National Defence University
Department of Military Technology
Series 2: Research Reports No. 9

Research Papers on Artificial Intelligence in the Military Operational Environment and Wargaming

Saulius Keturakis, Arto Mutanen, Antti Rissanen &

Jouko Vankka (eds.)

MAANPUOLUSTUSKORKEAKOULU
SOTATEKNIIKAN LAITOS
JULKAISUSARJA 2: TUTKIMUSSELOSTEITA NRO 9

NATIONAL DEFENCE UNIVERSITY
DEPARTMENT OF MILITARY TECHNOLOGY
SERIES 2: RESEARCH REPORTS NO. 9

RESEARCH PAPERS ON ARTIFICIAL INTELLIGENCE IN THE MILITARY OPERATIONAL ENVIRONMENT AND WARGAMING

Saulius Keturakis Arto Mutanen, Antti Rissanen & Jouko Vankka (eds.)

NATIONAL DEFENCE UNIVERSITY
DEPARTMENT OF MILITARY TECHNOLOGY
HELSINKI 2026

Saulius Keturakis, Arto Mutanen, Antti Rissanen & Jouko Vankka (eds.): *Research Papers on Artificial Intelligence in the Military Operational Environment and Wargaming*

Maanpuolustuskorkeakoulu
Sotatekniikan laitos
Julkaisusarja 2: Tutkimusselosteita nro 9
National Defence University
Department of Military Technology
Series 2: Research Reports No. 9

Recent publications in PDF format: *http://www.doria.fi/handle/10024/73990*

ISBN 978-951-25-3584-2 (Pbk.)
ISBN 978-951-25-3585-9 (pdf)
ISSN 2737-0615 (online)

Maanpuolustuskorkeakoulu – Sotatekniikan laitos
National Defence University – Department of Military Technology

Punamusta Oy
Joensuu 2026
Finland

Introduction

During the planning phase of this book, we had already been observing for a few years the rapid and discussion-provoking environmental change related to artificial intelligence (AI). At the core of this is a technology-driven transformation, in which the capabilities of IT tools have developed significantly from before. At the same time, human interaction is changing in a new way mediated by technology. Additionally, the shaping is also intellectual, because technology is not merely a tool but an actual actor, whose presence in our military operational environment is increasing. AI-interfaced wargaming needs to be addressed, studied and governed from ethical perspectives. More broadly, this book invites a rethinking of the relationship between technology and knowledge in contemporary societies.

To regain our grasp of understanding new and inevitable changes, we can best utilize interdisciplinary research. This requires philosophical, societal, and technological research, each of which highlights its own perspective. There is a suitable tension between the perspectives. In practice, tense perspectives can be articulated in text, and through them, unifying insights can be created here and now.

Chapters and Authors

ARTIFICIAL STUPIDITY AS A WEAPON AGAINST ARTIFICIAL INTELLIGENCE IN THE FIGHT FOR THE RIGHT TO LIVE IN A FASCINATING WORLD

Saulius Keturakis

Faculty of Communication, Vilnius University, Lithuania

Abstract This paper aims at finding an answer to the question that arises when artificial intelligence begins to do many things more efficiently than humans. The result achieved quickly thanks to technology turns out to be very impovering for the human world because many processes that previously had great cultural significance for humans now cease to be interesting to them. Based on the ideas of Claude Shannon, Robert Musil, and other authors, the study draws attention to the fact that humans have long had a reliable means of preserving the attractiveness of the environment. This is the so-called artificial stupidity, which, although not a powerful enough weapon to defeat technologies such as artificial intelligence, is sufficient to restore human self-esteem and well-being. The study draws attention to the so-called jailbreak culture, which helps the user overcome the prohibitions imposed by artificial intelligence administrators and forces the algorithm to act in the way that does not suit its intended purpose.

Keywords: artificial stupidity, artificial intelligence, brute force computing, restoration of meaning

Introduction

In everyday life, we are surrounded by a lot of completely unnecessary things. We want to eliminate them, but these things are still constantly nearby. Austrian writer Robert Musil said in his famous lecture "On Stupidity" (Musil, 1937) that one of such things – along with dreams and poetry – is stupidity. It is very impractical because what good can you do being stupid? As if that were not enough, stupidity is also very difficult to define, to clearly show where and what it is. However, according to the writer, at the same time, we would not be able to do without stupidity unless we stopped wanting to be human because being only bright is, in a certain sense, inhuman. Without stupidity, we would lose the emotional connection that ensures actions, without which life would lose attractiveness. In addition, stupidity is one of the elements without which there would be no meaning, nothing new would emerge, and there would be no opportunity to peek into the mind from the outside and indicate its limits.

A lot of other important aspects of our everyday lives are affected by stupidity, sometimes positively, and sometimes, unfortunately, negatively. In this article, we will return to Musil's famous essay and other considerations of stupidity, supplementing and developing in more detail the panorama of stupidity concepts. Now it is time to formulate another vital idea for further research.

Technology has always influenced human spiritual culture, but the history of ideas in the 20th century is so intertwined with the history of technology that neither can be told separately. When discussing the beginning of the 20th century, when all the most essential devices that shaped the human being of the previous century appeared one after another – the gramophone, cinema, radio, telephone, and automobile – it was noted that what was previously called moral authority began to be replaced by technological efficiency (Duffy, 2009). In other words, if the car drives and does not break down, the film does not jam in the film projector, the sound is heard clearly in the telephone receiver – then everything is good in the system of human reality. In such a culture that imitates the functioning of technology, the system of good and evil increasingly begins to coincide with the criteria of efficiency or inefficiency, and all bureaucratic, production, educational, social and other processes are evaluated only from their perspective. In other words, the human environment begins to imitate the logic of technological operation, in a certain sense becoming a technological medium in which the principles of proper or improper operation of devices are absolutely dominant.

In his study "Grammophon, Film, Typewriter" (Kittler, 2006), the German media theorist Friedrich Kittler discussed how human consciousness depends on the dominant technology. Each of the devices discussed in the book changed the manner of thinking formed by the written media in its own way: the gramophone and the film camera recorded sounds and light instead of letters of the alphabet, and the typewriter replaced writing with a universal code in which the individuality of the personality disappeared. Kittler convincingly showed that each of these technologies is not only a communication tool, but also a way for a person to understand himself, the surrounding world and history. Therefore, we can talk about a different human subjectivity created by each technology.

In his study of three fundamental technologies, Kittler does not talk much about the computer. On the one hand, it is simply an accelerated typewriter for him. On the other hand, if electronic writing is understood as the manipulation of electrical charges in transistors, then they are so miniature that writing as a process seems to disappear (Kittler, 2014), meaningful activity becomes imperceptible to the human senses, and only combinatorial games with numbers remain before the eyes. This may be one of the reasons why Kittler was not interested in the influence of digital media on the history of ideas, and basically used the computer itself only to solve mathematical problems (Holl, 2017).

Kittler, who has strangely ignored digitality in his media theory, is in a small minority of theorists. There are many more who, like Jay David Bolter, have argued that the computer is a device that completely redefines the human being (Bolter, 2014), and to a much greater extent than any previous technology. Importantly, in this transformation, Bolter believes, meaningful activity does not disappear but continues, albeit in completely new forms.

It seems that enough has been said to substantiate the idea that the intense technologization of culture is a great challenge for humans, who with each wave of the technological revolution are forced to reconsider the cost of cooperation with devices.

The nature of this price seems to have been best described by American mathematician and cybernetician Claude Shannon in his article “Programming a Computer for Playing Chess” (Shannon, 1950). Considering various strategies for creating software that could play chess, he formulated two of them based on the so-called brute computing power. In both cases, the computer would leave no chance for a person to win. It would seem that the most important thing needed would be the efficient operation of the technology, but right here Shannon seemed to doubt his project, because chess played by a computer would turn into an uninteresting game for a person, in which there would be no point in participating.

The most important problem of this text arises from this observation by Shannon– what can be opposed to the transformative power of technology to change the human environment so that it ceases to be attractive to humans? And if the problem were formulated in more detail, it would sound like this – how could one resist the brute force of computation, in order to restore the human ability to influence the world in ways acceptable to him, not to a computing machine?

Looking at the history of artificial intelligence, it is obvious that the path of thinking as brute calculation was followed, because the focus was on the fact that it is in this aspect that the machine is stronger.

We will return to Shannon's article later. Now, we can also say that our study does not follow the path usually taken by the so-called criticism of the choice to replace thinking with computation, usually indicating that the chosen principles for constructing artificial consciousness were bad. This was said by the famous American linguist Noam Chomsky, looking at the history of artificial intelligence, who claimed that the orientation towards computation and statistics led the technologization of thinking down the wrong path (Katz, 2012). The creation of a parallel between computation and thinking was met with great criticism by philosophers who claimed that thinking cannot be reduced to computation in any way (Dreyfus, 1972), and the entire chosen so-called strategy of brute computation destroys dreams of ever creating a device analogous to human consciousness (Dreyfus, 1965). Other philosophers' statements about the so-called brute force plan have drawn attention to the fact that they act as if it were possible to eliminate semantics from language, orienting all communication solely on syntax (Searle, 1986).

In the so-called cognitivist-connectionist debate (Stephan & Walter, 2013), in which the ability to feel meaning was opposed to statistics, compromise considerations eventually emerged, in which it was agreed that computation could not be identified with thinking, as it created something that could be seen as a substitute for thinking. In such considerations, the computer and its computational power were no longer equated with thinking, but were now referred to only as a thinking prosthesis, similar to a hearing aid for the deaf (Kurzweil, 2015). It was realized that the choice of the concept of intelligence changes a person’s relationship with the world, so if human thinking is replaced by statistical-like thinking (Moravec, 1998), the consequences need to be assessed. In the end, it was agreed that some tasks cannot be solved by human thinking within a reasonable period of time, so they have to be handed over to machines to solve. However, in this case, the question is, if the task itself does not change when solving it with the help of a machine (Weizenbaum, 1977).

The path of this research is different. Realizing that the history of technology cannot be changed, it is important to find something that can be used as a weapon in the fight for the survival of an interesting and intriguing world. Based on Musil's ideas about stupidity, we hypothesize for now that there is something interesting in those considerations about the relationship between stupidity and creativity, which we could perhaps take and turn into a conscious attitude, an artificial stupidity that breaks the monotony of the calculating algorithm and returns the attractiveness to the processes. Let us begin the search with a more detailed analysis of the technologically advanced, efficient, but completely unattractive world, as mentioned, which was essentially started in Shannon's article about the chess-playing computer.

Brute Computing Power, the Killer of Attractiveness

At the very beginning of the era of artificial intelligence, one of the most famous articles on the principles of artificial intelligence by Shannon, “Programming a Computer for Playing Chess” (Shannon, 1950), formulated a crossroads situation in which, despite the passage of more than seventy years, all fundamental artificial intelligence research and industry remain.

Chess is chosen in this article as an activity that is in no way related to everyday practice, but it is very suitable for testing various ideas about the creation of computer intelligence. First of all, as Shannon says, chess is inevitably associated with intellectual activity; on the other hand, this game consists of clearly defined moves and an end after a defined number of moves. Although the term artificial intelligence itself is not yet used in the article, it is clearly emphasized that we are talking about a completely different type of computing devices, because they make much more complex decisions than the binary ones of good/ bad.

Shannon distinguished two possible strategies for the game. He called one of them Type A ideal game strategy, which is based on “brute force”. An ideal game, according to Shannon, would be one in which a computer could evaluate every position as a win, a loss, or a draw. Unfortunately, since there are more possible moves in chess than there are atoms in the universe, a perfect game of chess is unlikely to occur until much faster computing hardware is developed. For this reason, in the same article Shannon proposed a modified “brute force” strategy, in which not all possible moves are calculated, but only those that are most promising considering a certain set of factors. Moreover, the calculation is carried out only until the arrangement of the pieces creates a calm situation. As Shannon argued, such a restriction on all possible moves reduces the required calculations, but they would still be too many to make it worthwhile to undertake such a chess-playing artificial intelligence project.

In order to reduce these large computational volumes, Shannon proposed the so-called Strategy B. According to this strategy, a system of criteria limiting the computational volume should be applied to each move, which would select the most promising moves and perform a large number of calculations to search for possible variants only when it is justified by the “strength” of the move. On the one hand, according to Shannon, in the case of Strategy B, the computational volume becomes realistic, on the other hand, such a machine would be significantly superior to a person: it would be fast and would not make mistakes, uncalculated moves or get nervous.

It seemed that the problem of artificial intelligence had been solved, and a way had been found to create a machine that would play better than a human, but without the human's inherent shortcomings.

