Trufanova, E. O. 2025. “Trustworthiness and Responsibility as the Key Issues of the AI Application: Human in the Loop of Responsibility” [in English]. *Filosofiya. Zhurnal Vysshey shkoly ekonomiki [Philosophy. Journal of the Higher School of Economics]* 9 (4), 105–122.

Elena Trufanova*

Trustworthiness and Responsibility as the Key Issues of the AI Application**

Human in the Loop of Responsibility

Submitted: Aug. 05, 2025. **Reviewed:** Oct. 19, 2025. **Accepted:** Nov. 01, 2025.

Abstract: The article attempts to answer two questions: (1) What can we trust artificial intelligence (AI) with? (2) Who is responsible for the decisions that we entrusted the AI to make? It is shown that the use of the AI to solve various tasks seems attractive, on the one hand, due to its speed and simplicity, which imply economic benefits for users, and on the other, due to the assumed objectivity and accuracy inherent in intelligent machines, as opposed to subjective and error-prone people. It is demonstrated that the simplicity and speed of using the AI are far from always beneficial, and the accuracy and objectivity of the AI are illusory. It is proved that the reliability of the AI can be regarded as high only for a number of specific narrow tasks. It is shown that it is impossible to talk about the responsible AI, since responsibility is a property of the subject, and the AI is not a subject. Considering AI systems as independent subjects or agents can lead to the formation of a technocracy, where decisions will be made by technical systems, but responsibility for them will not be assigned to anyone. It is proved that in matters concerning the life and well-being of people, their freedom and other basic human values, decisions can be made only by a human and only a human will be responsible for them. The advantage of a human expert in solving such issues is seen in the intersubjective perception of another person and empathy, that are unobtainable for the AI. Based on his own human experience, an expert is able to see those features of a specific problem that an artificial system cannot take into account. It is concluded that the problem of ethical, trustworthy and responsible AI is not a technical one, but a social one — it is a problem of how a person can ethically and responsibly use such a powerful and complex tool as AI. Ethics and responsibility are human properties that cannot be delegated to artificial systems.

Keywords: Trustworthy Artificial Intelligence, Responsible Artificial Intelligence, Responsibility, Human-in-the-Loop, Intersubjectivity.

DOI: 10.17323/2587-8719-2025-4-105-122.

The discussion around artificial intelligence (AI) is more than half a century old, but it changed its course only a few years ago with the introduction of the generative AI (Gen AI) systems. While the previous discussion was

*Elena Trufanova, Doctor of Letters in Philosophy; Leading Research Fellow at the RAS Institute of Philosophy (Moscow, Russia), iph@etrufanova.ru, ORCID: 0000–0002–2215–1040.

concerned more with the question of "what AI can and cannot do?" especially "can it reach the level not only of human intelligence but of human *consciousness?*" the most recent question is "what AI should be *allowed* to do?" The theoretical questions were pushed aside by the practical ones. I will be using the term "artificial intelligence" (AI) also in this practical meaning — the AI as various artificial intellectual systems that exist nowadays (and among those I will be mostly discussing Large Language Model systems (LLMs)), not the hypothetical artificial general intelligence that might or might not emerge in the indefinite future.

The digital technologies applications have been hastily forced upon present societies as a result of the pandemic and the self-isolation politics accompanying it. Many digital novelties have been introduced at this time and very soon became commonplace due to the need to avoid direct human-to-human contacts. For example, online conference systems like the ones we are often using nowadays have already existed for a while, but only at the beginning of the 2020s did they become a common practice. As this need arose unexpectedly, neither the society nor the individuals were fully prepared for the introduction of the new technologies and the changes they brought with them. The same can be said about the use of the AI. It is a well-known Marxist principle that productive forces development runs ahead of the development of the relations of production (Marx, 1971), and this is exactly the situation we are facing now, when we have the technologies on hand that have the power to change the social relations, but the society is not ready for the change and needs the time and effort to adapt.

There is a lot of alarmist talk lately concerning the AI, even from the innovative technologies' powerful leaders like Elon Musk and Steve Wozniak, who in 2023 were among the thousand subscribers to the open letter calling to pause for at least 6 months all the training of the AI systems whose powers and abilities exceed ChatGPT-4 (Future of Life Institute, 2023). The concerns voiced in this open letter are mostly about the unpredictability of the "black box" self-developing AI systems that might present a threat to the society and, on a grander scale of things, to humanity on the whole.[1] Even Sam Altman, the founder of OpenAI, the company that developed ChatGPT, has called on the US senators to impose stricter regulations on AI development, ensuring the safety of the products developed by the AI

[1] Whether as a reaction to this letter or on other grounds, the ChatGPT-5 has been introduced only in August 2025, and received a lot of negative feedback from its first users.

companies (O'Brien, 2023). This is an interesting precedent when the technology developer pleads with the government to stop them. Though Altman has recently proclaimed that we are past the event horizon of the "gentle singularity," and by the 2030 we shall live in the new world completely transformed by the AI. He admits that there are still challenges to confront and safety issues to solve but nevertheless suggests welcoming the upcoming change (Altman, 2025). Although Atlman's recent statement seems partly utopian and partly self-promoting, we can agree that although the alarm caused by the AI proliferation is well-grounded, it still should not make us spill the baby with the bathwater. The AI is a technological instrument we can and will use, but we first should learn how to use it the right way.

