Schoors, professor of economy at the University of Ghent (Belgium), points to the danger of the growing attraction of an AI-based technocracy, not hampered by the sluggishness of democratic decision-making. There is no shortage of politicians, he states, who are fed up with the inertia of the democratic model with regard to the developments in AI (Schoors, 2024, p. 189).

The question is not about whether or not or to what extent Harari, Hinton and others are right about their warnings. What intrigues me the most, as a philosopher of science, is: *why are so many people so adamant and resolute in brushing aside all these warnings? Why do we cling so tightly to the reassuring idea that AI is just a mere tool, totally under our control? What are the tacit assumptions which apparently make it very hard for us to find the blind spots in our rosy vision on AI?*

2 Literature review: AI and philosophy of mind

2.1 The Chinese room thought experiment

When I was still a graduate student in philosophy, way back in the 1980s, I remember we were discussing John Searle’s thought experiment of the Chinese room. Remember that 40 years ago, AI was still a very remote theoretical possibility, a popular theme in science fiction movies perhaps, but not something to be taken very seriously. Nevertheless, some philosophers gave it their attention, more often than not in order to concoct sophisticated arguments to show that artificial intelligence would not be possible. A machine, a computer, could never have a mind in the same way human beings can be said to have minds.

In brief, John Searle’s argument (Searle, 1980 and 1984) goes as follows. Someone is sitting in a room and receives a sheet of paper with Chinese characters from the left, through a slit in the wall. This person then meticulously follows a highly detailed instruction table, akin to a computer program, to transform the Chinese messages into other messages with different characters on another sheet of paper. Once the conversion is complete, this person then sends this new sheet of paper out to the right as output (likewise, through a slit in the wall). To an outside observer, it seems as if the person inside the rooms *understands* Chinese. However, in

reality, the person inside the room is just following instructions, he does not need to understand a single word of Chinese himself.

Searle wanted to show with this thought experiment that two people can be *functionally identical*: a native Chinese speaker and the person inside the Chinese Room. They both provide perfect answers to questions posed in Chinese. Yet, they have completely different *mental states* (one understands Chinese, while the other does not understand it at all). The bottom line of Searle's argument is: a computer will remain fundamentally different from a human being. A computer system will never “truly” understand what it is doing, whereas human beings obviously can.

2.2 Mary in the black-and-white room

Another famous example in this respect comes from the Australian philosopher Frank Jackson, who in 1986 published a very controversial and influential article, *What Mary didn't know* (Jackson, 1986). Even though it is almost forty years old by now, it is about a thought experiment which is still used today in discussions about the possibility of *artificial general intelligence* (AGI).

In short, Jackson's thought experiment goes as follows. This experiment is about a fictional scientist in a distant future, Mary. In that distant future, both physics and neurophysiology have reached a final state. That means that Mary knows everything these sciences have to say about perceiving colors.

But there is something special about Mary: she lives in a completely *colorless* room. Everything is black or white. So, Mary has never seen a color before in her entire life. Her knowledge of colors is therefore purely based on books about physiology, neurology, and the biochemistry of color perception.

Now it gets exciting. One day, Mary finds a secret door to the outside world. The first thing she sees when she steps outside is a *red apple*. The question Jackson poses is: *does Mary learn something new when she sees this apple?*

Jackson answers: yes, she learns a *new fact* — she learns what it is like to *experience* the color 'red'. This phenomenon is referred to in philosophy of mind as '*qualia*': 'individual instances of subjective, conscious experience' (*qualia* is Latin and is the plural of *quale*, which literally means 'such as').

Jackson aims to conclude that knowledge about *qualia* fundamentally relies on *subjective experiences*, unlike knowledge about physical states in our brains. By subjective, we mean how someone experiences or judges something from a personal perspective.

Although Jackson's knowledge argument was originally meant to question the philosophical position of *physicalism*, in short, the view that "everything is physical", that there is nothing "above" the physical (van Biezen, 2016), this thought experiment has also become a classic in discussions about artificial general intelligence (Wang, 2023); Baron, 2025; Renard, 2024). What it boils down to is that Jackson's argument about Mary in the black-and-white room is used to demonstrate that there will always be an unbridgeable gap between the human mind and artificial intelligence. In short, AI is Mary stuck in the black-and-white room, never having seen a color in her entire life. The human being is Mary when she walks into the outside world and sees a red apple.

2.3 The Copernican trauma

Somehow, I cannot shed the impression that these arguments resonate with older arguments pertaining to the difference between human beings and other animals. Time and again, throughout the ages, people have tried to pinpoint demarcation criteria to demonstrate that there is a fundamental difference *in kind* between human beings and animals, and certainly not a difference *in degree*. Part of the upheaval caused by the publication of Darwin's *On the Origin of Species* (Darwin, 1859) in 1859 was precisely that: the horrifying idea that human beings are just another animal in the tree of evolution. It seems that the gist of these arguments resurfaces in the discussions about the nature of artificial intelligence. Benjamin Bratton, philosopher of technology and professor at the University of California, San Diego, sees AI as the next phase in the series of Copernican human decentering of the once thought privileged position of human beings in the center of the world, a "Copernican trauma" (Bratton, 2024):

"What is today called 'artificial intelligence' reveals that intelligence, cognition and even mind (...) are not what they seem to be, not what they feel like and not unique to the human condition (...) *intelligence itself is artificializable.*" (Bratton, 2024)

2.4 Roger Penrose on minds and machines

One of the most famous defenders of the position that there is a fundamental chasm between the human mind and machines is the esteemed mathematician and mathematical physicist Sir Roger Penrose, who shared the 2020 Nobel Prize in physics for his ground-breaking work on black hole formation in the 1960s (which he accomplished together with Stephen Hawking who passed away in 2018, so, unfortunately, he could not be nominated anymore). As early as 1989, in his book *The Emperor's New Mind* (Penrose, 1989), Penrose argued that human consciousness is essentially *non-algorithmic*, which implies that human consciousness in principle could never be embedded by what is called a conventional Turing machine, i.e. the theoretical model which underlies every possible computer.