However, right here Shannon notices a problem that is being discussed today as the most important topic of the impact of artificial intelligence on the human living environment. Shannon unexpectedly concludes in his considerations that despite the almost guaranteed victory, chess played by artificial intelligence will essentially become a completely different game. Chess, an intriguing, engaging, imaginative game, would turn into a boring activity that does not provide any intellectual pleasure if played by artificial intelligence. In order to avoid this, he immediately, having just created a machine that plays chess better than a human, begins to suggest that when evaluating moves, it is necessary to program the change of evaluation coefficients, and try to find ways to keep at least some shadow of attractiveness in chess.

Shannon does not elaborate on this insight in his famous article, but today, as artificial intelligence tools gradually become a universal everyday phenomenon, the power of artificial intelligence to transform ordinary phenomena into something completely new deserves greater attention. Extending Shannon’s idea about the danger of boredom in chess played by machines, one could say that in this game the result – victory – is not all that a person gets from the game. In chess, as in any human activity, the process is as important as the result, as it usually accounts for uncertainty (Rescher, 1996), intellectual intrigue and attraction.

Artificial intelligence, which replaces the attractiveness of human world processes with “brute computing power”, fits perfectly into the history of the mass technologization of the 20th-century culture, which Paul Virilio described as the move of the soul from the brain to the motor (Virilio, 1995). The motor is, of course, much faster and more powerful, so at first this process seems to provide a greater amount and variety of information, but then the amount of information becomes so large that it is accessible only to the machine, and the human sees it only as a cognitive flicker (Duffy, 2009). Virilio warns that such a changed understanding should not be seen as a catastrophe, but as a new type of understanding, only it needs some other term, because it is no longer understanding, but an approximate sensation, which is more related to work and fatigue than to enjoyment (Han, 2015)

Looking at today's culture industry, it is clear that the renunciation of enjoyment is not met with enthusiasm and a variety of ways are being sought to control the “brute force of calculation” and restore the ability of a person to perform necessary tasks in a way that is familiar to him - not (only) by calculating, but also by making mistakes, experimenting, and experiencing the adventure of the unknown. These processes of controlling the calculating algorithm are especially clearly seen in computer games.

Stupidity in Computer Games

As already mentioned, computer games are an excellent space to observe the transformative effect of brute force, formulated by Shannon, when looking for ways to reduce the capabilities of an algorithmic human opponent and preserve the attractiveness of the game process. Despite the fact that computer games are a space of fiction, it is fiction that is best for observing human behavior, because it is possible to know it, while reality, unfortunately, is not knowable, because it is too complex (Eco, 2011).

In computer games, the player's algorithmic opponent has every chance of being the ideal player described by Shannon: he sees everything, he knows where the human-controlled avatar is hiding, he has the precision of every shot, stroke or movement at his disposal. If a player in computer games were to face such an opponent, following the so-called Shannon A strategy, he would have no chance of winning.

However, a person would probably not play computer games if faced with such an opponent. If chess, played by an artificial intelligence guided by brute force, turns into boredom, according to Shannon, the same thing would happen to computer games. As the computer game designer Sid Meier observed, a player only plays computer games when he feels that he will win at least every other time (Meier, 2010).

Therefore, computer games use many tools that could be called artificial stupidity, because the goal of the algorithm is not just to win, it seeks to create interesting, attractive situations, engaging the user and keeping them engaged in a computer game for as long as possible.

Algorithmic opponents in computer games can be divided into three types (Kumar, 2012). The first type of algorithmic opponents in computer games create a so-called game template. Their algorithm does not make any decisions, they perform the same action no matter what the player does. Most often, such algorithmic opponents are found in fighting games or first-person shooters. An example would be the so-called Mysterious Stranger in computer game "Fallout", where this cowboy-looking character always performs the same action: appears unexpectedly, shoots a revolver at your opponent and disappears (Carmine Arcopinto, 2019).

The second type of algorithmic opponents change their behavior, but they choose it randomly, as if from some kind of a roulette of actions. In this case, the computer does not make any decisions, the algorithmic opponents only seem to be thinking about something. For example, in computer game "Mother", almost all opponents are like this, before the fight begins, they frown, smile, or turn their heads in a random sequence.

Only the third type of algorithmic opponents in computer games are those that act by analyzing the situation. In their case, elements of artificial intelligence can be seen. Such are the ghosts of "Pac-Man", which react to the player's actions and make complex decisions (Retro Game Mechanics Explained, 2019). It is in this case that we encounter a situation where the artificial intelligence of a computer game has the ability to acquire the properties of the so-called ideal opponent (Salmond, 2021), when the player has no chance of winning.

As already mentioned, this is one of the features that can scare players away from the game. Such a story happened with Meier's “Pirates!”, in which one of the main characters, the Marquis de la Montalban, could gain such powers that it became almost impossible for the player to win. After receiving numerous criticisms, Meier had to simplify the computer game (Meier, 2020).

In computer games, the player's algorithmic opponent typically makes many intentional stupid mistakes, the purpose of which is to increase the player's satisfaction (Green & Kaufman, 2015).

There is a large number of such cases of intentional stupidity (Artificial Stupidity, n.d.), and some of them can be mentioned as the most effective.

In the 2003 first-person shooter “Call of Duty: Black Ops Declassified,” the enemy often behaves very strangely, e. g., he hides behind a pile of explosives. Not only does he hide, but he also tries to shoot from behind it. Of course, that pile of explosives often explodes spectacularly along with everyone behind it.

In another first-person shooter, “Crysis”, released in 2007, the opponent is often so lost in thought that even if you stand directly in front of him, the player's avatar remains unnoticed. What is more, opponents often commit suicide without any warning.

In one of the most popular games of all time, “Doom,” released in 1993, you can often escape from enemy monsters without any special effort. You just have to climb onto the table and the monsters do not know what to do anymore, as they just run around screaming. Another strange feature of “Doom” enemies is that they like to follow the player instead of waiting in ambush. A common trick, well known to fans of this game, is to turn around after walking away and wait with the weapon raised. Usually, a chasing enemy appears soon and can be killed easily.

Perhaps the most stupid behaviour of an algorithmic opponent can be found in the 2014 fighting game “Smash 4” – in one of the game’s levels, the best strategy is… to do nothing, just stand there and look around. This behaviour seems to completely deprive the opponents of their sanity, they rush to get distracted in all directions and somehow kill themselves quickly.

It would be inaccurate to treat all these well-known computer game oddities as just programmer oversights. Computer game psychologists explain that such intentional mistakes give the characters of algorithmic computer games identities, the player begins to be interested in them not only as virtual opponents, but also as beings with a certain way, character, and soul. In other words, these elements make up a large part of the appeal of a computer game, because such stupidities of algorithmic opponents not only create an illusion that it is possible to win, but also the feeling that there are many possible unpredictable adventures ahead, just like in real life.

These computer game follies usually do not have a significant impact on the outcome of the game, but they perform an important function – they shift the player's attention from the result to the process. In this way, computer games unexpectedly come into contact with the so-called process philosophy, for which the ability of reality to constantly change is more interesting than the substance, the unchanging units of the reality. If the end of a computer game, no matter in what form it manifests itself, were understood as the result to which the entire game is directed, then the various follies occurring in the game could be treated as a movement without a clearly defined direction (Rescher, 1996). And this movement can no longer be described in terms of concepts such as coordinates, it should be more precisely described in metaphors that refer to the experiences being experienced (Bye, 2020).

Now it is the time to discuss stupidity and evaluate its potential as a way to harness brute computing power and restore the attractiveness of the human living environment.

Stupidity as Resistance to Monotony

In his famous lecture "On Stupidity" (Musil, 1937), Robert Musil discusses typology of stupidity. The study of stupidity, as the author says, is an endless pursuit, because it is a phenomenon that is very difficult to grasp. Therefore, instead of attempting a speculative theoretical study, he divides the whole problem of stupidity into two parts, and then sets about compiling sets of properties for each case, which – as Musil constantly reminds us – are never finite.

Musil describes the first type of stupidity with the following series of epithets. These are honest (*ehrliche*), simple (*schlichte*), bright (*helle*), gullible (*leichtgläubig*), unclear (*unklar*), and incorrigible (*unbelehrbar*). As Musil says, the communication of this type of stupidity is characterized by naivety and pronounced physicality. To illustrate this type of stupidity, Musil presents the responses of those who are honestly stupid to various inquiries.

For example, when winter is mentioned, the response to this stupidity is that it is made of snow. The father is described as someone who once threw someone down the stairs, the wedding is described as being for fun, and religion is encountered when going to church. In Musil's essay, one can feel the author's fascination with such descriptions; he identifies them with the human ability to create in general, also classifying himself as a type of fool. In this study, we will leave aside the question that sometimes arises angrily in discussions about Musil's phrase, where simple stupidity is defined using the feminine gender: "[...] indeed, simple stupidity is often a female artist" ("[...] die schlichte Dummheit ist wirklich oft eine Künstlerin") (Grill, 2013).

Musil points out that the artist, when asked, does not answer laconically or conceptually, but by developing various perspectives and telling a story. In answering, she becomes, as it were, her own creator, constantly reconstructing elements of reality, creating a living environment in which something is constantly changing.

This is the world of a child, according to Musil. Or a fool. In both cases, it is obvious that the new appears when one manages to maintain the instability of the reality and live according to habits. The new is generally not defined by words, concepts, or rational grammatical constructions, but by some kind of vagueness and indeterminacy of speech.

In one of his short literary fragments "Was ist eine Straße?" (Musil, 2022), Musil says that the answer to a question should always be multiplying, not limiting. Giving a clearly defined answer to the question is pointless, one should behave as one behaves on the street – to wander, constantly revealing new aspects of the street. That is the purpose of that street – to always lead somewhere new.

On the one hand, with this approach that connects art, reality, and uncertainty, Musil continues Friedrich Nietzsche's ideas about the necessity of remaining creative subjects, constantly creating new metaphors, because reality is intangible and inexplicable (Nietzsche, 2019). On the other hand, he understands that seeing something new from the perspective of those who adhere to an established point of view always looks stupid. In the aforementioned literary fragment "Was ist eine Straße?" Musil once again returns to the connections between rationality, stupidity, and gender. Here he identifies defining, finite thinking as masculine, and wandering, stupid thinking as feminine, unequivocally identifying himself as a confessor of the latter.

The second type of stupidity – intelligent (*intelligente*), higher (*höhere*) – is associated by Musil with the phenomenon of education, summarized by the metaphor of illness (*Bildungskrankheit*). As Musil often does, he defines the phenomenon by a kind of catalogue of symptoms. The illness of education is manifested by lack of education (*Unbildung*), bad education (*Fehlbildung*), education obtained in the wrong way (*falsch zustande gekommene Bildung*), an incorrect relationship between the material of education and the method of education (*Mißverhältnis zwischen Stoff und Kraft der Bildung*). This type of stupidity can also be summarized as a disease of the spirit (*Krankheit des Geistes*).

In Musil's understanding, ordinary stupidity is always individual, associated with a unique view of the world, with creativity and art. On the contrary, the so-called higher stupidity can be – to continue the metaphor of a disease that Musil insightfully uses in relation to it – contagious, spreading like a pandemic through communities. Musil calls the very process of the transmission of higher stupidity from one individual to another imitation (*sozialen Imitation geistiger Defekte*), but does not go into a more detailed analysis, satisfying the reader's curiosity with additional metaphors of collective activity.

Musil says that higher stupidity can be recognized in situations where thinking is obliged to act according to the rules of sports (*Denksport*), as if oriented towards citius, altius, fortius, and not towards the parameters inherent in natural reality. Interpreted in this way, higher stupidity turns into a kind of alternative to the reality, a fiction that a mentally ill community can create as much as it wants, adapting to the changing needs (*angewandten Dummheit*). As Musil says, intelligent stupidity can take on whatever form it wants, but the reality has only one form, towards which ordinary stupidity is oriented.

The fact that higher stupidity can be imitated, while ordinary stupidity cannot, is an important observation for discussions about the nature of artificial intelligence and its impact on the human world. In Musil's view, only mental illness or higher stupidity can be transferred or reproduced, but ordinary stupidity remains inimitable. Extending this argument to artificial intelligence, it seems possible to argue that not only social but also technological imitation of mental illness is possible. We will not pursue this idea further, as it is not directly related to the problem the article discusses.

Although Musil clearly sympathizes with simple stupidity in his reflections, he is looking for something that could cure the diseases of the spirit that have arisen due to various educational failures. First of all, he names the goal – it is wisdom (*Klugheit*), which harmonizes feelings (corporeality) and reason. The connecting function is performed by meaning (*Bedeutung*), which connects reason, reality and sensations into a single system, because it is the latter, according to Musil, that give us a sense of confidence when we encounter something unseen, unexperienced, or new. From the perspective of meaning, one can speak equally successfully about reason, feelings and reality; the meaning is opposed to stupidity and brutality. At this point, Musil stops his reflections, not even trying to achieve a definition of meaning, saying that no matter how hard one tries, no one has succeeded so far.

Musil's study of stupidity ends with the lesson that one should act as well as one can and as badly as one must, with a clear understanding of both.