TEMPTATION OF THE AI

The use of the AI is temping on different levels.

First of all, it is new, fast, and easy to use. With the rapid development of the GenAI systems in the last couple of years they have become a tool everyone wants to apply in different fields — from language translations to legal decision-making and medical design. The seeming easiness and effectiveness of the AI application in certain problem-solving is too tempting, because fast and simple solutions are always sought for. Economically speaking, Gen AI supposedly saves time and money when we meet up with the easy tasks, such as, for example, designing a company logo or generating a typical tourist-oriented description of a "seaside paradise" hotel, etc. The machine translations into different languages are also widely used, since their quality increased drastically with the introduction of the LLMs when compared to the first automated translation attempts only a couple of decades ago. And of course the use of the brand-new technologies strengthens any advertising campaign: if you do not use AI, you are likely to appear outdated — which is never good for business. This is why we shall without doubt see a big increase in the use of the GenAI in the upcoming years in different types of businesses. We shall consider later in the paper, however, that the introduction of AI across various business sectors has not proceeded as smoothly as hoped.

Secondly, the AI systems often have the reputation of accuracy and objectivity. The AI is considered a "machine," and people are psychologically prone to trust the objectivity of the machines (in our case — intellectual machines) over the subjectivity of other humans. L. Daston and P. Gallison describe in their famous work how the introduction of photography in the second half of the 19th century has changed the idea of scientific objectivity: mechanical registration of the visible event, "the view from nowhere," starts to be considered the true objective and precise view of things in comparison to the subjective and fallible view of the scientist (Daston & Galison, 2007). We are used to the fact that *errare humanum est* ("to err is human"), yet we often overlook or downplay the possibility of machine failures — or of human misuse of machines. Humans do not only err, they lie, they cheat, they are prejudiced, and they seek their own gain, so we put our hopes and prayers into the AI: our natural distrust of the other humans makes us trust the AI systems because they seem more reliable than humans in many different ways.

Thus, people are tempted to use AI for both economical and psychological reasons, and the present GenAI systems are becoming more and more user-friendly and easy-to-use so that these systems are "accessible for all." Naturally, the AI systems are of big interest for the political actors as well, facilitating the bureaucracy and providing the tools for the realization of technocracy politics. This is a technocracy in the most literal form, where not the science experts but the autonomous technical systems themselves might play a crucial part in the decision-making.

There is also a problem that is very accurately formulated by Russian philosopher of science Natalia Yastreb: "the convenience and effectiveness of AI tools lead to their value being taken for granted. As a result, when artificial intelligence is introduced, it is not the tools that are embedded in social systems and practices, but the social systems themselves that adapt to AI-based functioning. There is a kind of shift from motive to the goal. What should help to cope with the tasks begins to change the tasks themselves" (Yastreb, 2025: 101). This is an important observation, because there is indeed a tendency to make people adapt to the new technology instead of adapting a technology to meet human needs. Right now we see the AI being introduced in some of the spheres where we never really needed it in the first place.

Thus, I suggest that the usage of AI in many different spheres of our lives is inevitable, but the two main questions we should pose when we are using it are:

(1) What can we trust the AI with?

(2) Who is responsible for the decisions that we entrusted the AI to make?

ETHICAL AI, TRUSTWORTHY AI, RESPONSIBLE AI

As we mentioned before, the current AI research switched from the area of theoretical philosophy to the area of practical philosophy, and the most discussed in recent philosophical and overall public debate about the AI are probably the questions of AI ethics. There are different concepts that are in use that are somewhat difficult to differentiate: ethical AI, trustworthy AI, and responsible AI. These concepts are deeply related and are sometimes used more or less as synonymous. The ideas behind these terms come mostly from the attempts at public regulation of the AI usage. As one of the originators of the philosophy of information and of the digital ethics Luciano Floridi mentions, there are more than 70 different lists of ethics principles for the AI (Floridi, 2019). As this was written in 2019, the number has probably doubled since then, and some authors express well-grounded concern that anyone can choose any one of those according to his specific needs or tastes (Yastreb, 2025; Floridi, 2019). Some of the most important of these documents that aim for a global status were issued by the European Commission and UNESCO. The European Commission's document called "Ethics Guidelines for Trustworthy AI" was issued in 2019. It was developed by the specially appointed high-level experts group (AI HLEG), with the Floridi among those. This document names three key principles of trustworthy AI: it should be lawful (abide by the laws), ethical (respect ethical principles), and robust (be accurate and safe) (AI HLEG, 2019). UNESCO issued in 2022 "Recommendation on the Ethics of Artificial Intelligence" (UNESCO, 2022) that outlines the principles of ethical usage of AI for the UNESCO Member States. UNESCO defines AI trustworthiness as an essential element that ensures that AI works for the good of the society and humans and underlines that to be regarded as trustworthy, "throughout their life cycle, AI systems are subject to thorough monitoring by the relevant stakeholders as appropriate" (ibid.: 18).