Penrose's argument is based on the so-called *first incompleteness theorem* by Kurt Gödel, undoubtedly one of the most important results in the foundations of mathematics of the twentieth century (for a comprehensive account, see Nagel & Newman, 1958). Put very simply: this theorem gives a mathematical proof that any consistent and sufficiently rich axiomatic system of ordinary arithmetic with natural numbers (0, 1, 2, 3,...) and basic arithmetic functions like addition (+) and subtraction (-) will always contain true statements about natural numbers which cannot be proved nor disproved *within that formal system itself*. In other words, these statements are *true*, but they cannot be derived from the axioms (that is, by a step-by-step procedure, i.e. an algorithm).

Penrose's argument boils down to his claim that, contrary to that formal system, a sufficiently skilled mathematician (i.e. a human mind) is indeed capable of arriving at and formulating those true statements (which are unprovable and underivable from within that system). According to Penrose, Gödel's first incompleteness theorem tells us that no computer which works within a formal system F can prove the sentence

$G(F) = \text{"This sentence cannot be proved in F."}$

But, Penrose continues, we humans can just "see" the truth of $G(F)$. Because if $G(F)$ were false, then it would be provable, which leads to a paradox, an absurd result. So, the human mind is capable of doing something which not a single computer can do. Therefore, Penrose concludes, consciousness can't be reducible to computation. This implies that machines can never have human-like consciousness and, therefore, artificial intelligence would be forever out of human reach.

Needless to say, there has been a lot of discussion on Penrose's argument, but that would be far beyond the scope of this article (for a brave attempt at refuting Penrose's use of Gödel's theorems, see Krajewski, 2015). The only reason why I bring it up here is the question: why is one of the smartest persons on this planet so persistent in trying to prove that the human mind will always "surpass" the capabilities of any possible machine?

The tacit assumption underlying all these approaches seems to be that artificial general intelligence systems should possess *consciousness* in order to be called "truly" intelligent. All these thought experiments and arguments serve the same strategy: machines, artificial intelligence, cannot possibly attain consciousness or "self-awareness", if you'd like, and, hence, machines, artificial intelligence, cannot possibly be called "really" intelligent.

2.5 "Can machines think?" (Alan Turing)

It is exactly this tacit assumption that intelligence somehow presupposes consciousness that I want to put into question. This is crucial to my argument, so I want to state it as explicitly as I can. For some reason or another, whatever it might be, some people still see computers as

inanimate, dumb machines which are incapable of thinking or feeling anything. And because computers cannot think or feel anything, they are not capable of making any decisions on their own.

This leads us to one of the key questions in the field of philosophy of artificial intelligence: the fascinating cross-border field between philosophy of mind and the philosophy of computer science (Brey & Søraker, 2009) that explores the implications of artificial intelligence for getting grips on concepts like ‘intelligence’, ‘consciousness’, ‘free will’ etc. (for a comprehensive account of the current state of affairs in the philosophy of artificial intelligence, see Müller, 2025). This key question is: *can something like artificial intelligence exist at all?* As Alan Turing, the father of the computer and of AI, put it succinctly: *can machines think?* (Turing, 1950, p. 442). Turing himself found this question simply “too meaningless” to deserve discussion:

“The original question, "Can machines think?" I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted” (Turing, 1950, p. 442)

The interesting question is *not* ‘Can machines think?’, but ‘Can we let machines do things for which human beings would need intelligence?’. Saying that AI systems cannot “truly” think, that is to say, think as human beings, is the equivalent of saying that planes cannot “truly” fly because they can’t flap their wings.

Admittedly, whether or not AI can really be called intelligent or not is still a controversial debate. Philosopher Luciano Floridi traces the issue down to a conceptual issue: either we enlarge our definition of what intelligence actually is so that it also includes artificial forms of it, or we widen our conception of agency so that it also encompasses artificial forms of agency which do not necessarily presuppose intelligence (Floridi, 2025). Floridi is in favor of the last option, viewing AI as “agency without intelligence” (Floridi, 2025).

Others are more clearly opposed to AI agency and its implications, considering AI systems as intelligent. Even as early as 2009, scholar Joanna Bryson unequivocally warned against the danger of humanising robots and declared that “robots are fully owned by us” (Bryson, 2009, p. 1). She argues: “In humanising them, we not only further dehumanise real people, but also encourage poor human decision making in the allocation of resources and responsibility” (Bryson, 2009, p.1).

Why, you might ask, is that so important? Because this reassuring and soothing position that AI cannot possibly be called “really” or “truly” intelligent entails the risk of creating a dangerous blind spot. As long as we keep on seeing AI systems as a mere tool, as a “dumb” instrument, passively waiting for human instructions to do something, we put ourselves in danger by

overlooking the fact that these systems are getting more and more *agency*, that is to say, more and more autonomy and decision-making capabilities. If those rapidly increasing decision-making capabilities of AI systems remain in our blind spot, we run a serious risk of continuing to outsource human decision-making to the point where it is no longer under our control. And this blind spot is related to the deeply ingrained tacit assumption that intelligence presupposes consciousness. As Yuval Harari strikingly noted, there is a widespread conviction that computers and AI systems simply are not capable of making decisions assumes that making decisions is predicated on having consciousness (Harari, 2024, p. 201). However, Harari continues, the fact that in human beings, as in other mammals, intelligence is often accompanied by consciousness does not allow us to extrapolate from humans and other mammals to all possible entities (Harari, 2024, p. 201).