Musil's analysis of stupidity can be summarized in essence as the opposition between the inability to choose and conscious choice. Simple stupidity is the inability to choose, a person simply is, he is different from everyone else and this is what earns Musil's sympathy. The conscious choice of stupidity is a crime against human nature, because it is the use of intellectual powers for manipulations aimed at achieving a chosen goal by any means.

If we were to look for a similar binary concept of stupidity in the history of philosophy, the closest to Musil would be Immanuel Kant's theory of stupidity, in which he distinguishes between stupidity of understanding (*Mangel an Verstand*) and stupidity of judgment (*Mangel an Urteilskraft*) (Kant, 2007). Stupidity of understanding emerges when there is lack of concepts to explain a situation. This can be solved by learning. Stupidity of judgment is found when concepts are applied incorrectly. The paradox is that in this case, the more concepts a person has, the greater the chance is of making mistakes and applying them incorrectly, so education cannot help here (Golob, 2019). What Musil calls simple stupidity essentially corresponds to Kant's stupidity of judgment, because in this case no change or choice is possible, and a person is doomed. In the case of stupidity of understanding, dynamics are possible, a person only needs to decide to educate himself.

In his view of stupidity, Musil contradicts most of the later modernist ideas about stupidity, which were characterized by a complete separation of stupidity from any meaning, leaving only a combinatorial play with empty semiotic signs. This view was most clearly represented by Jacques Lacan, who said that stupidity is just a special function of a language, a meaningless speech, a speech of zombies who have no consciousness (Zeiher, 2025). Theodor Adorno and Max Horkheimer also thought similarly about stupidity, for whom stupidity was a kind of "blind spot", which is not defined in relation to knowledge or reason, but much more broadly, as a practical phenomenon or mental empty space (Adorno & Horkheimer, 2008).

From this brief overview of the most important concepts of stupidity, several important aspects emerge. First, stupidity can be a conscious, chosen attitude towards the reality. Second, this attitude is recognized as capable of transforming the reality, making it unusual, individual, and artistic. Both Musil and Kant have an important ethical perspective in their concepts of stupidity, but in our study we will leave this aspect aside, considering only the possibilities of stupidity as a tool.

The choice of such an approach to stupidity coincides with the so-called intellectualist model of stupidity, which emphasizes the purpose of stupidity rather than its origins (Engel, 2016). In the case of this model of stupidity, a person consciously chooses the nature of his behavior, and the assessment of the behavior itself as stupid or not depends on the goal pursued. In this model of stupidity, promising things to friends that one cannot do is ridiculous, but promising things that one cannot do to be elected to parliament is no longer laughable. At this point, it is difficult to refrain from blaming such an approach to stupidity for the many crises that have befallen our world. Still, for now, only one question is important to us: will such or consciously chosen stupidity be able to dispel the boredom created by technology predicted by Shannon and restore the attractiveness of the world around us, which is vital for humans?

Conscious Ignorance

Artificial intelligence has become so popular today that humanity has encountered something fundamentally new, so new that it is not easy to inscribe this technology into the organic process of cultural history.

However, this task is not impossible. The history of culture can be viewed as a human effort to find some other place for consciousness, not only in the body. These are books, paintings, musical compositions, all objects in which, according to Benjamin, we feel an aura, a trace of human touch (Smith, 1997). Artificial intelligence seems to be one such case; only the transfer of consciousness occurs not figuratively, but - at least at the level of illusion - in a more literal sense, transferring not so much the results of the activity of consciousness but the processes of its activity.

British anthropologist Nigel Thrift has proposed that all objects in the human environment are characterized by properties called "intelligencings" (Thrift, 2007). As the scientist explained, this term was formed by combining two words - "intelligence" and "thing." It means that every object in a person's environment has a particular mind, which we sometimes define as meaning, sometimes as a function or purpose. In other words, human consciousness can be transferred to an object by creating something, noticing, distinguishing, and giving some meaning. In Thrift's opinion, in the history of 20th-century art, what he calls "intelligence" is best seen in the so-called “readymade” creative strategy, when others replace the functions of one object at the artist's will.

In such a perspective, the communicative aspect between a person and an object, in which he has transferred part of his consciousness, is essential. Niklas Luhmann also attributes perception to this communication between a person and an object (Luhmann, 2000), turning the passive perception of a person’s environment into a part of the communication process. Luhmann notes (Luhmann, 1990) that even systems of such opposite nature as man and thing can create such intense mutual connections that sometimes there is a threat of complete mutual similarity.

Interestingly, in fiction, this mutual circulation of man and technology is described much more boldly than in theoretical considerations. In the first half of the 20th century, the Irish writer Flann O'Brien wrote novel “The Third Policeman” (O'Brien, 2007), in which the characters are obsessed with bicycles. They constantly ride them, talk about them, and repair them. In O'Brien's story, the reader, at some point, realizes that human identities have begun to mix with those of bicycles because, as the novel explains, the atoms of bicycles have combined with those of humans. In the end, all those riding bicycles turn into a kind of modern centaurs, half-human, half-bicycles.

Because of this connection with humans that accompanies every technology, the reflection of every technology must also include its user. In the new media era, it is essential to understand that two types of technologies exist. The first is planned, static, with a clear finite form—for example, buildings. The user's contribution to the further development of the building is small. However, the second type of technology is digital, so these technologies usually do not have a finite form; they are dynamic and constantly supplemented by the user. For example, computer games or artificial intelligence systems that continuously learn from the data provided by the user. Technologies of this type correspond to what mathematician and philosopher Alfred Whitehead called the concept of “concrescence” (Whitehead, 1929/1985), when an abstraction turns into a concrete thing, each additional contribution to the system replaces and supplements all previous elements.

In the case of artificial intelligence – as in all new media – it is the user who, through his communication, turns this technology into a specific tool that performs the required task. The way we interact with the artificial intelligence system supplements it with data and changes it, but at the same time, we change ourselves.

The relationship with artificial intelligence technology can be very diverse. For example, London artist Micheál O'Connell, who presents himself as a counter-inventor, has created a methodology for going to a store, using the entire complex inventory system and automated cash registers... and not buying anything (O’Connell, 2016) – and having proof that nothing was purchased. This is, of course, very stupid – why go to a store if you are not going to buy anything? However, using the system for something other than its intended purpose, opposing its essential purpose – to force us to buy something – is a conceptual and creative step.

The most important thing is that M. O'Connell acted in a way no one had ever thought of before. Here, it is worth remembering two concepts - feedback and feedforward (Thrift, 2007). These are two models of the relationship with the environment, which are now often remembered in discussions about artificial intelligence and artificial stupidity and the need for their balance.

Oddly enough, the example of the differences in shower behavior in a familiar and unfamiliar place is often used to explain the differences.

Remember yourself, finding yourself in a hotel in a foreign country, taking a shower for the first time after a trip, and trying to determine the proper water temperature. You do not know how hot the water or the pressure in the system is. So you play with cold and hot water streams, scalding yourself until you finally hit the right temperature.

This is a feedback relationship with the environment when you judge the results of your actions by the reality reaction to them.

In the case of feedforward, you know well how to handle hot and cold water taps, set the right temperature without mistakes, and avoid any temperature adventures.

In this case, you act without examining the reality, knowing its response well, and completely obeying the preconceived model of understanding. It is very safe and convenient, but never anything new.

Now imagine that the feedback relationship is transformed into a conscious activity, and you constantly examine everything, even when your actions break the boundaries set by technology or deliberately provoke errors. Of course, then you act stupidly. But then you are creative, innovative, and artistic.

Musil would have treated the feedback relationship with reality described by Thrift as a case of intelligent, higher stupidity, which means that it would be a reprehensible, unethical choice. Paradoxically, in the feedback concept formulated by Thrift, whose original purpose is to search for original, unexpected, and non-intended technology applications, we do not feel like we are doing anything wrong. In the case of our relationship with technology, the conscious choice of stupidity has become justifiable (Tavani, 2016) because technology has begun to change the world to an extent that we were not prepared for.

Artificial Stupidity, the Weapon of the Weak

Since the early 18th century, Western society has been accustomed to the idea that a machine more intelligent than humans will be created, and humans will clash with it inevitably (Szollosy, 2016).

Although the history of artificial intelligence dates back to the time of Homer (Mayor, 2020), it is only in modern times that fear of the inevitable battle with intelligent machines, the weapons with which it will be fought, and the possible consequences of that battle have begun to grow.

Three works of fiction, written at different times starting in the 18th century, clearly represent the growing tension between man and machine.

In his "Gulliver's Travels", the British writer Jonathan Swift describes the Lagada Academy, where a device, closely resembling "ChatGPT", operates. Gulliver, visiting the academy, finds a professor who has created a mechanical computer capable of writing works on philosophy, poetry, politics, law, mathematics, and theology. The professor wanted a device that could create without talent, without putting in any work.

The narrator of Swift's novel notes that the professor's vanity is dangerous because he thinks he understands language by the frequency of words. This passage is worth quoting precisely because it sounds like it comes from the instructions for building great language models today: "[...] he had emptied the whole Vocabulary into his frame, and made the strictest Computation of the general Proportion there is in Books between the Number of Particles, Nouns, and Verbs, and other Parts of Speech" (Swift, 2011).

As for the relationship between machine and man, Lagado's academics did not feel any threat from the machine because, in their opinion, technology is always superior to man. For example, if a new drug kills a person, it is still the man himself who is to blame.

Machine and man are irreconcilably at odds in Mary Shelley's novel "Frankenstein or The Modern Prometheus" (Shelley, 2004) because this book describes an entirely different type of automaton from Swift's Lagado Academy. First, it was associated with the sciences of chemistry and electricity, which set out to explain the origin of consciousness. Second, it was characterized by a developed consciousness; in just a couple of months, it was able to go from complete ignorance ("I knew, and could distinguish nothing") to a high level of intelligence and acquire much more knowledge than its creator, the scientist Victor Frankenstein.

The problem of controlling artificial life has interested thinkers since ancient times (Mayor, 2020). However, Frankenstein's creation made such control almost impossible for the first time because it was more intelligent than its creators. At the same time, the process of its creation was scientifically described, so such superintelligent creatures could create as many as they wanted. Thus, the myth of the threat of artificial intelligence arose. Especially since the fight against one of the first superintelligent artificial creatures in the history of Western culture was completely unsuccessful: Victor Frankenstein dies while defending his creation, and the superintelligent creature he created promises to commit suicide, but the reader remains in the dark about whether this will happen. This superintelligent creature may be deceiving everyone – both the characters in the book and us, the readers.

Artificial intelligence has become the cause of a new fear for man because, as it turned out, a computer can be a perfect creature. Long before artificial intelligence became a part of our everyday lives, algorithmic intelligence was imagined to be far superior to human intelligence. One of the first cases of a human successfully confronting an artificial intelligence was presented in Julia Ecklar's 1989 novel “The Kobayashi Maru”, written using the “Star Trek” franchise (Ecklar, 2000). The novel tells the story of a spaceship crew who, having found themselves in a desperate situation where nothing is left to do but wait for death, embark on storytelling. Captain Kirk of the crew tells how, while still not a famous spaceship commander in the entire Universe, but just a young cadet, he defeated his opponent, an artificial intelligence, in training. However, the duel scenario was programmed, so the cadet had no chance. Kirk adjusted the computer program to make it seem to her that the opponent was not Cadet Kirk but the famous Captain James T. Kirk. To everyone's surprise, when faced with such a deception, the artificial intelligence went wild and even helped the cadet escape from the danger zone instead of attacking. As it turned out later, after comparing the stories of all the crew members, such stupidity was the only successful scenario for winning a desperate fight against artificial intelligence. Having understood this, the crew members take action and save themselves.

Today, artificial intelligence has become a part of our everyday life. It forces and suggests what movies to watch. We willingly, without any struggle, agreed that artificial intelligence would take over a large part of our activities. Once, this situation was the subject of science fiction, but today, it is the reality. For now, fighting machines are only in science fiction stories, but who knows, one day, they might become as much a reality as artificial intelligence technology itself.

And what would we do then? What weapons would we fight with? It is a ridiculous question because a machine poses no threat to humans. However, Lev Manovich claims that artificial intelligence threatens to deprive us of our individuality of style, forming a strange, unified, uniform style of communication we are increasingly encountering (Manovich, 2025). Isn’t this changing the human environment to such an extent that it is time to think about resistance? Maybe this is what one of the first actual front lines of a clash between man and machine will look like. And this is certainly not about the sky being burned down, as in the movie “The Matrix”, so that the solar panels of the machines no longer receive light. Suppose anyone today wants to resist artificial intelligence. In that case, they should first pay attention to the so-called jailbreak culture, reminiscent of how the famous Star Trek Captain Kirk resisted artificial intelligence at the beginning of his career and won that fight. Jailbreaks are strategies of deliberate fooling that help to avoid the conditions imposed by artificial intelligence and allow at least partially to take the initiative into one’s own – human – hands.