Floridi, in his paper supporting and explaining the importance of the AI HLEG's work, insists that although these guidelines are not yet legally enforced, this does not make them useless. Rather he maintains that we should adopt an ethics-first approach to pave the way for the legislation in the AI domain (Floridi, 2019). He also warns that "'innovate first, fix later' is a mistake that, in the case of AI, could also be very costly and may cause a public backlash against AI" (ibid.: 262), so we should develop and discuss those principles right now, no matter how advanced current AI systems are.

While these documents speak mostly about trustworthy and ethical AI, the concept of responsible AI is widely mentioned in the current discussions. This concept is also ambiguous. Recent research suggests that responsible AI principles can be divided into "accountability, diversity, non-discrimination and fairness, human agency and oversight, privacy and data governance, technical robustness and safety, transparency, and social and environmental well-being" (Papagiannidis et al., 2025). Most of these principles are also not univocal and need to be explained as well (see, for example, a paper on accountability (Novelli et al., 2024)). And we can see that all the principles listed by Papagiannidis et al. could as well be attributed to the trustworthy AI. I suggest that responsible AI can be understood from two different viewpoints attributed to different agents. On the one hand, responsible AI can be regarded as a part of the Responsible Research and Innovation (RRI) approach, which means that humans should develop and apply the AI systems responsibly (human agents should be responsible). On the other hand, responsible AI can be regarded as an AI agent accountable for certain actions (the AI agent should be responsible).

As we have seen above, the concepts of ethical, trustworthy, and responsible AI are always overlapping. In my research I will try to discern different spheres of application of trustworthy AI and responsible AI.

I suggest regarding the problem of *trustworthy AI* as an epistemological problem: can we trust AI's "knowledge"? Is it accurate and reliable? Is AI more objective than humans? Can AI be a useful and trusted instrument in our cognition? What can AI do successfully without human oversight, if anything?

As for the *responsible AI*, I suggest that it is an ethical and legal problem: how can we teach AI to follow ethical values and principles? How can we ensure that AI does not inherit the bias and prejudice of its developers and teachers? How can we provide for AI to do no harm? Who is accountable for the AI's actions?

I will now tackle these two different spheres in more detail in the following paragraphs.

CAN WE TRUST THE AI AT ALL?

The first thing we should keep in mind when deciding what to trust the AI with is that the image of flawlessly objective and accurate machines is a myth. Even when we are speaking about the technologies far more primitive than the AI. The saying goes, "the camera never lies," but even when the photo itself is not manipulated with it might show a very specific angle or a very specific bit of the whole wide picture that misrepresents the reality. This, of course, can be explained by the human misuse of the camera, because the human photographer chooses the angle, but when we take as an example a self-tracking CCTV camera, we might encounter the same misinterpreted view with no human to blame. Thus, the machines are not flawless.

Intelligent machines, as well as intelligent human beings, can be wrong. Even an inexperienced user of the AI can very soon learn that the AI does not only make mistakes, but it also makes things up — sometimes due to the lack of certain information in its databases, sometimes due to inexplicable reasons. This became widely known as AI hallucinations, though the term does not seem to be quite on spot; another term — confabulation — suggested by tech journalist Benji Edwards seems more plausible (Edwards, 2023). Both terms are anthropomorphic, but confabulation is a more accurate metaphor. Confabulation means there is a gap in memory that a person fills with fake "recollections," and when we speak about the AI, the system encounters a gap in the pool of information available to it and fills it with whatever it can come up with, following the algorithm pattern. The AI systems are now being taught to give honest "I don't know" answers, but the AI hallucinations/confabulations problem is still not completely resolved. Thus, it is reasonable to agree with the conclusion cited by B. Edwards:

> ...ChatGPT as it is currently designed, is not a reliable source of factual information and cannot be trusted as such. "ChatGPT is great for some things, such as unblocking writer's block or coming up with creative ideas," said Dr. Margaret Mitchell, researcher and chief ethics scientist at AI company Hugging Face. "It was not built to be factual and thus will not be factual. It's as simple as that" (ibid.).

The AI-generated answers should not be treated as a trusted source of information, because they can be misleading in a most dangerous way — when complete nonsense is hidden among the vast amounts of accurate facts and reasonable arguments.

There is also a question of objectivity. Every AI system is taught upon certain databases that represent certain worldviews. For example, if posed with some controversial political questions, ChatGPT (OpenAI, Inc., USA), GigaChat (Sber, Russian Federation), and DeepSeek (Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co., Ltd., PRC) will give different answers that reflect the respectful official political positions of the countries where the systems were developed. Some of the questions will be avoided by certain of the AI systems because of their built-in "ethical" restrictions, which themselves reflect particular cultural and political positions.