Admittedly, the issue of what 'consciousness' precisely is remains very slippery. Nevertheless, as scholars Patrick Krauss and Andreas Maier recently pointed out, most biologists nowadays consider consciousness as a gradual phenomenon which, in different levels of complexity, can also be found in animals (Krauss & Maier, 2025). In this respect, according to the integrated information theory (Giulio Tononi), the level of consciousness, the level of consciousness depends on the structure of the underlying substrate (i.e. the brain, for humans and other animals). The more coherent or connected a system is, the more conscious it is. In short, consciousness is related to the mutual interconnectedness of a system (Krauss & Maier, 2025).

To cut all these ramifications short, *let us simply erase 'consciousness' from the equation*. The rapid evolution of AI compels us to radically rethink our understanding of concepts like 'intelligence' and 'consciousness'. As Blaise Agüera y Arcas and James Manyika put it succinctly: "We're in paradigm-shifting territory" (Agüera y Arcas & Manyika, 2025). Meanwhile, we should focus on reducing our blind spot and getting past this pernicious stumbling block of keeping on seeing AI as a mere tool, a mere instrument which cannot operate without human supervision and which we have under our control.

3 Methodology

In this article, although I bring in different elements from various disciplines, the methodology followed is that of a philosophical argumentation, as the central thread of the questions examined are most closely related to philosophical issues in the field of philosophy of mind.

4 Findings: The risk of our blind spot

4.1 AI gaining in agency

AI gaining in agency, becoming more and more autonomous and capable of independent decision-making, is not about Terminator-like robots rampaging through the streets in a killer

spree. As Mustafa Suleyman, the co-founder of DeepMind, one of the leading AI research laboratories, already pointed out in his book *The Coming Wave* (Suleyman & Bashkar, 2023):

“Many technologies and systems are becoming so complex that they’re beyond the capacity of any one individual to truly understand them (...) In AI, the neural networks moving toward autonomy are, at present, not explainable. You can’t walk someone through the decision-making process to explain precisely why an algorithm produced a specific prediction.”
(Suleyman & Bashkar, 2023)

When it comes to artificial *general* intelligence (AGI), the risk of the blind spot impeding us to see the real potential of danger becomes even greater. When bringing up the mere possibility of artificial general intelligence, you easily get your share of disbelief and laughter, conjuring up the famous image of the super-intelligent HAL 9000 computer on board of the spaceship U.S.S. Discovery in the motion picture *2001: A Space Odyssey* (1968, Stanley Kubrick), refusing to open the air-lock, in that very calm, soothing tone: “I’m sorry Dave, I’m afraid I can’t do that”.

By brushing aside the possibility of artificial general intelligence, we run the risk of turning a blind eye to the *real* dangers of the rapid development of current AI technology. The position we are in right now reminds me of the horrifying images of the 2004 tsunami in South-East Asia. Remember the footage of the receding water from the beaches, with people laughing at this strange phenomenon, small boats, sloops, suddenly lying dry at the sea floor, maybe some fish flopping in a remaining shallow pool, children still playing around. Even when they saw the wall of water looming in the distance, people’s warning systems still didn’t seem to be triggered, they curiously kept on staring in the distance, instead of deciding to run to higher grounds as quickly as possible.

Lest I be dismissed as another doomsday prophet, I want to add a critical note to the rosy picture of AI being hailed as the next Industrial Revolution, lifting the future of the human workforce to a whole new level. A bit like in the lyrics of that song of Timbuk 3, “The future’s so bright, I gotta wear shades”. Admittedly, there are quite a number of economic and labor studies highlighting the risks of automation and AI for skills mismatch and workforce displacement. However, voices pleading for some caution are easily drowned out as too overly pessimistic.

4.2 “AI is set to surpass us in speed and understanding” (Geoffrey Hinton)

Moreover, there are some recent developments taking an ugly turn. At a recent conference, Geoffrey Hinton made a very interesting point, which strengthens me in my argument that we are put on the wrong track, set on the wrong foot, so to speak, when we continue to convince ourselves that AI systems are just dumb tools, not capable of “really” understanding what they are doing. Hinton’s point was that even with today’s large language models (LLMs), there is a

key difference between them and the way human memory works, and that is “AI’s unmatched ability to share knowledge” (Saso, 2025). Human beings pass information in small pieces, whereas AI has the ability to synchronize trillions of bits in the blink of an eye:

“(...) It’s no competition,” he said. If intelligence is about learning and sharing knowledge, AI is set to surpass us in speed and understanding. It’s a “very scary conclusion,” Hinton said—a warning that highlights the need for consensus on AI’s capabilities. (...)” (Saso, 2025)

In a recent paper in *Science* on how to manage extreme AI risks amid rapid technological progress, Hinton warns unequivocally:

“(...) Increases in capabilities and autonomy may soon massively amplify AI’s impact, with risks that include large-scale social harms, malicious uses, and an irreversible loss of human control over autonomous AI systems. (...)” (Bengio, Hinton et al., 2024)

Now, let us not get carried away by popular science fiction ideas of evil AI systems taking over the world. The danger of AI I want to talk to you about does *not* come from systems like HAL 9000, the super-intelligent computer, eliminating all humans on board of the U.S.S. Discovery. The danger I want to shed a light on is coming from *us*. It is *we* who are putting ourselves in danger because *we* are turning a blind eye to the growing *agency* of AI systems. And we are doing so because we are held captive by an image, the image of an AI system as a mere tool, the iconic image of a computer as a simple box with a keyboard. Because we tacitly assume that agency presupposes consciousness, and since we are firmly convinced that AI systems cannot possibly acquire consciousness, they will not acquire autonomous agency, so, there is nothing to worry about, isn’t it?