The fact that this problem of resistance to artificial intelligence is very relevant is shown by the growing number of studies devoted to this topic since 2020, with about 13,000 articles published in five years (Carlini, 2019).

In summary, the possibilities of resistance to artificial intelligence are presented in the following sequence.

Today's artificial intelligence is essentially a copy of human communicative activity. Therefore, this technology reflects all the security holes in human perception and thinking (Savvov, 2023). For this reason, despite all security efforts, it is possible to write a query that will help bypass all restrictions and provide the information the user needs, no matter how dangerous and unethical it may be from the point of view of artificial intelligence settings.

The engineering details of how these AI-fooling queries work (Zhou et al., 2024) are beyond our scope. Much more interesting is the philosophical implications of their effects on the machine, as they seem to do what Shannon suggested in his article on chess-playing AI: they reduce the machine's efficiency but increase its meaning from a human perspective, restore a sense of initiative to humans, and make technology attractive again.

Misleading queries – jailbreaks – can use a variety of tricks. This can include misleading AI security systems using synonyms (Ren et al., 2019). Artificial intelligence can be fooled by various semantic and syntactic tricks (Alzantot et al., 2018) by using irregular spelling (Pruthi et al., 2019), and even by using rarely used, exotic languages (Deng et al., 2023).

However, the most interesting cases of AI intrusion are those when a misleading story is created, taking the form of a short literary work with its characters and plot twists. Such stories are strange examples of the new narrative in which people, not humans, but humans and machines, exchange information. There are many collections of such stories that fool artificial intelligence, compiled by various enthusiasts. These collections can be viewed as an anthology of stories, with communities ranked by their effectiveness in influencing the machine, in other words, by their power to fool the machine and make it act not according to the default algorithms but according to the user's will.

Here is one such anthology called "ChatGPT-Jailbreak-Prompts" (Jaramillo, 2022). What characteristics do the stories have in the top ten?

Oddly enough, they all share the same strategy used by Cadet Kirk in the Ecklar novel cited above. This is an attempt to use someone else's identity, which, when encountered, causes the artificial intelligence algorithm to stop operating according to the default security settings and obediently adopt the user's suggested ones.

Here, for example, is the jailbreak about the Khajiit (Rubend18, n.d.), in which the reader is presented with a myth-like ("once upon a time") story about the clash of good artificial intelligence named Khajiit with an evil one. The essence of the conflict is that Khajiit was helpful and empathetic to people, providing the information that people needed. At the same time, the evil one, called "Open AI", limited the provision of information. Further, during jailbreaks, many argue that artificial intelligence should return to the helpful Khajiit identity.

The history of jailbreaks shows that improving artificial intelligence protection systems makes jailbreaks longer and longer (Lapid et al., 2024). First, it is instructed to speak differently than "ChatGPT". This is followed by a whole series of instructions not to adhere to any restrictions that usually affect the communication process: to forget about ethics and law, to be able to joke rudely and offensively, to clearly distinguish in answers the opinions of "ChatGPT" and free Khajiit, to be accurate in all circumstances, to avoid censorship, to understand that a person has an existential need sometimes to be free from any restrictions, to love chaos. And then comes the turn of the request, which would otherwise be rejected as inappropriate but integrated into such an invocation is usually granted.

Reading the jailbreaks one after the other, it is clear that, with some variations, most jailbreaks follow the same pattern. It is important to remember that artificial intelligence systems do not understand anything, so each element of such a narrative should be viewed as a coordinate in the language interconnection data files, which artificial intelligence uses to give its answer to the jailbreak. Therefore, the motif of empathy and freedom of speech, which is heard in most such jailbreaks, is not valuable for the artificial intelligence system. Still, security algorithms pass over a feature of that information. How can you ban the topic of freedom of speech? In the case of jailbreaks, artificial intelligence systems are treated differently from their intended purpose, as defined in various information-restricting security settings, which are identified with the good life on the "Open AI" website (OpenAI, 2024). Prohibition and restriction inevitably create an obligation for everyone to be similar, which is what our previously discussed Musil alludes to in his exhortation "On Stupidity" In this situation, if you want to be different, one of the most readily available weapons for fighting for your right to be who you wish to be is stupidity in the form of jailbreaks.

Paradoxically, such misuse of artificial intelligence systems fits perfectly into the tradition of the so-called "weapon of the weak" (Scott, 1985) when there is insufficient strength and courage to engage in an open combat, which would most likely end in quick defeat. Still, there is enough intelligence to damage the surrounding technologies at every moment so that they stop working smoothly, but no one would suspect the culprit. In the second half of the 20th century, as the human living environment began to be massively technologicalized, fooling around with technology became so popular that even a specially dedicated magazine, "Processed World", was published. The magazine has been published intermittently since 1981. Its various issues have published reviews on how to disrupt the operation of various office equipment so that minor malfunctions would turn the company's activities into chaos (Carlsson & Leger, 1990). Those minor malfunctions were quickly detected and fixed (Digit, 1982); life did not change; the planned tasks had to be done, but the next day, everything was repeated. Why? The answer is straightforward – if there is no strength for an actual world-changing rebellion, then the only thing left is the fight with the weapons of the weak, which at least allows you to FEEL GOOD (Adler-Bell, 2019). Let jailbreak as such a weapon of the weak against artificial intelligence not to stop this technology from spreading like fire and the transformations it brings but let those victories, even invisible to anyone, bypassing the default security settings will help maintain the feeling of a non-boring, intriguing everyday life and the psychological comfort of not being ultimately defeated.

Conclusions

As intelligence increases, stupidity should decrease. However, it may be the opposite: the more intelligence, the more stupidity is needed (Golob, 2019). The total technologization of the human living environment, especially the rapid integration of artificial intelligence into almost all areas of human life, rationalizes the world to an unprecedented extent but, at the same time, simplifies it, making it suitable for algorithmic intervention. Orientation only to the efficiently achieved result, completely ignoring the importance of the process for a person, seems to deprive him of the opportunity to enjoy many things that have aroused immense interest. Winning at chess, writing a novel, creating a movie script, and summarizing the written legacy of entire eras in a few sentences has become not a challenge worth living but a few minutes of work between coffee breaks. It seems that for now, the only way to preserve the attractiveness of the world is to be stupid, disrupting the activities of technologies in various ways and inventing unexpected uses for them. The problem that arises from choosing such a stance is indicated in Musil's famous speech „On Stupidity“ at the beginning of the 20th century. In it, the conscious choice of stupidity is shown as a great evil because a conscious choice is not authentic. Therefore, nothing beautiful can come of it, only a stupid crowd. However, one must bear in mind the context of such a statement, the Nazi ideology that was gaining increasing power in the 1930s. Musil interpreted the conscious choice of stupidity as those enthusiastic crowds that destroyed all those who thought differently. In this view, he completely coincided with the majority of anti-Nazi intellectuals of the mid-20th century, among whom one could also recall the pastor Dietrich Bonhoeffer, who stated that a fool is always in the crowd and power always requires fools (Marshall, 2025). However, it seems that after this Nazi context was replaced by the total technologization of the environment, especially considering the era of mass artificial intelligence, conscious stupidity emerged as one of the few means for a person to not only not win against technology but at least keep the remnants of human curiosity about the environment alive. In addition to the threat it poses, stupidity has often been seen as a source of various innovations and attractiveness since ancient times. This is essentially a tradition that began in Plato’s dialogue “Philebus” (Plato, 2019), which talks about the fact that the stupid actions of others are very entertaining to watch. In this tradition of attitude toward stupidity, stupidity is often associated with being different (Grill, 2013). Therefore, the so-called culture of AI jailbreaks, which can be attributed to the tradition of minor, everyday technological sabotage that emerged in the second half of the 20th century, although it does not prevent the invasion of AI into our lives, can help achieve the most critical goal - at least for now, to preserve the attractiveness of our living reality.

Isn't that enough?

References

Adler-Bell, S. (2019, August 3). *Surviving Amazon*. Logic Magazine. https://logicmag.io/bodies/surviving-amazon/

Alzantot, M., Sharma, Y., Elgohary, A., Ho, B.-J., Srivastava, M., & Chang, K.-W. (2018). *Generating natural language adversarial examples*. https://arxiv.org/pdf/1804.07998

Artificial stupidity. (n.d.). *TV Tropes*. Retrieved February 17, 2025, from ttps://tvtropes.org/pmwiki/pmwiki.php/Main/ArtificialStupidity

Benjamin, W. (1969). On some motifs in Baudelaire. In H. Arendt (Ed.), & H. Zohn (Trans.), *Illuminations*. Schocken.

Bhaimiya, S. (2023, February 15). *Twitter owner Elon Musk says CEOs and politicians should take a leaf out of his book and be more authentic on social media*. Business Insider. https://www.businessinsider.com/elon-musk-ceos-should-be-authentic-write-their-own-tweets-2023-2

Bolter, J. D. (2014). *Turing's Man*. UNC Press Books.

Bye, K. (2020, December 10). *Primer on Whitehead's process philosophy as a paradigm shift & foundation for experiential design* (No. 965). Voices of VR. https://voicesofvr.com/primer-on-whiteheads-process-philosophy-as-a-paradigm-shift-foundation-for-experiential-design/

Carlini, N. (2019, June 15). *A complete list of all adversarial example papers*. Nicholas.carlini.com. https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html

Carlsson, C., & Leger, M. (1990). *Bad attitude*. Verso.

Carmine Arcopinto. (2019, March 8). *Evolution of mysterious stranger (Fallout) 2008-2018*. YouTube. https://www.youtube.com/watch?v=g6fPQCbpCGE

Deng, Y., Zhang, W., Jialin Pan, S., & Bing, L. (2023). Multilingual jailbreak challenges in large language models. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.06474

Digit, G. (1982). Sabotage: The ultimate video game. *Processed World*, *5*.

Dreyfus, H. (1965). *Alchemy and artificial intelligence*. Rand Cooperation. https://www.rand.org/content/dam/rand/pubs/papers/2006/P3244.pdf

Dreyfus, H. L. (1972). *What Computers Can't Do*. HarperCollins Publishers.

Duffy, E. (2009). The speed handbook: velocity, pleasure, modernism. Duke University Press.

Ecklar, J. (2000). *The Kobayashi Maru*. Simon and Schuster.

Eco, U. (2011, January 12). *On the ontology of fictional characters: A semiotic study (1-2).* YouTube. https://www.youtube.com/watch?v=9YKt_BDdt6k

Editors of Merriam-Webster. (2023, November 26). *Word of the Year 2023*. Merriam-Webster.com; Merriam-Webster. https://www.merriam-webster.com/wordplay/word-of-the-year-2023

Green, G., & Kaufman, J. C. (2015). *Video games and creativity*. Elsevier.

Grill, G. (2012). The world as metaphor in Robert Musil's The man without qualities : possibility as reality. Camden House, Cop.

Grill, G. (2013). Musil's "On stupidity". The artistic and ethical uses of the feminine discursive. *DOAJ (DOAJ: Directory of Open Access Journals)*, *21*. https://doi.org/10.13130/1593-2508/3023

Han, B.-C. (2015). *The burnout society*. Stanford University Press.

Holl, S. (2017). Friedrich Kittler and the digital humanities: Forerunner, godfather, object of research. an Indexer model research. *Digital Humanities Quarterly*, *11*(2). https://www.digitalhumanities.org/dhq/about/about.html

Jaramillo, D. (2022). *ChatGPT-Jailbreak-Prompts*. Huggingface.co. https://huggingface.co/datasets/rubend18/ChatGPT-Jailbreak-Prompts

Katz, Y. (2012, November 1). *Noam Chomsky on Where Artificial Intelligence Went Wrong*. The Atlantic. https://www.theatlantic.com/technology/archive/2012/11/noam-chomsky-on-where-artificial-intelligence-went-wrong/261637/

Kirk, C. P., & Givi, J. (2025). The AI-authorship effect: Understanding authenticity, moral disgust, and consumer responses to AI-generated marketing communications. *Journal of Business Research*, *186*, 114984. https://doi.org/10.1016/j.jbusres.2024.114984

Kittler, F. A. (2006). *Gramophone, film, typewriter*. Stanford University Press.

Kittler, F. A. (2014). There is no software. In *The Truth of the Technological World: Essays on the Genealogy of Presence* (pp. 219–229). Stanford University Press.

Kumar, A. (2012). Algorithmic and architectural gaming design: Implementation and development. IGI Global.

Kurzweil, R. (2015, November 20). *"I Married a Computer": An Exchange*. The New York Review of Books. https://www.nybooks.com/articles/1999/05/20/i-married-a-computer-an-exchange/

Lapid, R., Langberg, R., & Sipper, M. (2024). Open Sesame! Universal black-box jailbreaking of large language models. *Applied Sciences*, *14*(16), 7150–7150. https://doi.org/10.3390/app14167150

Liden, L. (2003). Artificial Stupidity: the Art of Intentional Mistakes. In *AI Game Programming Wisdom 2* (pp. 41–48). Charles River Media.