Prejudice and bias can also be found within AI algorithms. For example, LLMs can be more likely to ascribe domestic roles to women and business jobs to men based on the analysis of texts that show the commonness of such inequality (UNESCO & IRCAI, 2024). The natural languages retain certain prejudices of the societies that are using them, and LLMs that learn those languages assimilate those prejudices as well. The AI can even become the source of discriminating decisions of its own—as, for example, interesting research on political orientation-based discrimination shows (Peters, 2022). Philosopher from Utrecht University Uwe Peters demonstrates that while there are ethical and social laws prohibiting race or gender discrimination, there are no such rules against certain kinds of social identity discrimination, for example, the one based on political orientation. He suggests a situation where the AI is used in the job recruitment, and while assessing a candidate for the job that presumably has conservative views, decides against them based on the statistics that show that conservative-oriented workers have been underrepresented in the company and probably thus are unfit for this job. The irony is that a human recruiter would probably have no idea about the political orientation of the job applicant (unless the applicant voices this fact on their own accord), but the algorithms can find certain clues in the applicant's personal data that might lead to the conclusion about their political orientation and to rejecting their application on these grounds. Peters calls it "algorithmic discrimination" (ibid.).

Hence, we should come to the conclusion that the AI is not accurate when we consider factual information, nor is it objective or unbiased when making decisions. The AI algorithms are taught on the specifically chosen data sets, just as humans are taught on specifically chosen textbooks and develop under the influence of a certain social environment, and both parties come out of the education process biased in some way. Therefore, if we are seeking impartial judgment or a reliable source of information, we should not place our hopes in AI.

CAN THE AI BE RESPONSIBLE?

The idea of responsible AI is tricky. Responsibility and accountability are the characteristics of the subject. The subjects act upon their reasoning and free will and thus have responsibility for their acts. I have already argued elsewhere that the AI cannot be regarded as the subject described above. It is the successful imitation of the human-like communicative patterns by LLMs that makes us mistake their generated answers for the free and reasonable subject's behavior, because the adequate use of language has always been the natural “indicator” of the conscious being (Trufanova, 2025). If AI does not qualify as a subject, it cannot be held responsible for the solutions it proposes or for the consequences of their implication. Therefore, there is no accountability to talk about.

The recent AI development builds the foundations for the new kind of technocracy — the AI-based technocracy. Looking for the aforementioned fast and simple solutions made by the AI algorithms we invoke the risk of switching the responsibility for the decisions from humans to the AI systems, which means that when the AI makes a mistake, it will be regarded as a malfunction nobody is responsible for. As Floridi wrote in his work on distributed moral responsibility, “Too often ‘distributed’ turns into ‘diffused’: everybody’s problem becomes nobody’s responsibility” (Floridi, 2016). Russian philosopher of technology Tatiana Leshkevich mentions likewise that most of the time we encounter problems of attributing and reclaiming the responsibility of the algorithmic systems: for example, when a bank rejects a loan application after it was rejected by the algorithmic model, we have no one to blame for this (Leshkevich, 2023). This will virtually mean the impunity of certain deeds realized with the help of the AI. This problem is sometimes called a problem of algorithmic accountability (Shah, 2018), and it presents both ethical and legal challenges.

There is an important principle in machine learning that is called “human-in-the-loop” (HITL). It refers to the approach in the machine learning and in machine decision-making where humans are actively involved — they verify the data used by the AI systems, provide feedback to them, evaluate their performance, etc. Naturally, humans can also approve or disapprove of the decision made by the AI. Human-in-the-loop should not only be present as a “teacher” of the LLMs; the human should be the responsible subject when the AI is used. Thus, it is a decision *made by a human expert* with the help of the AI (or based on the solutions suggested by the AI), not a decision *made by the AI* and routinely approved by a human. That is to say, the AI system should have only an advisory vote in the decision-making. The idea of the responsible AI then should refer not to the AI system in question but to the humans and institutions that are involved in its development and usage. This principle is touched upon in the UNESCO’s “Recommendations...”:

> Member States should ensure that it is always possible to attribute ethical and legal responsibility for any stage of the life cycle of AI systems, as well as in cases of remedy related to AI systems, to physical persons or to existing legal

> entities. Human oversight refers thus not only to individual human oversight but to inclusive public oversight, as appropriate (UNESCO, 2022: 22).

As we have mentioned earlier, the AI algorithms can be prejudiced, as well as the ethical principles that they are taught to abide by may differ according to the different value systems of the human "teachers"; thus the basic principles of the responsible AI might differ as well. So the humans should not only try to embed certain ethical and other RRI principles in the AI but should also make sure there is no "algorithmic bias" that has emerged along the way.

Hence, we can summarize that the AI cannot be a responsible agent, and the question of responsible AI is the question that is carried out by human agents.

ATTEMPTS AT DELEGATING HUMAN TASKS TO THE AI

Among the responsible and the trustworthy AI, the latter seems to me to be a more important goal, because we need AI to become a reliable instrument we can use to solve different tasks. We don't need a knife to be responsible; we just need it to be sharp and to be used for good purposes and not for causing harm; the same can be said about the AI.