However, reality is catching up fast with us. Even now, as we speak, AI systems analyze tremendous amounts of data, they take decisions in fractions of a second, more and more without any human intervention. Just think about all the sophisticated algorithms which manage the content feed on social media. Think about the trading by algorithms on financial markets, medical diagnostics driven by AI systems. The point is: we rely more and more on AI systems to make decisions that once were the sole province of human beings.

4.3 The need for vigilance

As we are getting more and more comfortable towards these new technologies, as we are embracing them more deeply, we tend to become less vigilant over the technology, not to say downright lazy with regard to the use of all those systems. The principle of ‘*least effort*’ is a strong predictor of human behavior (Anderson & Rainie, 2023). It makes me think of my former life, when I was still working as a computer scientist for IT companies. In those days, you had to

be able to write sophisticated and elegant search queries in SQL-type languages to retrieve relevant information from a database. Nowadays, we just shout “Hey, Google!”, followed by a simplified query, often a very inelegant one, *and we accept the result at face value, without giving it a second thought*. The near infinite capacity of human beings to take things for granted will never fail to amaze me. Aren’t we lulling ourselves to sleep too quickly?

A striking example of how this lazy attitude, blindly accepting what AI systems regurgitate, might lead us straight into disaster has been reported recently by Bastian Leibe, professor at the RWTH Aachen University in Germany (Leibe, 2025). When President Donald Trump from the United States announced his reciprocal tariffs on April 2nd, 2025, a number of people noticed that these proposed tariffs are not related to the *actual* tariffs these countries charge on imports from the United States. Instead, they have been shown to correspond to the United States’ trade deficit divided by the United States import volume from that country, which, according to economists, does not make any sense at all from an economic perspective (Leibe, 2025).

Admittedly, at the time of writing of this article (early April 2025), president Trump had just announced his tariff plan. No scientific analyses or academic articles on this topic had been published as yet, as it was just discovered and signaled by a few academic scholars who are well versed in the field, like Bastian Leibe. Since April 2nd, a lot of economists have been trying to find out how on earth someone could come up with such an insane strategy. Until, Leibe continues his argument, someone found out that if you pose the question of tariff tables to current LLMs (like ChatGPT version 4o, Gemini 2.5pro et al.), they all propose tariffs which turn out to be very close to the tariffs in the list of president Trump. Leibe concludes that the most likely conclusion is that the Trump administration simply based its tariffs “on the unchecked outputs of an LLM” (Leibe, 2025):

“This has real-world consequences. It is already sending economies into turmoil and it will cause worldwide harm and suffering. And it is sadly nothing that AI safety research could have prevented -- because the problem lay in front of the screen.” (Leibe, 2025)

We see it happening all around us, as we speak. Society is becoming more and more complex. We are delegating human decision-making to sophisticated systems which involve storing our digital data and automating decision rules at an ever-increasing pace.

In other words, we are lured into outsourcing our cherished decision-making and autonomy to AI systems step by step, and as these smart systems, powered by sophisticated machine-learning, will rapidly augment their sophistication level in the next, say, ten years from now, we might lose the ability to keep on making decisions independently of these systems. Barry

Chudakov, founder and principal of Certain Research, sees the relationship between human beings and AI systems as

“(…) a struggle between the determined fantasy of humans to resist (‘I’m independent and in charge and no, I won’t give up my agency!’) and the seductive power of technology designed to undermine that fantasy (‘I’m fast, convenient, entertaining! Pay attention to me!’)” (…)” (Anderson & Rainie, 2023)

5 Debate: Towards a Human–AI Symbiosis

What is the way forward? Is the future so bleak as some of these findings tend to suggest? No, I think not. But it is time that we realize that we will have to face difficult questions, now that there is still time to do so. *What are the things we, as human beings, really want agency over?* What should we list as the conditions under which we will turn to AI to help us in making decisions? And, the most pernicious issue, under what conditions and tight control mechanisms are we prepared to outsource certain precisely defined decisions to AI systems? We do not have the luxury of turning a blind eye to these difficult questions.

A very promising view on a possible way forward is offered by Somendra Narayan, a professor of Strategy and Innovation at the University of Amsterdam, The Netherlands. His idea is that if we want to come to grips with understanding the impact of AI on human agency, it is helpful to see this interaction in terms of a *symbiosis* (Narayan, 2024). Instead of looking at AI as “an external force eroding our autonomy”, it would be better to look at it in terms of “an augmentation of human capability, a co-evolution where humans and machines are learning from each other and influencing each other’s decision-making” (Narayan, 2024). Narayan admits: yes, AI presents certain risks for human agency, but also opportunities. In order to ensure that AI augments human autonomy instead of diminishing it, we must build systems “where transparency and ethical considerations are central to how AI operates” (Narayan, 2024). His core argument is: *human agency does not exist in isolation*. Human agency is part of a larger socio-technical system. Human decisions do not arise in isolation, they are shaped by both technological tools and societal structures. In his view, AI is simply the most recent layer in this system.