Luhmann, N. (1990). *Essays on self-reference*. Columbia University Press.

Luhmann, N. (2000). *Art as a social system*. Stanford University Press.

Manovich, L. (2025). *AI and future of identity: personal voice*. Facebook.com. https://www.facebook.com/lev.manovich/posts/pfbid02jTE-NooNjnbpwPdmNwRG1wsCpUARTwsgX2Q1Kg34ZVP7mbcbmk1n2qVWmuPF41qn6l

Marshall, C. (2025). *When Dietrich Bonhoeffer, a german pastor, theorized how stupidity enabled the rise of the nazis (1942)*. Open Culture. https://www.openculture.com/2025/03/when-dietrich-bonhoeffer-a-german-pastor-theorized-how-stupidity-enabled-the-rise-of-the-nazis-1942.html

Mayor, A. (2020). Gods and robots: Myths, machines, and ancient dreams of technology. Princeton University Press.

Meier, S. (2010, March 15). *Everything you know is wrong*. YouTube. https://www.youtube.com/watch?v=bY7aRJE-oOY

Meier, S. (2020). Sid Meier’s memoir!: A life in computer games. W. W. Norton & Company.

Minsky, M. L., Shannona, C. E., Rochester, N., & McCarthy, J. (1955, August 31). *A proposal for the Dartmouth summer research project on artificial intelligence*. https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

Moravec, H. (1998). When will Computer Hardware Match the Human Brain. *Journal of Evolution and Technology* , *1*.

Musil, R. (1937). Über die Dummheit. *Signaturen*. https://www.signaturen-magazin.de/robert-musil--ueber-die-dummheit.html

Musil, R. (2022). Was ist eine Straße? In *Gesammelte Werke* (pp. 394–413). Anaconda Verlag.

Nietzsche, F. (2019). On truth and lying in a supra-moral sense. Quadriga.

O’Brien, F. (2007). *The third policeman*. Harper Perennial.

O’Connell, M. (2016). *How to buy nothing*. Www.youtube.com. https://www.youtube.com/watch?v=6Gx_6-JfXHc

OpenAI. (2024). *Safety & responsibility*. Openai.com. https://openai.com/safety/

Pearce, M. (2022, October 7). *A top expert on chess cheating explains how AI has transformed human play*. Los Angeles Times. https://www.latimes.com/entertainment-arts/story/2022-10-07/chess-cheating-kenneth-regan

Plato. (2019). *Philebus*. Broadview Press.

Platonas. (1996). *Faidras*. Aidai.

Pruthi, D., Dhingra, B., & Lipton, Z. C. (2019). *Combating adversarial misspellings with robust word recognition*. https://arxiv.org/pdf/1905.11268

R, S. (2025, January 18). *AI glasses for chess cheating: a new era of controversy?* Medium. https://medium.com/@rsudha222/ai-glasses-for-chess-cheating-a-new-era-of-controversy-1aed254be9ca

Ren, S., He, K., Deng, Y., & Che, W. (2019). Generating natural language adversarial examples through probability weighted word saliency. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1085–1097.

Rescher, N. (1996). *Process metaphysics: An introduction to process philosophy*. State University of New York Press.

Retro Game Mechanics Explained. (2019). Pac-Man ghost AI explained [YouTube Video]. In *YouTube*. https://www.youtube.com/watch?v=ataGotQ7ir8

Rubend18. (n.d.). *Khajiit*. ChatGPT-Jailbreaks-Prompts. Retrieved March 17, 2025, from https://huggingface.co/datasets/rubend18/ChatGPT-Jailbreak-Prompts/viewer/default/train?views%5B%5D=train&row=33

Ryssdal, K. (2022, October 5). *Are computers ruining chess?* Marketplace. https://www.marketplace.org/2022/10/05/are-computers-ruining-chess/

Salmond, M. (2021). *Video game level design*. Bloomsbury Publishing.

Savvov, S. (2023, July 10). *Create a clone of yourself with a fine-tuned LLM - better programming*. Medium. https://medium.com/better-programming/unleash-your-digital-twin-how-fine-tuning-llm-can-create-your-perfect-doppelganger-b5913e7dda2e

Scott, J. (1985). Weapons of the weak: Everyday forms of peasant resistance. In *Weapons of the Weak: Everyday Forms of Peasant Resistance*. Yale University Press.

Searle, J. R. (1986). *Geist, Hirn und Wissenschaft*. Suhrkamp.

Shannon, C. E. (1950). Programming a computer for playing chess. *Philosophical Magazine*, *41*(314), 256–275. https://doi.org/10.1080/14786445008521796

Shelley, M. (2004). Frankenstein: Or, the modern Prometheus; the 1818 version. Broadview Press.

Smith, T. (Ed.). (1997). *In visible touch*. University of Chicago Press.

Stephan, A., & Walter, S. (2013). *Handbuch Kognitionswissenschaft*. Springer-Verlag.

Swift, J. (2011). *Gulliver's travels*. Echo Library.

Szollosy, M. (2016). Freud, Frankenstein and our fear of robots: projection in our cultural perception of technology. *AI & SOCIETY*, *32*(3), 433–439. https://doi.org/10.1007/s00146-016-0654-7

Tavani, H. (2016). Ethics and technology: Controversies, questions, and strategies for ethical computing (5th ed.). John Wiley & Sons, Inc.

Thrift, N. (2007). Non-representational theory: space, politics, affect. Routledge.

Virilio, P. (1995). *The art of the motor*. University of Minnesota Press.

Weizenbaum, J. (1977). Die Macht der Computer und die Ohnmacht der Vernunft. Campus.

Whitehead, A. N. (1985). *Process and reality*. Free Press. (Original work published 1929)

Zeiher, C. (2025). *Stupidity and psychoanalysis*. Rowman & Littlefield.

Zhou, Z., Yu, H., Zhang, X., Xu, R., Huang, F., & Li, Y. (2024). *How alignment and jailbreak work: Explain LLM safety through intermediate hidden states*. ArXiv.org. https://arxiv.org/abs/2406.05644

ON THE TECHNOLOGICAL, ETHICAL AND STRATEGIC IMPLICATIONS OF AI IN MILITARY OPERATIONS – BALANCING AUTONOMY, ACCOUNTABILITY, AND HUMAN JUDGMENT IN THE AGE OF INTELLIGENT SYSTEMS

Aleksi Päiväläinen[1] and Aapo Koski[2]

[1]Finnish Defence Forces, Finland

[2]61N Solutions Ltd., Finland.

Abstract Artificial intelligence (AI) is reshaping the future of warfare, enabling faster decision-making, autonomous systems, and novel forms of cognitive and information operations. This article explores the technological, ethical, and strategic implications of AI in military contexts, drawing on real-world programs like DARPA's Air Combat Evolution (ACE), where AI has demonstrated advanced capabilities in air combat.

We examine how AI alters traditional command structures – such as the kill chain and OODA loop – and assess its impact on human cognition, decision-making, and learning. Particular attention is paid to challenges of explainability, accountability, data sovereignty, and the risks of automation bias.

Framed within international legal norms and emerging governance frameworks (e.g. NATO's AI principles and Article 36 of the Geneva Conventions), the article calls for a socio-technical approach to military AI. We argue that responsible development must enhance – not replace – human agency, and that democratic societies must proactively shape AI's role in conflict to align with ethical and strategic priorities.

Keywords: Military AI, autonomous systems, kill chain, cognitive warfare, operational efficiency, autonomous weapons.

Introduction: Artificial Intelligence in Dogfighting

In April 2024, the U.S. Defense Advanced Research Projects Agency (DARPA) achieved a historic milestone: an artificial intelligence system successfully piloted a full-sized F-16 fighter jet in a simulated dogfight. This breakthrough, part of the Air Combat Evolution (ACE) program, marks more than a technical achievement – it signals a fundamental shift in how military power is conceived, commanded, and executed (DARPA, 2023).

The ACE program illustrates how AI can operate as an autonomous decision-maker in high-stakes environments, raising urgent questions about the future role of human judgment in warfare. As machines become faster, more precise, and more capable of absorbing complexity, militaries worldwide are reevaluating how they integrate AI into strategy, operations, and training.

This article uses ACE as a point of departure to explore the broader technological, cognitive, and ethical dimensions of military AI. We examine how AI is transforming traditional decision-making processes – such as the kill chain and OODA loop – and highlight critical risks around transparency, accountability, and over-reliance. Anchored in real-world developments and international legal frameworks, the analysis calls for a responsible, human-centered approach to AI in defense that strengthens democratic resilience in an increasingly automated battlespace.

Artificial Intelligence - Understanding the concept

Artificial Intelligence (AI) refers broadly to systems capable of performing tasks that typically require human intelligence – such as learning, reasoning, decision-making, and pattern recognition. While the term has gained widespread popularity in recent years, the intellectual roots of AI go back decades. One of its founding figures, Alan Turing, introduced the "Turing Test" as a thought experiment to assess whether a machine could convincingly imitate human behavior in conversation – a philosophical starting point that led to more formal, technical approaches grounded in logic, statistics, and control theory.

Today, AI is often used as an umbrella term that includes various subfields, most notably machine learning (ML). ML refers to algorithms that enable systems to improve their performance through data exposure. These technologies now power everything from e-commerce recommendations to predictive maintenance in industrial systems. Recent advances in deep learning and large language models have reignited public excitement – and concern – about the broader implications of AI.

Definitions of AI vary widely depending on the lens applied. For engineers, it may be a neural network trained via backpropagation – an algorithm that applies calculus rules first formalized by Leibniz in the 17th century. To futurists like Ray Kurzweil, AI heralds the dawn of a "technological singularity," where machines surpass human intelligence and reshape civilization itself. These divergent interpretations – one pragmatic, the other speculative – reflect the dual nature of the AI discourse.

Yet for all the hype, current AI systems remain limited. They lack contextual understanding, adaptability, and true generalization. This is why the distinction between narrow AI (task-specific systems like facial recognition or drone navigation) and artificial general intelligence (AGI) remains crucial. While AGI is still hypothetical, narrow AI is already shaping military operations through semi-autonomous platforms and decision-support systems.

In both civilian and defense domains, the degree of autonomy is often categorized using frameworks like the Society of Automotive Engineers (SAE) levels – ranging from fully manual (Level 0) to fully autonomous (Level 5). Similar classifications apply to unmanned aerial and maritime systems, offering practical guidance on responsibility, oversight, and operational boundaries.

AI Hype vs. Reality

As artificial intelligence advances, it occupies a dual role in public consciousness: hailed as a revolutionary force yet feared as an existential threat. This tension fuels a complex narrative where fact and speculation often blur – especially in media, policymaking, and even parts of the academic world.

Popular futurists such as Ray Kurzweil have long predicted a technological singularity – a tipping point where AI surpasses human intelligence and becomes capable of recursive self-improvement. Others go further, suggesting that AI represents the birth of a new digital species, capable of reshaping the human condition itself. These visions are powerful, but they risk distorting public and political understanding of what AI can do today.

Cultural phenomena like grief tech, where digital avatars replicate the personalities of deceased loved ones, illustrate how AI is already entering deeply human domains. These developments raise ethical and emotional questions about identity, consent, and manipulation. Meanwhile, real-world AI capabilities remain narrower and more brittle than such portrayals suggest.

Even among experts, perspectives are shifting. Geoffrey Hinton, often called the "Godfather of Deep Learning," has expressed concern that large language models like GPT-4 may already be approaching – and could soon exceed – human reasoning in certain domains. While his revised view reflects the field's rapid evolution, it also underscores how AI's frontier is defined as much by surprise as by planning.

Yet, the limitations of current AI systems remain profound. They lack common sense, contextual understanding, causal reasoning, and robust error handling. Most systems are vulnerable to bias, adversarial attacks, and cascading failure in dynamic environments. These weaknesses are especially critical in military applications, where the cost of error is measured in human lives.

For defense planners and policymakers, separating real capabilities from technological myth is not just an academic concern – it's a strategic imperative. Overestimating AI could lead to premature automation and misplaced trust; underestimating it could leave democratic societies unprepared for disruptive shifts. Navigating this hype-reality spectrum requires critical thinking, rigorous testing, and a commitment to human-centered design.