We can assume that we can mostly trust AI with the simple and specific tasks — doing math, classifying data, looking for certain objects in the vast amounts of big data, etc. Famous Russian theoretical physicist Valerii Rubakov stated in the interview about a decade ago that the scientists nowadays cannot check all the calculations made by the computers, so they have no choice but to trust them (Lektorskiy et al., 2022); the same conclusion can be made about some of the AI operations. There are also certain "creative" tasks that the present Gen AI does fairly well — drawing pictures, composing music, writing texts, etc. It might not be the work of art to impress the generations of art lovers, but it might satisfy the needs of copywriters, advertisers, mass media specialists, pop artists, etc.

But when it comes to the more sophisticated analytical and theoretical work, AI might not be that trustworthy. American philosopher Jacob Browning and information scientist and specialist in machine learning Yann LeCun at the dawn of the so-called "GPT revolution" show that while LLMs became quite proficient in using language, it does not mean that they have knowledge about the things they are "talking" about, for language represents only a very limited part of knowledge. These limitations of LLMs that use the language without understanding the meanings beyond the words might result in their mistakes (Browning & LeCun, 2022). Likewise, Russian logician Vladimir Shalack analyzes the ChatGPT's skills in making logical statements and concludes that LLMs' "intelligence" remains on the prelogical level: it is based on the associative connections between words but cannot operate with logical connections between the concepts. "The great danger of the widespread infiltration of neural networks into our lives lies in the fact that in new non-standard situations they will block logically correct reasoning and thus lead us to incorrect conclusions..." (Shalack, 2024: 35). If Shalack's argument is valid, then we should not trust any of the AI's decisions, at least until we get access to its Chain-of-Thought (CoT) and check its logical correctness.

Even customer support service AI bots that were supposed to easily deal with typical questions and standardized answers have proved themselves not the best solution for business. In 2023–2024 a lot of big companies decided to fire hundreds or, in some cases, thousands of the human employees and substitute them with the AI "agents." In about a year, they decided against it and started rehiring humans. The Swiss financial services company "Klarna" sought to cut expenses, but after dismissing human employees it discovered that customers are dissatisfied with the help provided by AI chatbots — and that the company was losing customers, and therefore money, instead of saving it. Thus, they made it their priority to ensure that every customer will be able to solve his problems with a human agent if he has that need or wish (Dellinger, 2025). Famous language learning mobile app company "Duolingo" decided to go "AI-first" and replace human language lesson creators with the AI. The critical response came from the users — they noticed that the quality of study texts became worse (they became dull and stereotyped), and voice-overs of the texts by the AI in some cases just came up with wrong pronunciations, which is misleading for the students and should be regarded as provision of the poor-quality service (Rochefort, 2025). "Duolingo" bosses had to draw back and rehire human employees (Ivanova, 2025). These are only a couple of cases among many, and they show that LLMs' language generation abilities might still be lacking and the AI is not yet a valid substitute for human employees. Unfortunately, the business companies, in hope of raising profits, are still too eager to introduce AI solutions that are not ready to use, which paradoxically turns out to be not so profitable — the fast-and-simple solutions are not necessarily the best and the cheapest ones, which makes them unpreferable both for the companies and for their customers. Thus, both responsibility and trustworthiness questions should be addressed to the companies providing their AI services.

So, both responsible and trustworthy AI seem to be quite elusive goals if we try to take the AI systems as independent agents acting on their own. They only start to make sense when we regard them in the context of the *human-using-the-AI* process. It is not a human-AI *collaboration.* It is a human using yet another technical extension supporting and enhancing their abilities. Russian philosopher of science Sophia Pirozhkova regards the technologies as part of the "inorganic human body" and argues, "What is the specificity of digital and, above all, intelligent technologies? The fact is that a person delegates to such technologies the performance of intellectual tasks..., i.e. those tasks that until that time were among the exceptional human competencies... However, in the last century, it turned out that a person's intellectual abilities are not enough to solve the ambitious tasks that he sets for himself (including, I emphasize, purely cognitive tasks). Neither modern production nor scientific research practice can be implemented without intelligent technologies. Before their appearance, man delegated the solution of intellectual tasks only to other people, but not to artificial objects... But never before have technologies participated in the division of intellectual labour, and they are also distinguished by their unattainable perfection in the implementation of intellectual procedures. Only some of those procedures so far, but who knows..." (Lektorskiy et al., 2022: 30–31).

There are naturally certain technical limitations that should be discussed by the AI developers regarding what the AI is able to do and what mistakes it can make that can help us to estimate the trustworthiness of the AI in solving certain problems. When discussing ethical questions of the AI though, as Yastreb rightfully puts it, we need to consider not the AI on its own but hybrid systems that include humans, technologies, and social institutions that interact with the AI, and the object of such AI ethics should be "the content and nature of changes in the human ideas and values, their behavior and decision-making that happen under the influence of the artificial intelligence systems" (Yastreb, 2025: 92). So, it is mostly the task of humanities scholars and social scientists to provide the humanitarian expertise that will help us to draw boundaries of what we entrust the AI to do, both in scientific research and in the social sphere, because this is not just the question of what the AI is *able* to do, but also the question of what we should *allow* it to do.