Tyler Suard, an AI researcher and developer, formerly at Apple and Meta, places a critical note on this sunny idea of a fruitful AI-human combination. He gives several examples of experiments where AI turned out to perform better on its own than when working together in a combined human-AI team. He concludes: if AI can indeed outperform human-AI teams, massive job loss will be lurking around the corner (Tyler, 2024). He calls for a reality check: we need to prepare for that possibility where AI will be a powerful independent entity in the workforce (Tyler, 2024). He pleads for governments and organizations to start working on policies

and regulations that address potential job displacement. He thinks about “social safety nets, retraining programs, and incentives for industries that create new job opportunities” (Tyler, 2024).

I want to join this positive but cautious attitude towards the future of the human workforce (Boudry & Friederich, 2024). No, these developments certainly do not mean the end of human work. However, we must see to it that the transition to human-AI cooperation is not too disruptive, as this could be a threat to an inclusive and just division of labor, possibly even destabilizing society (van Biezen, 2024). In the past few years, the *World Economic Forum* stresses time and again in its annual jobs report that both analytical thinking and creative thinking are considered as the most wanted core skills in order to face the challenges of the oncoming transformation of the labor market (van Biezen, 2024, p. 58). In other words, companies focus on very high-level profiles with abstract cognitive skills as critical for that “brave new world” (Huxley, 1932). But what about the medium-level and low-level occupations? We need to be aware of the urgency of scaffolding social protection and support for those who run the risk of being forced out.

The research group *Inclusive Society* (department Research & Expertise) at UCLL University of Applied Sciences (Leuven, Belgium) attaches a lot of importance to inclusion when it comes to the future of the human workforce. We focus on the question: *how can we contribute to building a more inclusive and fairer world?* From this perspective, we plea for the need of continuing policy debates on AI regulation taking explicitly into account the phenomenon of growing AI agency.

6 Conclusions

The study has shown that the prevailing assumption of artificial intelligence (AI) as a mere tool—an inert instrument incapable of independent reasoning or decision-making—no longer holds in light of current technological developments. Through a philosophical and conceptual analysis, it was demonstrated that AI systems are gaining increasing levels of autonomy and agency, allowing humans to outsource parts of their decision-making to machines. This process, if left unchecked, risks creating social and ethical blind spots with potentially destabilizing consequences for labour structures and democratic decision-making. At the same time, the transition toward human–AI cooperation presents opportunities for innovation, inclusion, and new forms of collaboration—provided it is governed by ethical oversight and clear regulatory frameworks.

The paper contributes to the interdisciplinary dialogue between philosophy, technology studies, and the social sciences by reframing the debate on AI not in terms of technical functionality but in terms of agency. It advances a conceptual bridge between classical philosophical arguments about consciousness and contemporary issues of algorithmic autonomy, offering a framework

for understanding AI as an emergent actor within socio-technical systems. This philosophical contribution strengthens the conceptual foundations of ongoing debates on AI governance and the future of work.

For management and organizations, the findings underline the importance of recognizing AI systems as active participants in decision-making processes rather than as passive tools. Managers and policymakers are encouraged to develop governance models that balance efficiency with ethical responsibility, ensuring human oversight, transparency, and accountability in the implementation of AI technologies. On a broader societal level, the paper highlights the need for continued public and policy debates on AI regulation, particularly regarding its implications for inclusion, social justice, and democratic control.

The study is theoretical in nature and does not include empirical or quantitative data. Its findings are based on philosophical reasoning and conceptual synthesis, which limits its direct applicability to specific organizational contexts. The rapid evolution of AI technologies also means that some empirical examples may quickly become outdated.

Future research should empirically investigate how AI agency manifests in real organizational and societal contexts. Comparative studies across industries or sectors could provide insight into how different forms of AI autonomy affect human decision-making, trust, and accountability. Interdisciplinary research combining philosophy, organizational studies, and AI ethics would help to refine theoretical models and translate them into practical guidelines for governance.

References

1. Agüerra y Arcas, B. & Manyika, J. (2025). *AI Is Evolving – And Changing Our Understanding of Intelligence*. In *Noema*, Berggruen Institute,
2. <https://www.noemamag.com/ai-is-evolving-and-changing-our-understanding-of-intelligence/>.
3. Anderson, J. & Rainie, L. (2023). *The Future of Human Agency*. In *Pew Research Center*, <https://www.pewresearch.org/internet/2023/02/24/the-future-of-human-agency/>.
4. Baron, S. (2025). *Are a Machine's Thoughts Real? The Answer Matters Now More Than Ever*. In *Science Alert*, <https://www.sciencealert.com/are-a-machines-thoughts-real-the-answer-matters-now-more-than-ever>.
5. Bengio, Y, Hinton, G. et al. (2024), *Managing Extreme AI Risks Amid Rapid Progress*, in *Science*, Vol. 384, Issue 6698, pp. 842-845.
6. Boudry, M. & Friederich, S. (2024). *The Selfish Machine. On the Power and Limitation of Natural Selection to Understand the Development of Advanced AI*. In *Philosophy of Science*, PhilSci-Archive, preprint, <https://philsci-archive.pitt.edu/23903/>.