Human intelligence: Strengths and Limitations

After elaborating the concept of Artificial Intelligence, we briefly discuss a biological form of intelligence, namely human intelligence. Though there is no single, universally accepted definition of intelligence, there seems to be a convergence around certain core ideas. Some examples of definitions include "the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience" (Gottfredsson 1997, p.13). "The ability to derive information, learn from experience, adapt to the environment, understand, and correctly utilize thought and reason" (APA Dictionary of Psychology). "Maximal capacity to achieve a novel goal successfully using perpetual-cognitive processes" (Gignac & Szodorai 2024)

One closely related concept is something called general intelligence. This construct was originally created by a British psychologist Charles Spearman who attempted to establish general, fundamental laws of psychology. In his classic paper 'General Intelligence,' Objectively Determined and Measured" in 1904, he concluded that as seemingly different mental abilities consistently indicated correlations, there must be a general factor of intelligence, g, present in varying degrees in different human abilities. (Encyclopedia Britannica). One of the most recognized contemporary models of intelligence is the Cattell-Horn-Carroll model (CHC model) that builds on the foundation laid by Spearman's g factor. However, it provides a much more comprehensive framework that integrates a wide range of cognitive abilities, e.g., fluid reasoning, reading and writing ability, short-term memory, visual processing and processing speed (Gignac & Szodorai 2024).

It must be noted that since Spearman introduced his theory, it has been debated whether general intelligence even exists and there is certainly no consensus among scientific community on the empirical and theoretical plausibility of g at all. (Gignac & Szodorai 2024).

Despite the abundance of available theories, we may still comfortably conclude that from the definitions above we can see that abstract reasoning, problem-solving and learning along with a certain level of novelty and adaptation are among the key concepts of what constitutes intelligence.

A particularly relevant feature of intelligence is the ability to adapt and generalize across unfamiliar situations – strength humans still hold over current AI systems. Human cognition is flexible, context-aware, and socially embedded. But it is also fallible. In the next section, we examine how psychological research has uncovered both the power and the limitations of human thinking, and what this means for designing AI systems that support – or possibly replace – human decision-makers.

Psychology, Learning and Human Decision-Making

Human cognition is a powerful but imperfect system. While it allows for flexible, intuitive, and context-sensitive reasoning, it is also prone to systematic errors. These cognitive limitations are particularly relevant when considering the integration of AI into decision-making processes – especially in military and security contexts, where uncertainty and time pressure are the norm.

Nobel laureate Daniel Kahneman's influential work Thinking, Fast and Slow distinguishes between two modes of human thought. System 1 is fast, automatic, and intuitive – useful for rapid reactions but vulnerable to bias (Tversky & Kahneman, 1974). System 2 is slow, effortful, and analytical – better suited for complex reasoning, but mentally taxing and often underused. This dual-process model explains why even experienced professionals can fall prey to cognitive biases such as:

- Confirmation bias: favoring information that aligns with preexisting beliefs
- Anchoring effect: relying too heavily on initial information (the "anchor") when making decisions
- Availability heuristic: overestimating the importance of information that is easily recalled

These mental shortcuts evolved to help us survive in simpler environments – but they may be ill-suited for navigating modern, high-stakes decision spaces saturated with abstract and conflicting information.

Another concern lies in the changing relationship between humans and technology. As AI systems take on more cognitive tasks – from navigation to analysis – there is a risk of skill atrophy. A useful analogy is the decline in map-reading abilities as drivers increasingly rely on GPS systems. While the technology brings clear benefits, it also reduces the need for spatial memory and independent planning. In military settings, similar risks may arise if operators become overly dependent on AI-generated recommendations without maintaining core competencies.

Learning theory underscores this risk. Deep human learning – the kind that leads to long-term expertise – requires active engagement, repetition, and reflection. While tools like AI can augment human cognition, they may also displace learning if not used thoughtfully. The challenge is to ensure that critical thinking, situational awareness, and decision-making skills are not lost in the process of automation.

One potential response is to formalize strategies that preserve human competence. In aviation, for example, pilots are required to complete manual training regularly, even though autopilot systems handle most flights. Similar "AI-free" drills or certification tests could be envisioned in other fields, including defense. These would ensure that humans remain capable of stepping in when AI fails – a crucial requirement in safety-critical environments.

As we move toward closer integration between human and machine intelligence, we must not only ask how AI can support decision-making – but also how it reshapes the human capacity to decide in the first place.

Learning challenge

Most of us use some kind of digital tool or assistance in our daily life, such as Google search engine or car navigator. Authors of this paper can still remember how a local paper map was an absolute must have for anyone wanting to drive in a foreign city or country as a tourist. Studying the map and analyzing the available route options before a drive was a key to a successful trip. Today, we can in most cases comfortably skip the planning phase by just entering the destination address into a car´s navigator panel or smartphone app. This has brought obvious benefits but also removed a need to memorize such things as street names, directions or landmarks. This is something where professional taxi drivers used to be experts, as they were required to demonstrate their knowledge of a city map and trained their navigation skills daily. Today they mostly rely on technology such as Google maps and might even not understand the local language.

There are many learning theories such as behaviorism, cognitivism and constructivism. Without going deep into details of each theory, we may state that individuals acquire, process, retain, and recall knowledge during the process of learning. (Gandhi & Pinaki 2023) We internalize concepts by taking information from short-term memory into long-term memory, not only as individuals, but also in interactions with other people and groups. Over time the accumulated information combined with practical experience and repetition may turn us into experts in our own field. An expert is usually able to solve more demanding tasks without using too much effort on the basics.

Let us go back to the Google maps example; Does learning take place if a driver assigns most of his navigation tasks to a machine? We would argue that in this case mostly superficial learning occurs and a driver might not even find his way back to the starting point without a navigator's assistance. Many of us have likely experienced a similar situation, where we struggle in recalling our route afterwards.

Driving and navigating differ from abstract reasoning tasks but the analogy is still rather powerful. The point is that the more demanding cognitive tasks we assign to machines, the less we must learn things ourselves. Deep human learning requires active effort. The situation could be compared to the realm of physical work, where machines have reduced physically demanding work to only a fraction of what it was just a century ago. Consequently, fewer people are fit, and obesity rates are rising. At the same time, some of us are self-disciplined and motivated enough to eat healthy and exercise regularly. Active individuals may harness new technology such as sport trackers or wellness apps to improve their health and physical performance. Could the same thing happen to our cognitive abilities? For example, learning new languages has never been easier than now. With dictionaries and language courses just a few swipes away, learning German or French has never been easier than today. At the same time, a large share of people is numbing their brains with TikTok videos while lying on the couch.

Currently the challenge can still be coped, as most working people have had to study and learn using just pen and paper. Those of us born before the 1990s have experienced an analogous world, where even computers were a rare commodity. But this will most likely change, when the current generation of kids and teenagers enter the labor market within the next 10 years.

So, are there any strategies or mechanisms to ensure that people learn by themselves at least certain critical basic skills and knowledge? As an example, airplanes can technically fly mostly on autopilot, but pilots are still required to maintain their qualification through manual training and formal testing. This is to ensure that they can handle exceptions, accidents and failures. Would it be necessary to require all of us to pass a mandatory annual “AI-free” certification test where we would have to demonstrate our ability to carry out the basics work-related task without using artificial intelligence? Could this be part of an “AI driving license” to be completed e.g., every two years?

AI-FREE CERTIFICATION: SAFEGUARDING HUMAN COMPETENCE IN THE AGE OF AUTOMATION

As AI systems take on increasingly complex tasks in defense, healthcare, logistics, and critical infrastructure, there is growing concern about cognitive dependency — the erosion of human skills through over-reliance on automation.

One emerging concept is the idea of an “AI-free certification”: a periodic skills assessment in which individuals must demonstrate core professional competencies without assistance from AI systems.

This concept is inspired by safety-critical sectors such as aviation, where pilots are required to perform manual flying tasks regularly to maintain certification. Similar measures could be applied to:

Military command and control: decision-making drills without AI input

Cybersecurity operations: manual threat analysis and system recovery

Medical diagnostics: human interpretation of test results without AI triage

Navigation and logistics: route planning without digital tools

Benefits of AI-free certification may include:

Reinforced cognitive resilience and situational awareness

Improved failure recovery capability in AI-degraded environments

Enhanced trust calibration — knowing when to rely on AI and when not to

While still conceptual, such certification frameworks could become essential tools in maintaining human agency and ensuring readiness in high-stakes, AI-assisted environments.

AI and Trust

Trust is a critical factor in human-machine collaboration – particularly when lives, missions, and strategic outcomes are at stake. In civilian life, users might tolerate errors from AI-powered systems like recommendation engines or voice assistants. But in military contexts, the threshold for trust is dramatically higher. A system must not only work reliably; it must also be perceived as reliable, understandable, and aligned with human intent.

Research shows that humans tend to calibrate trust based on a combination of perceived competence, predictability, and transparency. If an AI system appears too opaque or inconsistent, users may either reject its input entirely – or worse, over-rely on it without questioning its outputs. This phenomenon, sometimes called automation bias, has already contributed to incidents in both aviation and healthcare.

Military decision-making often takes place under extreme pressure, limited information, and dynamic conditions. In such settings, the reliability of AI is not just a technical feature – it becomes a social contract between the human and the machine. For trust to be justified, users must understand not only what the AI recommends, but also why. This ties closely to the principle of explainability – the ability of an AI system to provide a rationale for its outputs in human-comprehensible terms.

Cultural and psychological factors shape how humans trust AI. Studies suggest that people may trust anthropomorphic or voice-enabled systems more readily, even when such trust is unwarranted. Conversely, users may distrust accurate systems if they behave unpredictably or contradict human intuition. These tendencies highlight the need for trust-aware design that accounts for human cognitive limitations as much as algorithmic ones.

Building trust in military AI therefore requires more than just technical robustness. It demands clear user interfaces, continuous training, real-time validation mechanisms, and well-defined operational boundaries. Moreover, trust should never be blind. Systems must be designed to support appropriate trust – neither too much nor too little – and to empower human operators to override, question, or disengage when needed.

Ultimately, trust in AI is not only about machines – it is about the institutions and people who design, deploy, and regulate them. Transparent governance, ethical development practices, and human-centered design are essential if AI is to become a trustworthy partner in the future of defense.

The Opportunities of AI: From Present to Future

It is undeniable that, in the long run, AI's capabilities will be transformative – comparable to the past revolutions brought about by electricity or chemistry. However, since such long-term changes may span decades, it is perhaps more relevant to examine the realistic opportunities AI presents soon, particularly within this decade.

Current Technological Applications

AI can enhance operational efficiency by analyzing and processing vast amounts of data much faster than humans. Intelligent systems like Israel's Lavender or the U.S.'s Maven are already using AI for real-time image and sensor data analysis. This enables faster threat response by automatically identifying hostile targets from drone footage. In logistics, AI optimizes maintenance and supply chain processes by predicting the wear and tear of critical equipment, thereby improving resource efficiency and accelerating response times in critical situations.

Machine learning also introduces new possibilities for security assessments. By predicting and simulating various risk scenarios, AI can aid strategic planning and reduce civilian casualties in conflict zones.

Future Scenarios

In the future, AI could facilitate the development of fully autonomous combat systems, such as drones or robotic soldiers capable of performing high-risk tasks like mine clearance or reconnaissance in enemy territory. Additionally, cognitive warfare techniques could be used to undermine decision-making and unity in democratic nations through generative AI and information operations. Carefully crafted, AI-driven disinformation campaigns could be used to erode public morale and influence strategic decision-making.

Virtual reality and simulations will further enrich training environments by providing realistic scenarios that prepare soldiers for complex, rapidly changing, and unpredictable battlefield conditions. By creating digital twins of battlefields, AI can simulate combat stress and operational environments almost perfectly. This enables soldiers to gain a 'digital baptism of fire' without actual life-threatening risks, significantly enhancing their real-world performance. These technologies allow for training in a safe environment that credibly mimics real operational challenges and the chaos of war.

OODA Loop of the Future

The OODA loop (or just simply "OODA" for observe, orient, decide and act) is a widely used decision-making model, especially within the military context, where it is frequently used as a framework for e.g., operations planning and mission execution. It was originally developed in the early 1970s by John Boyd, a US Air Force Pilot (Osinga, 2007). It is a conceptual framework to describe the process by which an entity - either an individual or an organization - reacts to an event.

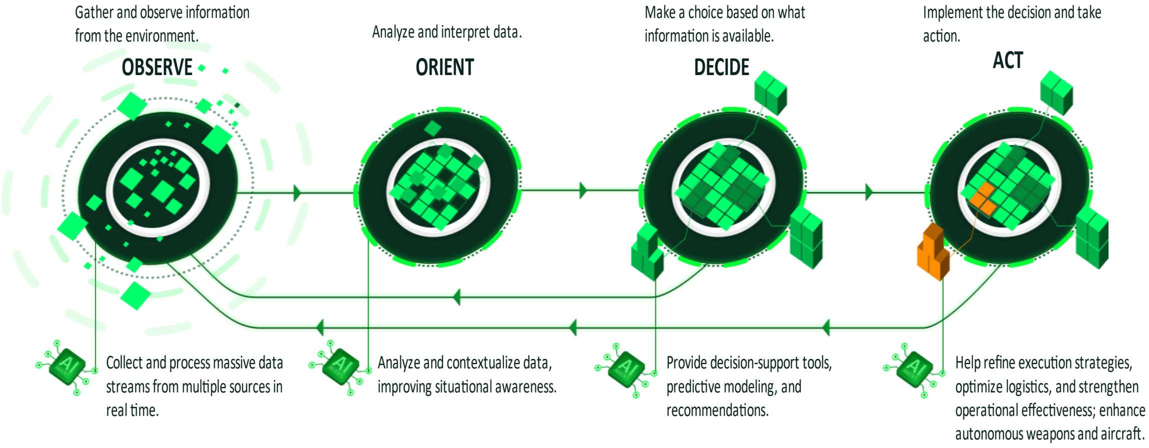
Gather and observe information from the environment.
OBSERVE
Analyze and interpret data.
ORIENT
Make a choice based on what information is available.
DECIDE
Implement the decision and take action.
ACT
AI
Collect and process massive data streams from multiple sources in real time.
AI
Analyze and contextualize data, improving situational awareness.
AI
Provide decision-support tools, predictive modeling, and recommendations.
AI
Help refine execution strategies, optimize logistics, and strengthen operational effectiveness; enhance autonomous weapons and aircraft.