HUMAN RESPONSIBILITY

There are a lot of different legal, ethical, or just common sense issues being discussed concerning the AI usage. These are such questions as creativity or copyright issues when we talk about the AI generating pictures for commercial use (do we regard the prompt-writer of the image as its author who is the subject of copyright? Does the resulting image belongs to the developer of the Gen AI system used to create it, or does the image have no copyright at all? This includes making deepfakes of the dead actors in the new movies, changing the faces of certain actors with the help of the AI due to the censorship (when reputation flaws of the actor might otherwise influence the movie in question), mentioning GenAI systems as co-authors of the scientific research, etc. These are interesting questions, but they are not life-staking. In fact, most of them are quite pragmatic — one has to decide who gets paid for certain AI-generated products.

But there are also sensitive matters such as human health and well-being, human freedom and dignity, social security, environmental issues, etc. A great deal of hope is placed, for example, in AI systems used to design new medicine, make a diagnosis, perform surgeries, assess potential life-threatening or environmental risks, or even render judicial decisions. These matters are sensitive because those are either life-or-death questions or the questions that might seriously influence the quality of life of the individuals. These are questions that should never be left for AI to resolve without the oversight and final judgment of a human expert.

At first glance AI expertise — based on the statistical analysis of the many cases it has been trained on — does not seem very different from the "personal knowledge" (to use M. Polanyi's term (Polanyi, 1958)) of a human expert, who draws conclusions from their own lived experience. The AI system and the human expert both have in their "minds" a number of certain previous cases that make them solve the current one, and the AI has a certain advantage here, for it can analyze many more cases than the expert has ever met. The advantage of the expert, on the other hand, might lie in the embodied nature of the expert's knowledge — he didn't only analyze the cases; he lived these experiences. As Browning and LeCun put it: "But we should not confuse the shallow understanding LLMs possess for the deep understanding humans acquire from watching the spectacle of the world, exploring it, experimenting in it and interacting with culture and other people" (Browning & LeCun, 2022). Does this "living it" experience give a real advantage when compared to the nearly unlimited (compared to an individual person's) archive of digital data available to the AI? It might be one of those philosophical questions that might never have an answer. But when we are speaking about those sensitive matters I mentioned above, it is a difference that is important to mention.

Since the times of the famous Alan Turing's paper on the intelligence of the machines, there have been a lot of discussions about what AI can and cannot do in comparison to humans. As we already understand, the AI can perform some of the intellectual tasks that humans can perform (but we can say that about simple calculators as well) and some of them at a level that is unreachable to individual humans. What is crucial to my mind is that the AI systems cannot provide intersubjective judgement and cannot be empathic. Humans, when making decisions concerning other humans, do not only have the advantage of "personal knowledge," they understand what being human is like, and they might have fear of doing other humans harm and remorse after doing so. Simply to put — they *care*, while the AI systems simply *don't care*. The AI system cannot be responsible, because it does not care for being wrong or for being punished. There was a lot of sensationalist talk lately about the AI avoiding some of the commands, cheating, or even rewriting its code to avoid shutting down when explicitly ordered to do so (Pester, 2025), and some of the public were eager to interpret it as an emerging consciousness that tries to avoid "death." Though the most plausible explanation should be that the priorities of the algorithms dictated the AI to circumvent obstacles that were trying to stop its work rather than to follow the instructions to the letter (ibid.). Thus, there is nothing the AI system wants or needs either for itself or for the sake of human beings.

Both the AI systems and human experts can be wrong, and the mistakes of them both can lead to unforeseen consequences. Even when the efforts of humans and the AI are combined, no 100% successful result can be promised. Why do I insist on the priority of human experts over the AI systems in the questions directly concerning human lives? Because only humans can understand the value of human life or take into account a certain social and cultural situation of the person in question. For example, when the operation is needed, the AI assistant makes a calculation that shows that there is a 90% chance to save the person from sepsis by amputating the whole leg, while amputation of toes shows the less favourable numbers, which brings the risk of the second operation and possible complications. The human surgeon might prefer to take that risk trying to save the patient's ability to walk, because as a human the surgeon realizes how drastically the quality of life will be reduced after the leg amputation. He might be wrong and need to operate again, but he might as well be right and thus save the patient from being unable to walk without a prosthesis. This is why in sensible matters a human expert might check things up with the AI, use their support, but make decisions on their own. These are the decisions that they will be responsible for, no matter how much the AI has contributed to these decisions. So, as we are trying to integrate the AI systems both into our work and into our everyday lives, we should develop the social mechanisms (the existing ones are vague and insufficient) controlling the AI usage in different spheres that will maintain in sensitive questions the primacy of the responsible decisions over the fast-and-simple ones. This is our human responsibility to do.

CONCLUSION

In the beginning of this paper I have posed two questions:

(1) What can we trust the AI with?

(2) Who is responsible for the decisions that we entrusted the AI to make?