7. Boudry, M. (2025). *The Selfish Machine. Will Humanity Be Subjugated by Superintelligent AIs?*. In Maarten Boudry's Substack, <https://maartenboudry.substack.com/p/the-selfish-machine>.
8. Bratton, B. (2024). *The Five Stages of AI Grief*. In *Noema*, Berggruen Institute, <https://www.noemamag.com/the-five-stages-of-ai-grief/>
9. Brey, Ph. & Johnny H. Søraker, J. H. (2009). *Philosophy of Computing and Information Technology*. In *Philosophy of Technology and Engineering Sciences*, edited by Antonie Meijers, Amsterdam, Elsevier, pp. 1341–1407.
10. Bryson, J. (2009). *Robots Should Be Slaves*. Published at Joanna Bryson Publications, <https://www.joannabryson.org/publications/robots-should-be-slaves-pdf>.
11. Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life*. New York, D. Appleton and Company, 1861 (first edition 1859).
12. Dhondt, S. & Dessers, E. (eds.) (2022). *Robot zoekt collega*. Uitgeverij Lannoo. [In Dutch; English title: *Robot seeking colleague*].
13. Douglas Heaven, W. (2023). *Deep learning pioneer Geoffrey Hinton quits Google*. In *MIT Technology Review*, <https://web.archive.org/web/20230501125621/https://www.technologyreview.com/2023/05/01/1072478/deep-learning-pioneer-geoffrey-hinton-quits-google/>.
14. Edwards, B. (2025). *What does "PhD-level" AI mean? OpenAI's rumored \$20,000 agent plan explained*. In *Ars Technica*, <https://arstechnica.com/ai/2025/03/what-does-phd-level-ai-mean-openais-rumored-20000-agent-plan-explained/>.
15. Ferguson, N. (2025). *The Doom Nexus*. In *Niall Ferguson's Time Machine*, <https://niallferguson.substack.com/p/the-doom-nexus>.
16. Floridi, L. (2025). *AI as Agency without Intelligence: On Artificial Intelligence as a New Form of Artificial Agency and the Multiple Realisability of Agency Thesis*. February 12, 2024. Available at <http://dx.doi.org/10.2139/ssrn.5135645>.
17. Foreman, J. T. (2024). *How to Make it as a Doomsday Prophet*. In *The Metaphor*, <https://www.taylorforeman.com/p/how-to-make-it-as-a-doomsday-prophet>.
18. Ginnis, V. (2025). *Is er nog iemand bekommerd om de gevaren van AI?* In *De Standaard*, 15 February 2025, https://www.standaard.be/cnt/dmf20250214_96655287 [In Dutch; English title: *Is there still anyone concerned about the dangers of AI?*].
19. Harari, Y. N. (2024). *Nexus. A Brief History of Information Networks from the Stone Age to AI*. Vintage Publishing, Kindle Edition.
20. Huxley, A. (1932). *Brave New World*. Pdf edition, Coradella Collegiate Bookshelf, 2004, <http://collegebookshelf.net>.
21. Jackson, F. (1986). *What Mary Didn't Know*. In *The Journal of Philosophy*, Vol. 83, No. 5 (May, 1986), pp. 291-295.
22. Kahneman, D. (2011). *Thinking Fast and Slow*. New York, Farrar, Straus and Giroux.

23. Krajewski, S. (2015). *Penrose's Metalogical Argument is Unsound*. In Ladyman, J. et al. (eds.)(2015). *Road to Reality with Roger Penrose*. Kraków (Poland), Copernicus Center Press, p. 87-104.
24. Krauss, P. & Maier A. (2025). *De geest in de machine*. In *EOS Psyche & Brein*, June 2025, pp. 20-25 [In Dutch, English translation of the title: *The Ghost in the Machine*].
25. Ladyman, J. et al. (eds.)(2015). *Road to Reality with Roger Penrose*. Kraków (Poland), Copernicus Center Press.
26. Leibe, B. (2025). *Post on LinkedIn*,
<https://www.linkedin.com/feed/update/urn:li:activity:7313873939691130880/>.
27. Lim, D. (2024). *Why Yuval Noah Harari's AI Doomsday Prophecies Are Misleading*. In *Medium*, <https://medium.com/@don-lim/why-yuval-noah-hararis-ai-doomsday-prophecies-are-misleading-5541504ec3ab>.
28. Molek, N., Pulinx, R. & van Biezen, A. (eds.)(2024). *Analysis of the State of the Art on the Future of Human Workforce. Scientific Report*. Transform, European Union.
29. Molek, N., van Biezen, A. & Velez, M. J. (2025), *Book of Abstracts. International Interdisciplinary Conference Transform "The Future of Human Workforce"*. Novo Mesto (Slovenia), FOS.
30. Müller, V. (2025). *Philosophy of AI. A Structured Overview*. In Smuha, N. (ed.)(2025). *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence*. Cambridge University Press, p. 40-58.
31. Nagel, E. & Newman, J. R. (1958). *Gödel's Proof*. New York, New York University Press.
32. Narayan, S. (2024). *AI and the Future of Human Agency: Are We Outsourcing Decision-Making or Evolving with Machines?*. In *Medium*,
<https://medium.com/@narayan.somendra/ai-and-the-future-of-human-agency-are-we-outsourcing-decision-making-or-evolving-with-machines-78da6ba4475f>.
33. Newman, S. et al. (2019). *AI & Agency*. In *2019 Summer Institute on AI and Society*, in *AI Pulse*, 26 September 2019, <https://aipulse.org/ai-agency/?pdf=417>.
34. Palazzolo, S. and Weinberg, C. (2025). *OpenAI Plots Charging \$20,000 a Month For PhD-Level Agents*. In *The Information*, <https://www.theinformation.com/articles/openai-plots-charging-20-000-a-month-for-phd-level-agents>.
35. Pelley, S. (2024). *"Godfather of Artificial Intelligence" Geoffrey Hinton on the promise, risks of advanced AI*. In *CBS News*, <https://www.cbsnews.com/news/geoffrey-hinton-ai-dangers-60-minutes-transcript/>.
36. Penrose, R. (1989). *The Emperor's New Mind. Concerning Computers, Minds and The Laws of Physics*. Oxford, Oxford University Press.
37. Renard, V. et al. (2024). *Mary Steps Out: Capturing Patient Experience through Qualitative and AI Methods*. In *NEJM AI*, Vol. 1 No. 12, <https://ai.nejm.org/doi/10.1056/Alp2400567>.