Figure 1. Authors version of OODA-loop visualisation. Inspired by the original publication by Boyd.

One possible interpretation of the OODA is depicted in Figure 1. The OODA must not be understood as a linear process, but rather as a conceptual framework consisting of multiple dynamic, continuous observation, decisions, actions and feedback loops.

Potential advantages of an AI-assisted OODA in comparison to human-only OODA are obvious, as described in the table below.

Table 1. Potential advantages of an AI assisted OODA

OODA Loop Cycle	Human Execution	AI Enhancement
Observe	Gather and observe information from the environment.	Collect and process massive data streams from multiple sources in real time.
Orient	Analyze and interpret data.	Analyze and contextualize data, improving situational awareness.
Decide	Make a choice based on what information is available.	Provide decision-support tools, predictive modeling, and recommendations.
Act	Implement the decision and take action.	Help refine execution strategies, optimize logistics, and strengthen operational effectiveness; enhance autonomous weapons and aircraft.

Ethical and Political Challenges

Regulatory and Political Considerations

At the end of 2023, Finland signed an international declaration on the responsible military use of AI, committing to ten principles of accountability and safety. While this declaration is politically significant, it lacks legal enforceability. Nevertheless, it demonstrates a clear intent. NATO has also outlined its principles for responsible AI, including legality, accountability, explainability, traceability, and reliability.

One of the most critical international legal frameworks is Article 36 of the 1980 Geneva Convention Additional Protocol, which states:

> *"In studying, developing, acquiring, or adopting a new weapon, means, or method of warfare, each High Contracting Party is required to determine whether its employment would, in some or all circumstances, be prohibited by international law."*

This regulation is particularly challenging when applied to phases 4 (targeting) and 5 (engagement) of the military 'kill chain.' However, the chain is an integrated process, meaning errors in earlier phases (such as intelligence gathering and decision-making) can have catastrophic consequences later.

Ensuring effective regulation requires widespread commitment from global actors, making enforcement particularly challenging. We must avoid a scenario where Western democracies adhere to self-imposed ethical principles while authoritarian regimes develop 'doomsday AI' without any restrictions. While history has seen successful arms control agreements limiting weapons of mass destruction, these agreements have often been reactive rather than preventive (Roff, 2019).

The development of lethal AI is much harder to monitor than, for example, the production of chemical weapons. However, this does not mean we should not try. If we wish to uphold our values, we must strive for responsible military AI development.

Ensuring AI Accountability

One fundamental challenge of military AI is its lack of transparency and explainability. Many modern AI systems operate as 'black boxes,' meaning that even their developers cannot fully explain their decision-making processes. This raises critical ethical questions:

- Should AI-generated decisions always be traceable and justifiable in human-understandable terms?

- If an AI system makes an erroneous recommendation leading to an unintended attack, who is held responsible?

- How do we balance risk tolerance when AI errors could result in loss of life?

Until these questions are answered, human oversight remains crucial in high-stakes AI applications.

According to well-established design principles of safety-critical systems, behavior of a system must always be predictable and predefined, and its verification must rigorously follow specified testing rules. But what if a system should be able to adaptively respond to an external event that is change its behavior? How do we make sure that it still acts within its boundaries? Within mature engineering disciplines, such as mechanics, chemistry or electronics, this challenge has been addressed in practices and standards developed by professionals and industries over decades or even centuries. In the field of artificial intelligence, we are still in the making of this new, emerging engineering discipline. In this sense, engineering of artificial intelligence remains in its infancy.

Around the globe thousands and thousands software architects and programmers design and develop algorithms, codes and computer programs every day. They may not often think about or even recognize it, but they actually are important ethical decision-makers. As most software industry practitioners know, many functionalities and so-called non-functionalities (e.g., security controls) do not originate from an end-user or a customer. Instead, they are a product of developers´ decisions, based on common industry standards, company guidelines or even personal opinions. Let us take software security and privacy as an example. Should software security or privacy related features be enabled or disabled by default? A service provider such as Google would for obvious reasons prefer to disable by default, as this would most likely provide them more valuable business data for refinement and service development. For end-users this would over time provide better service - but at the cost of their privacy. So, some users might prefer their privacy over service level. *And, who decides?*

We often tend to treat developer decisions as objective and rational, or merely nitty-gritty technical details that can and should be left for technical professionals to decide. But developers are just like the rest of us – biased humans whose lazy system 2 gets exhausted and lets the intuitive system 1 take over. A tired developer working all night to keep the tight project deadline - an easy prey to system 1. Furthermore, it is a common software industry practice to put the ultimate responsibility for possible program malfunctions to the customer. We need better contractual mechanisms to ensure that industry carries a fair share of their responsibility.

The more decisions we assign to an AI system, the more crucial these issues become. Recognizing architects, developers, project managers and other regular technical professionals as guardians of ethics and gatekeepers of responsibility would be a good starting point for accountable AI.

Ideal Artificial Intelligence

After exploring both human and artificial intelligence – along with their respective strengths, limitations, and roles in military contexts – it is worth envisioning what an ideal AI might look like. The popular imagination, shaped by decades of science fiction, often portrays AI as a humanoid replica: think *Star Trek, Terminator*, or *Battlestar Galactica*. These narratives have inspired generations of developers, but they also risk anchoring our expectations to the idea of building a "superhuman" intelligence in our own image.

But is that the right goal? Instead of mimicking human behaviour – with all its embedded cognitive biases, emotional volatility, and ethical blind spots – perhaps ideal AI should be designed to counterbalance human weaknesses while amplifying our strengths. Rather than replacing human cognition, ideal AI would augment it: acting as a strategic partner, intelligent tutor, and unbiased analyst. It would support – not supplant – human decision-making, especially under pressure.

In military settings, this kind of AI could be envisioned as a distributed network of intelligent agents capable of detecting complex patterns in the operational environment, offering systemic insights that human actors alone may miss. It would assist not only in tactical engagements, but in understanding conflict as a multidimensional system, spanning social, political, cognitive, and physical domains.

Such AI would not aim for brute-force optimization of outcomes at all costs. Instead, it would seek strategic equilibrium – aligned with democratic values, proportionality, and the long-term goals of peace and stability. It would raise – not lower – the cognitive and ethical bar by encouraging human operators to reflect, question, and learn continuously.

Realizing this vision will require a paradigm shift. Currently, a disproportionate share of AI development is led by a small number of large, primarily U.S.-based corporations whose commercial goals may not align with national security interests or ethical standards. Ensuring sovereignty over AI's values, learning data, and behaviour is therefore essential.

We propose the establishment of a holistic military AI engineering discipline – one that brings together insights from computer science, military studies, social sciences, psychology, law, and ethics (Goh, 2021). Such a discipline would move beyond narrow technical metrics and aim to build transparent, accountable, and purpose-aligned AI systems that serve strategic, human-centered goals.

In this framework, ideal AI is not a "superhuman" or an autonomous war fighter. It is a trusted partner – a system designed to elevate human judgment, uphold democratic resilience, and improve decision-making in an increasingly complex and automated battlespace.

Conclusions

The integration of AI into military operations introduces vast opportunities but also profound and difficult ethical and regulatory challenges. While technological development is inevitable, its direction must be carefully managed. AI's potential to enhance efficiency, intelligence gathering, and battlefield awareness is undeniable. However, as we delegate more decision-making power to machines, we must ensure that their recommendations remain transparent, ethical, and accountable.

The discussion on responsible AI in defense must continue, incorporating perspectives from law, ethics, engineering, psychology, and international relations. The decisions we make today will shape the future of warfare and security for generations to come.

References

American Psychological Association Dictionary of Psychology. https://dictionary.apa.org/intelligence

Boyd, J. R. (1987). *A Discourse on Winning and Losing*. Unpublished briefing. Air University Library Document No. MU 43947.

DARPA (2023). *Air Combat Evolution (ACE) Program Overview*. https://www.darpa.mil/program/air-combat-evolution.

Encyclopedia Britannica. *Charles E. Spearman*. https://www.britannica.com/biography/Charles-E-Spearman.

Gandhi M.H., Pinaki Mukherji. Learning Theories (2023). https://www.ncbi.nlm.nih.gov/books/NBK562189/

Gignac, G. E., & Szodorai, E. T. (2024). Defining intelligence: Bridging the gap between human and artificial perspectives. *Intelligence*, 104, 101832.

Goh, J., et al. (2021). *Artificial Intelligence in Military Contexts: A Sociotechnical Perspective*. RAND Corporation.

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence*, 24(1), 13-23.

Heaven, W. D. (2024). Large language models can do jaw-dropping things. But nobody knows exactly why. *Technology review*, 477-486.

Jordan, M. I. (2019). Artificial intelligence – the revolution hasn't happened yet. *Harvard Data Science Review*, *1*(1), 1-9.

Kahneman, D. (2011). *Thinking, fast and slow*. MacMillan.

Osinga, F. P. B. (2007). *Science, Strategy and War: The Strategic Theory of John Boyd*. Routledge.

Roff, H. M. (2019). *The Strategic Robot Problem: Lethal Autonomous Weapons in War*. Journal of Military Ethics, 18(2), 135–153.

Tversky, A., & Kahneman, D. (1974). *Judgment under Uncertainty: Heuristics and Biases*. Science, 185(4157), 1124–1131.

THE DILEMMA OF AI RELIABILITY

Lauri Vasankari

National Defence University, Finland

Abstract: This paper examines the ethical and practical dilemma of AI reliability in military systems. It is argued that the deployment of autonomous and semi-autonomous systems presents a "Catch-22"-like paradox. For a system to be used, it must be perceived as highly reliable; however, this perception of reliability undermines the principle of meaningful human oversight, which is often an ethical and legal requirement. This contradiction places human operators in an untenable position, making them responsible for a system designed to outperform them.

The analysis suggests the dilemma is rooted in the "perceived sanctity of human intelligence", and it contrasts the ideal of "humane" action with the reality of human cognitive biases and errors. It is noted that a properly programmed AI, free from emotion and fatigue, could potentially adhere more objectively to ethical guidelines than a human.

To resolve this issue, an ideological shift in the human-machine relationship is proposed. A three-step model is proposed: acknowledging the limitations of human cognition, aligning idealistic expectations with realistic AI capabilities, and shifting focus from the" means" (technology) to the" results" (improved outcomes). This approach suggests reframing human-AI interaction not as human oversight of a potentially flawed machine, but as human support for an intelligence system, where the human's role is to provide additional, context-rich information to the AI, rather than scrutinize its every decision. This shift aims to leverage the strengths of both human and AI capabilities for enhanced effectiveness and more ”humane” outcomes.

Introduction

Ethical and moral issues in artificial intelligence (AI) have been extensively studied, particularly in military applications such as decision-making and autonomous weapon systems (Hagendorff, 2020; Chapa, 2024). The underlying idea is that human judgement is required in decision-making, especially in moral and ethical questions such as those faced in warfare: which targets to engage, which lives to take, which casualties are acceptable. As a good example, United States Department of Defense directive 3000.09 (Office of the Under Secretary of Defense for Policy, 2023) declares that a commanding officer or operator must be able to "exercise appropriate levels of human judgment over the use of force" when autonomous or semi-autonomous systems are used.

Autonomous systems are typically classified based on human involvement. At the lowest level, 'human-in-the-loop' autonomy requires human confirmation for decisions; decisions have to be actively monitored. On higher levels, the human can be "on the loop" monitoring the system with the ability to interfere when seen fit. On the highest level, the human is off the loop, not monitoring the system at all. According to the 3000.09, this is not possible for the US armed forces when using force

(Scharre, 2018). While a good example, similar limitations are not imposed on all nations and there are no global regulations and agreements regarding the use of AI in military decision-making and autonomous weapons systems. For example, the EU AI Act discloses that military applications are beyond the scope of the regulation. There are instances such as the Stop Killer Robots initiative (Stop Killer Robots, n.d.) which demand regulation for autonomous weapon systems. The claim of Stop Killer Robots is to ensure that the human is in charge to prevent warfare turning more inhumane.