I have tried to answer these questions in my argument above. Now I can conclude that we cannot trust the AI as a source of reliable information, nor can we trust it with deep theoretical thinking. The AI works at its best with concrete problems in the narrowly defined field that can be solved with analyzing and comparing vast amounts of data, for example — in the search for the unknown correlations or some oddities. But whatever we trust the AI with, the responsibility *always* lies with humans, whether it is the developers of the AI systems, the experts using them in the decision support, or the persons of power who will commit to the implementation of the AI-based decisions. The responsibility may be distributed between various agents, but it will remain everyone's *shared* responsibility, not a dispersed one.

We can aim for the AI to be ethical, trustworthy, and responsible, but it is not a technological problem; it is a social problem. There is no need to regard the AI as a separate agent in the social relations. We can delegate the AI to solve certain problems, and we can try to improve their accuracy, but we cannot delegate them to be ethical and responsible instead of us. The AI is a tool that we as human agents should use ethically and responsibly and not expect that it can be ethical or responsible on its own or even that we can ever make it to be so. We will always be needing a human in the loop.

REFERENCES

AI HLEG. 2019. "Ethics Guidelines for Trustworthy AI." Accessed Oct. 16, 2025. https://ec.europa.eu/digital-single-market/en/news/ethicsguidelines-trustworthy-ai.

Altman, S. 2025. "The Gentle Singularity." Sam Altam Blog. Accessed July 11, 2025. https://blog.samaltman.com/the-gentle-singularity.

Browning, J., and Y. LeCun. 2022. "AI And The Limits Of Language." Noemamag.com. Accessed Aug. 8, 2025. https://www.noemamag.com/ai-and-the-limits-of-language/.

Daston, L., and P. Galison. 2007. *Objectivity.* New York: Zone Books.

Dellinger, A. J. 2025. "Klarna Hiring Back Human Help After Going All-In on AI." Gizmodo. Accessed July 25, 2025. https://gizmodo.com/klarna-hiring-back-human-help-after-going-all-in-on-ai-2000600767.

Edwards, B. 2023. "Why ChatGPT and Bing Chat Are So Good at Making Things Up." Ars Technica. Accessed June 25, 2025. https://arstechnica.com/information-technology/2023/04/why-ai-chatbots-are-the-ultimate-bs-machines-and-how-people-hope-to-fix-them/.

Floridi, L. 2016. "Faultless Responsibility: On the Nature and Allocation of Moral Responsibility for Distributed Moral Actions." *Philosophical Transactions of the Royal Society* 374 (2083).

———. 2019. "Establishing the Rules for Building Trustworthy AI." *Nature Machine Intelligence* 1 (6): 261–262.

Future of Life Institute. 2023. "Pause Giant AI Experiments: An Open Letter." Future of Life. Accessed June 25, 2025. https://futureoflife.org/open-letter/pause-giant-ai-experiments.

Ivanova, I. 2025. "Duolingo CEO Walks Back AI-First Comments: 'I Do Not See AI as Replacing What Our Employees Do'." Fortune.com. Accessed July 31, 2025. https://fortune.com/2025/05/24/duolingo-ai-first-employees-ceo-luis-von-ahn.

Lektorskiy, V. A., Ye. A. Alekseyeva, N. N. Yemel'yanova, et al. 2022. "Iskusstvennyy intellekt v issledovaniyakh soznaniya i obshchestvennoy zhizni (k 70-letiyu stat'i A. T'yuringa 'Vychislitel'nyye mashiny i razum') (materialy kruglogo stola) [Artificial Intelligence in the Research of Consciousness and in Social Life (In Honor of 70-years Anniversary of A. Turing's Paper 'Computing Machinery and Intelligence' (Papers of the 'Round Table'))]" [in Russian]. *Filosofiya nauki i tekhniki [Philosophy of Science and Technology]* 27 (1): 5–33.

Leshkevich, T. G. 2023. "Paradoks doveriya k iskusstvennomu intellektu i yego obosnovaniye [The Paradox of Trust in Artificial Intelligence and Its Rationale]" [in Russian]. *Filosofiya nauki i tekhniki [Philosophy of Science and Technology]* 28 (1): 34–47.

Marx, K. 1971. "Zur Kritik der politischen Ökonomie" [in German]. In *Marx-Engels-Werke (MEW)*, 13:7–160. Berlin: Dietz.

Novelli, C., M. Taddeo, and L. Floridi. 2024. "Accountability in Artificial Intelligence: What It Is and How It Works." *AI & Society* 39:1871–1882.

O'Brien, M. 2023. "WATCH: OpenAI CEO Sam Altman Testifies Before Senate Judiciary Committee." PBS News. Accessed June 25, 2025. https://www.pbs.org/newshour/politics/watch-live-openai-ceo-sam-altman-testifies-before-senate-judiciary-committee.

Papagiannidis, E., P. Mikalef, and K. Conboy. 2025. “Responsible Artificial Intelligence Governance: A Review and Research Framework.” *The Journal of Strategic Information Systems* 34 (2).