38. Sapunov, G. (2023). *Turing, "Intelligent Machinery. A Heretical Theory", 1951*. In *Gonzo ML*, https://gonzoml.substack.com/p/turing-intelligent-machinery-a-heretical?utm_campaign=post&utm_medium=web.
39. Saso, E. (2025). *The path to safe, ethical AI: SRI highlights from the 2025 IASEAI conference in Paris*. In *Schwarz Reisman Institute for Technology and Society*, University of Toronto. <https://srinstitute.utoronto.ca/news/the-path-to-safe-ethical-ai>.
40. Satyanarayan, A. and Jones, G. M. (2024). *Intelligence as Agency: Evaluating the Capacity of Generative AI to Empower or Constrain Human Action*. In *An MIT Exploration of Generative AI - From Novel Chemicals to Opera*, <https://mit-genai.pubpub.org/pub/94y6e0f8/release/2>.
41. Searle, J. (1980). *Minds, Brains and Programs*. In *Behavioral and Brain Sciences*, 3, pp. 417-517.
42. Searle, J. (1984). *Minds, Brains and Science*. Cambridge, Mass., Harvard university press.
43. Schoors, K. (2024). *Alles wordt anders*. Gent, Borgerhoff & Lamberigts. [In Dutch; English title: *Everything Will Be Different*]
44. Smuha, N. A. (ed.)(2025). *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence*. Cambridge University Press.
45. Suard, T. (2024). *The Future of Work: AI May Not Need Us After All*. In *Medium*, https://medium.com/@ceo_44783/the-future-of-work-ai-may-not-need-us-after-all-5df8eae52ed9.
46. Suleyman, M. & Bhaskar, M. (2023). *The Coming Wave. Technology, Power, and the Twenty-First Century's Greatest Dilemma*. New York, Crown.
47. Turing, A. (1950). *Computing Machinery and Intelligence*. In *Mind*, 49, pp. 433-460.
48. Turing, A. (1951). *Intelligent Machinery. A Heretical Theory*. <https://gwern.net/doc/ai/1951-turing.pdf>.
49. von Hoffman, C. (2025). *Smarter AI means bigger risks – Why guardrails matter more than ever*. In *MarTech*, <https://martech.org/smarter-ai-means-bigger-risks-why-guardrails-matter-more-than-ever/>.
50. van Biezen, A.F. (2016). *A Case for Naturalism*. In van Biezen, A.F., *The Torch of Discovery*, <http://alexanderfvanbiezen.blogspot.com/2016/05/a-case-for-naturalism.html>.
51. van Biezen, A.F. (2022). *Top-Down Cosmology and Model-Dependent Realism. A Philosophical Study of the Cosmology of Stephen Hawking and Thomas Hertog*. Brussels, VUB Press.
52. van Biezen, A. (2024). *Emerging Skills for the Future Workforce*. In Molek, N., Pulinx, R. and van Biezen, A. (eds.)(2024), *Analysis of the State of the Art on the Future of Human Workforce. Scientific Report.*, Transform, European Union, p. 50-62.
53. Van Biezen, A. (2025a). *Abstract of 'AI is Not a Tool'*. In Molek, N., van Biezen, A. & Velez, M. J. (2025), *Book of Abstracts. International Interdisciplinary Conference Transform "The Future of Human Workforce"*. Novo Mesto (Slovenia), FOS, p. 8.

54. van Biezen, A.F. (2025b). *AI is not just another tool. What keeps us in the blind spot?*. In van Biezen, A.F., *The Torch of Discovery*, <https://alexanderfvanbiezen.blogspot.com/2025/04/ai-is-not-just-another-tool.html>.
55. Verbinnen, L. (2025), *AI-gebruik stijgt, maar ook onze bezorgdheid: 'Techno-optimisme maakt plaats voor technorealisme'*. In *EOS Wetenschap*. [In Dutch; English title: AI usage rises, but so does our concern: 'Techno-optimism gives way to tech realism'.] https://www.eoswetenschap.eu/technologie/ai-gebruik-stijgt-maar-ook-onze-bezorgdheid-techno-optimisme-maakt-plaats-voor?utm_source=ActiveCampaign&utm_medium=mail&utm_campaign=eos_515.
56. Walther C.C. (2025). *Hybrid Intelligence: The Future of Human-AI Collaboration*. In *Psychology Today*, <https://www.psychologytoday.com/us/blog/harnessing-hybrid-intelligence/202503/hybrid-intelligence-the-future-of-human-ai-collaboration>.
57. Wang, X. (2023). *The Possibility of Artificial Qualia*. In *Communications in Humanities Research*, <https://doi.org/10.54254/2753-7064/6/20230083>.
58. Wiggers, K. (2025). *OpenAI reportedly plans to charge up to \$20,000 a month for specialized AI 'agents'*. In *TechCrunch*, <https://techcrunch.com/2025/03/05/openai-reportedly-plans-to-charge-up-to-20000-a-month-for-specialized-ai-agents/>.