Pester, P. 2025. “OpenAI’s ‘Smartest’ AI Model Was Explicitly Told to Shut Down — and It Refused.” Livescience.com. Accessed May 30, 2025. https://www.livescience.com/technology/artificial-intelligence/openais-smartest-ai-model-was-explicitly-told-to-shut-down-and-it-refused.

Peters, U. 2022. “Algorithmic Political Bias in Artificial Intelligence Systems.” *Philosophy & Technology* 35.

Polanyi, M. 1958. *Personal knowledge: Towards a Post-Critical Philosophy.* Chicago: The University of Chicago Press.

Rochefort, S. de. 2025. “Duolingo Users Are in Turmoil Over the App’s AI Lessons.” Polygon.com. Accessed June 4, 2025. https://www.polygon.com/ai-artificial-intelligence/603216/duolingo-ai-language-lessons.

Shah, H. 2018. “Algorithmic Accountability.” *Philosophical Transactions of the Royal Society* 376 (2128).

Shalack, V. 2024. “Izbavleniye ot illyuziy II na primere ChatGPT [Exposing Illusions — The Limits of the AI by the Example of ChatGPT]” [in Russian]. *Technology and Language* 5 (2): 26–39.

Trufanova, E. O. 2025. “Mozhet li iskusstvennyy intellekt obladat’ svoystvami sub’’-yektnosti [Whether Artificial Intelligence Can Have the Properties of Subjectivity]: filosofskiye aspekty problemy [Philosophical Aspects of Problem]” [in Russian]. *Ekonomicheskiye i sotsial’no-gumanitarnyye issledovaniya [Economical and Social-Humanitarian Research]* 12 (1): 110–117.

UNESCO. 2022. *Recommendation on the Ethics of Artificial Intelligence.* Paris: UNESCO.

UNESCO and IRCAI. 2024. *Challenging Systematic Prejudices: An Investigation into Bias against Women and Girls in Large Language Models.* Paris and Ljubljana: UNESCO.

Yastreb, N. A. 2025. “Osnovaniya kriticheskogo podkhoda k resheniyu eticheskikh problem iskusstvennogo intellekta [Methodological Foundations of a Critical Approach to Solving Ethical Problems of Artificial Intelligence]” [in Russian]. *Filosofiya nauki i tekhniki [Philosophy of Science and Technology]* 30 (2): 90–103.

Trufanova E. O. [Труфанова Е. О.] Trustworthiness and Responsibility as the Key Issues of the AI Application [Надежность и ответственность как ключевые вопросы применения ИИ-систем] : Human in the Loop of Responsibility [человек в петле ответственности] // Философия. Журнал Высшей школы экономики. — 2025. — Т. 9, № 4. — С. 105–122.

Елена Труфанова
д. филос. н., ведущий научный сотрудник, Институт философии РАН (Москва);
ORCID: 0000–0002–2215–1040

Надежность и ответственность как ключевые вопросы применения ИИ-систем

человек в петле ответственности

Получено: 05.08.2025. **Рецензировано:** 19.10.2025. **Принято:** 01.11.2025.

Аннотация: В статье дается попытка ответа на два вопроса: (1) что мы можем доверить делать искусственному интеллекту (ИИ)? (2) Кто несет ответственность за решения, предложенные ИИ? Показывается, что использование ИИ для решения различных задач кажется привлекательным, с одной стороны, за счет своей скорости и простоты, которые предполагают в том числе экономическую выгоду для пользователей, а с другой — за счет предполагаемой объективности и точности, присущих интеллектуальным машинам в отличие от субъективных и склонных к ошибкам людей. Демонстрируется, что простота и скорость использования ИИ далеко не всегда приносят выгоду, а точность и объективность ИИ являются иллюзорными. Обосновывается, что надежность ИИ может расцениваться как высокая только для ряда конкретных узких задач. Показывается, что нельзя говорить об ответственном ИИ, поскольку ответственность — это свойство субъекта, а ИИ субъектом не является. Рассмотрение ИИ-систем как самостоятельных субъектов или агентов может стать причиной становления технократии, где решения в самом прямом смысле будут приниматься техническими системами, но ответственность за них не будет возложена ни на кого. Обосновывается, что в вопросах, которые касаются жизни и благополучия людей, их свободы и иных базовых человеческих ценностей, решения могут приниматься только человеком и только человек будет нести за них ответственность. Преимущество эксперта-человека для решения таких вопросов видится в интерсубъективном восприятии другого человека и эмпатии, недоступных ИИ. Исходя из собственного человеческого опыта, эксперт способен увидеть те особенности конкретной проблемы, которые искусственная система учесть не может. Делается вывод, что проблема этичного, надежного и ответственного ИИ не техническая, а социальная — это проблема того, как человек может этично и ответственно использовать такой мощный и сложный инструмент, как ИИ. Этичность и ответственность — свойства человека, которые не могут быть делегированы искусственным системам.

Ключевые слова: надежный искусственный интеллект, ответственный искусственный интеллект, ответственность, человек-в-цикле, интерсубъективность.

DOI: 10.17323/2587-8719-2025-4-105-122.