Alexander van Biezen is a philosopher of science, graduated from the Free University of Brussels (Vrije Universiteit Brussel) as a Doctor of Philosophy and Moral Sciences. He specialized in the philosophy of cosmology, with a doctoral dissertation on the cosmological models of Stephen Hawking and Thomas Hertog. His book *Top-Down Cosmology and Model-Dependent Realism* (2022, VUB Press) is freely available online through the research portal of the Free University of Brussels (<https://researchportal.vub.be/>). He has an additional background in religious studies and in computer science. Currently, he is employed as a teacher of philosophy and religion with the Arcadia school group in Aarschot, Belgium. His main areas of interest are cosmology and the philosophy of artificial intelligence.

Alexander van Biezen je filozof znanosti, doktoriral je na Svobodni univerzi v Bruslju (Vrije Universiteit Brussel) kot doktor filozofije in moralnih znanosti. Specializiral se je za filozofijo kozmologije, z doktorsko disertacijo o kozmoloških modelih Stephena Hawkinga in Thomasa Hertoga. Njegova knjiga *Top-Down Cosmology and Model-Dependent Realism* (2022, VUB Press) je prosto dostopna na spletu preko raziskovalnega portala Svobodne univerze v Bruslju (<https://researchportal.vub.be/>). Ima tudi dodatno izobrazbo na področju religijskih študij in računalništva. Trenutno je zaposlen kot učitelj filozofije in religije v skupini šol Arcadia v Aarschotu v Belgiji. Njegova glavna področja zanimanja sta kozmologija in filozofija umetne inteligence.

Povzetek**UI ni zgolj orodje****Vpliv naraščajoče tvornosti umetne inteligence na prihodnost dela**

Raziskovalno vprašanje (RV): Katere so temeljne filozofske predpostavke, ki oblikujejo sodobno razumevanje umetne inteligence (UI) kot zgolj orodja, in kako te predpostavke vplivajo na naše dojetanje naraščajoče tvornosti UI ter njenega možnega vpliva na prihodnost dela?

Namen: Članek kritično preučuje razširjeno domnevo, da sistemi UI ostajajo pasivna orodja, popolnoma pod nadzorom človeka. Raziskuje, kako nove oblike tvornosti UI – razumljene kot avtonomne oziroma polavtonomne sposobnosti odločanja – izpodbijajo to predstavo in kakšne posledice ima ta premik za človeško delo, etiko in družbeno stabilnost.

Metoda: Raziskava uporablja filozofsko in konceptualno metodologijo, utemeljeno v filozofiji duha in filozofiji znanosti. Opira se na klasične miselne poskuse (Searlov “kitajski sobi”, Jacksonovo “Mary v črno-beli sobi” in Penroseove argumente o nealgoritmični zavesti) ter vključuje sodobne interdisciplinarne razprave o tvornosti, avtonomiji in zavesti UI. Analiza temelji na kritičnem pregledu literature, ki združuje filozofske, tehnološke in družbenopolitične vire.

Rezultati: Ugotovitve kažejo, da predpostavka o UI kot »neum-nem orodju« ne vzdrži več. Dokazi o naraščajoči avtonomiji UI potrjujejo, da se postopki odločanja, ki so bili nekoč izključno v domeni človeka, vse pogosteje prenašajo na stroje. Takšno postopno prenašanje človeške tvornosti lahko povzroči družbene in etične slepe pege ter vodi do neenakih transformacij dela in izzivov upravljanja. Vendar pa lahko nadzorovan prehod k sodelovanju med človekom in UI spodbuja inovativnost in vključenost, če temelji na etičnem nadzoru in ustreznih regulativnih okvirih.

Organizacija: Za organizacije raziskava poudarja potrebo po pravočasnem predvidevanju sprememb v delovnih strukturah in procesih odločanja, ki jih povzročajo sistemi UI z naraščajočo tvornostjo. Menedžerje in oblikovalce politik spodbuja k oblikovanju upravljavskih okvirov, ki ohranjajo človeški nadzor in hkrati omogočajo odgovorno sodelovanje z UI.

Družba: Na družbeni ravni raziskava poudarja nujnost odprtega političnega in etičnega dialoga o regulaciji UI. Naslavljanje posledic avtonomije UI je ključno za ohranjanje človeške tvornosti, demokratične odgovornosti in družbene pravičnosti v digitalni dobi.

Originalnost: Članek prispeva k povezovanju filozofske refleksije in družbeno-tehnične analize, saj ponovno opredeljuje UI ne zgolj kot tehnološko orodje, temveč kot nastajajočega akterja v sistemih človeškega odločanja. Razvija koncept »tvornosti UI« kot osrednjo analitično perspektivo za razumevanje preobrazbe dela.

Omejitve/nadaljnje raziskovanje: Raziskava je konceptualne narave in ne vključuje empiričnih podatkov. Nadaljnje raziskave bi morale empirično preučiti, kako organizacije in delavci v praksi doživljajo tvornost UI – na primer s pomočjo etnografskih ali organizacijskih študij primerov – ter raziskati politične in regulativne instrumente, ki bi lahko omilili tveganja, povezana z avtomatizacijo in tehnokratskim upravljanjem.

Ključne besede: umetna inteligenca, tvornost UI, zavest UI, delovna sila, prihodnost dela, filozofija znanosti, filozofija duha.



This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/).



This journal is published by [Faculty of Organisation Studies in Novo mesto](https://www.fos.unm.si/).