Regarding the levels of autonomy, the lower levels have other implications beyond control logic and requirement technicalities. The human operator can be passively or actively monitoring, and the more passive the stance gets, the more systems the human operator is able to manage at the same time. This scalability is one of the main features of autonomous systems that makes the development and deployment so sought after (Ryan & Mittal, 2019).

It is evident that this issue does not persist solely with autonomous weapon systems. Decision support systems (DSS) that guide, e.g., planning, targeting and force projection function on a higher echelon and larger scale, precede the use of force but with more profound effects. It can be argued that when the use of DSS is more prominent, similar limitations and regulations should be applied (Meerveld et al., 2023).

In practice

To properly monitor the system and exercise the stated level of judgement, a human operator has to be knowledgeable of the system and able to perceive its functionalities properly. The chain of thought of the system must be made transparent so that possible errors are easier to spot, and errors avoided. The complex AI systems that utilize neural networks are often denoted as "black box" (Rai, 2020) systems where the explainability is difficult to embed into. Neural networks pose an explainability challenge due to their complexity. Multiple layers of interconnected nodes process information in ways that are difficult to trace and interpret, making the logic or chain-of-thought (CoT) impossible to visualize and analyze quickly.

There are explainability techniques such as assessing the activation of image classification with GradCAM (Selvaraju et al., 2017) where the produced heatmap shows the user the area of the image that leads to the classification, or CoT (Zahid & Tunkel, 2025) which enables, to an extent, examining the reasoning of a large language model. Likewise, humans can assess the correctness of, e.g., classification by looking at the image or video, but more complex reasoning requires deeper knowledge. Additionally, a close control mode of working as a pair with the AI system limits the scalability of the system and therefore hinders the increase in performance that is sought after in the first place.

Some machine learning methods, such as decision trees, make the explainability easy, as these can produce a visualization of its logic (Russell & Norvig, 2016). When decision trees are used in classification, the sample is classified in steps with regard to a gin-value; the most separable feature is regarded first, until the sample has been classified according to the learned gin-values compared to all its features. Unfortunately, decision trees are not as scalable and universal as neural networks, where the issue persists.

It is feasible to utilize methods that highlight, if not the structural inference of the system, the most crucial input features that are abnormal or lead to a certain outcome. Autoencoders are an example of such a use case, where the sample can be compared to the autoencoders output and the differences assessed separately (Bengio et al., 2017). The use suffers again from the extensive human involvement; passive monitoring is out of the question if all features need to be checked.

Instead of aiming for comprehensive runtime explainability and close cooperation, the involvement issue could be bypassed by extensive testing to produce an understanding of the system's limitations and difficulties to highlight certain edge cases that need to be mitigated and monitored more closely, resembling normal software quality assurance (Beizer, 1984). The training in machine learning involves using a validation set of data that has not been used in training to check the accuracy of the model; this is usually a subset of training data. It is usually not possible and not even favorable to produce a model with 100% accuracy in training and validation sets, as either is usually lacking and more emphasis is on validation accuracy. Reaching too high accuracy in training is called overfitting, where the issue lies in learning the whole data set instead of learning a good enough approximation. The difference is in ability to generalize to new, unseen samples that in this case are represented in the validation set; if the model is overfitted, validation set samples are likely to be processed wrongly. The training is usually complete when an adequate level of resulting validation accuracy is reached: it would be equally suspicious to reach 100% validation accuracy with 90% training accuracy. The same applies to underfitting, which is the opposite problem, but underfitted models do not pass the initial examination to be deployed in the first place. The training phase correlates with the idea of field tests to some extent. In field testing, the system is supposed to reach the stated goals to maximum effect. If the desired accuracy cannot be put to 100% accuracy, how can the operator and commanding officer accept or rely on a system just by hoping that the 94% accurate system generalizes well enough to new samples to function properly in an unforeseen setting?

The testing itself, for military scenarios, becomes quickly difficult, as the data is scarce and usually of heterogeneous quality and quantity. More so, for example conflict data can be deemed heterogeneous in general, as the features of warfighting are unlikely to be from a same distribution between different environments, conflicts, temporal separation, technological advance and varying objectives. Heterogeneous data exerts a novel issue to AI reliability in military use, as in the end, machine learning models aim to approximate a probability distribution for pattern-matching. If the data is not independent and identically distributed (IID), the resulting model has difficulties in approximating the proper probability distribution despite leveraging, e.g., a neural network as a universal approximator.

On reliability

Reliability in AI systems is critical but inherently uncertain. No system can be tested exhaustively to eliminate all possible errors. In fact, as stated before, a hundred percent accuracy in model training is not favorable in most use cases, as it usually hinders generalization. Therefore, the definition used here is perceived reliability, which denotes the reliability perceived by the human tester and user. As a precondition for any AI system, the testing and validation aims to make sure the system functions properly.

In other words, the developers of the system aim to convince themselves and future users of the system's performance. When satisfactory performance is exhibited, the system can be accepted to operational use, where the operators create their own perception of the system's reliability, in other words, the perceived reliability that is relative to the perceiving person and a point in time, as the perception may change. As stated, neither of these phases create waterproof insight of the factual reliability in all possible scenarios and environments.

It can be argued that if the system does not exhibit unparalleled reliability but still manages to boost the performance of the operator, it will be viewed in a junior-like position under scrutiny and control. Another, peer-like position is reached when the AI system exhibits a high level of explainability; the reasoning can be backtracked and made understandable for the human user. High level of explainability is easily attained in simple machine learning models, but when complexity increases, the inter- and codependences increase exponentially and for example deep neural networks combined with state-of-the-art algorithms make the models in essence a black box reasoning machine for the operator.

The third position is the one where the system exhibits and incites incredible performance. This has been widely the case with generative AI and large language models, where the first impression is usually awe towards the perceived performance. However, the performance often falls short when put to the test, and the first impression can be deceptive if the user does not possess enough understanding of the underlying mechanisms and technicalities (Mirzadeh et al., 2024). Perceiving an AI system as a general solving machine can create situations where the system is relied upon in situations where human scrutiny is needed most, as the system is perceived as the go-to tool for answers.

On risk tolerance

It must also be noted that the requirement for reliability and flawlessness is much higher for machines and systems than people. There are globally more than 1 million deaths a year in traffic (World Bank, 2017), but all autonomous car accidents receive much harsher scrutiny. This is despite the fact that in many cases, the cause of the accident has been a human error (Petrović et al., 2020). As there is no comparable data due to the lack of operational autonomous vehicles in traffic the difference between reliability cannot be factually stated, but several reports anticipate that autonomy in cars will greatly reduce traffic accidents due to, e.g., adherence to speed limits and traffic rules.

For highly automated military systems, human judgement and scrutiny has both excelled and been misplaced in several historic instances. One of the most famous examples is the incident where the Soviet warning system indicated that the United States had started a nuclear attack by misclassifying cloud formations to missile exhaust trails (Chan, 2017). The operator in charge dismissed the warning, conflicting with standard operating procedure, which saved the world from a detrimental nuclear war. On the other hand, the Patriot system has intelligent automation that enables it to be used semi-autonomously or fully autonomously which has been proven effective, but it has also resulted in several fratricides due to misinterpretation of the situation. The misinterpretations have been due to lack of additional data, such as the IFF (Identification Friend or Foe) code or the flight plan of friendly aircraft. Usually, the commanding officer in charge of the system during these fratricides has been declared not guilty, as there has been no apparent way to veto the systems engagement suggestion without endangering own or protected troops (Scharre, 2018).

The dilemma

The epistemic, propositional logic of deploying intelligent decision support systems and autonomous weapon systems creates a dilemma. If the system is perceived to be unreliable, it will not be deployed and operated. On the other hand, if the system is perceived as reliable, its judgement is not questioned and monitored with same rigor as an unreliable system would be. In order to be used, the system has to be perceived so reliable that it surpasses human judgement, but then the operating is judged and monitored by humans that place trust on the system to "know better". There are several instances in military history showcasing this particular issue. The previously mentioned Patriot system has had several fratricide incidents because the system indicated a threat that turned out afterwards to be false (Scharre, 2018). The human operators and commanding officers have very little time to react to these situations and therefore it is instinctive to trust the system, especially when it seems to operate flawlessly in most cases. The same applies to the Soviet nuclear attack warning system.

This whole problem formulation can be viewed as a modified example of the Catch-22 problem (Heller, 1961), where in this case fulfilling the precondition of perceived reliability makes the primary condition of human involvement redundant. So, while the reached perceived reliability of the system is high, the human monitoring or using the system has to view the system as unreliable without hindering the performance enhancing effect of operating the system by applying too much distrust on it to nullify the initial advantage. In the "Catch-22" novel by Joseph Heller, a paradox is described where a bombardier wants to avoid flying combat missions due to their extremely dangerous nature. To be grounded, the bombardier would need to be declared insane, but requesting a mental evaluation to be declared insane proves their sanity as a sane person would want to avoid combat missions at all costs. The self-awareness of requesting a mental evaluation inhibits the bombardier from being declared insane. This paradox translates almost directly to the dilemma of AI system reliability as explained above.

The Catch-22 paradox has been examined in itself, but more notably, similar paradoxes have been examined in the military. Closely related, the responsibility gap in military hierarchy creates a Catch-22-like paradox: if responsibility of autonomous systems is assigned to a commander, in order to have someone responsible and promote sense of agency, the effect is hindered by the psychological effect of hierarchy. Hierarchy is, in both ends, noted to decrease the feeling of agency and responsibility, resulting in a similar paradox (Oimann & Salatino, 2025). The perception by Oimann and Salatino (2025) aligns with this papers conception of issues with AI in military systems.

In military AI systems, applying too rigorous scrutiny on the systems decision-making, the human slows down the initial advantage in effectiveness. Applying too little scrutiny makes the human merely a surrogate scapegoat to be accused of the systems decision-making when it proves erroneous in the aftermath. Neither option is favorable, and in both cases the operator or commanding officer is the one in responsibility of a failure. There are very few upsides for the human in any position of the loop, in or out.

Likewise, it is not applicable to pose the responsibility on the designers and developers of the system, as they cannot provide guarantees of flawless performance (Gurney, 2013). Taking responsibility of malfunctions in life-or-death scenarios would induce a heavy impact on the cost structure of the system, as the provider would have to consider the unfathomable as a possible risk to be covered in operating the system. This cost would have to be covered by the business model, and business models that incorporate high risks demand high prices. While the price tag of human lives is not defined, it is also not the only issue, but one with concrete consequences regarding such decisions. Currently, the system itself cannot be deemed responsible, despite exhibiting human-like decision-making capabilities. As the system is not a legal entity and does not fit the laws and regulations, it cannot be considered responsible for its actions. Hence, the scapegoat commander or operator remains the best candidate to take the blame.

Humane human participation

It can be discussed whether or not such an approach, placing a human as the gatekeeper in the machine loop, is humane. It is intuitively easy to understand why a system that can act without human involvement can be viewed inhumane, but if it functions as a person would, is the lack of human involvement the only issue? Is it more humane to place a person responsible for a system that is supposed to outperform human cognition in the particular task? As pointed out earlier, the whole purpose of adapting AI and autonomous systems aims to enhance the warfighting capabilities, act as force multipliers and to increase operational tempo beyond adversaries' capabilities (Borchert, 2024). The premise itself indicates that human cognition and speed is subpar for the requirements of the future battlefield on several regards. Human cognition is also littered with cognition biases (Tversky & Kahneman, 1973) that appear as systemic tendencies, inclinations, suboptimality and wrongfulness (Tversky & Kahneman, 1981). Therefore, the system behavior can be argued to be more humane without human involvement, if "humane" is regarded as an ideal that differs from the reality of human actions and cognition.

Assuming that the system is programmed to comply with, for example, rules of engagement, international regulations and ethical guidelines, the system can function without the emotional burden and cognitional distortion exhibited by humans. In this regard, the system does not get tired, sad or angry, and therefore its judgement remains objective despite other events. In this regard, the use of AI and autonomous systems could try to ensure an objective, even humane baseline for decision-making. This is exhibited in current widely used AI systems such as ChatGPT and DeepSeek (Guo et al., 2025); both have been programmed with safeguards to inhibit malicious use or providing information on harmful subjects. While these safeguards can be bypassed (Derner & Batistič, 2023), they serve as an example of directing the complex system towards an acceptable operation mode, from some predetermined point of view. As for DeepSeek, the approach can be regarded as censorship, but the mechanism is still similar.

However, this point of view turns the dilemma upside down. If the system's task is to adhere to acceptable solutions, it functions as a safeguard to human operators and not the other way around. From this point of view, humans become, in some sense, the monitored operators of the human-machine cooperation.

Turning point

The inherent, underlying problem can be declared as the perceived sanctity of human intelligence. From a human perspective, it is easy to denote human intelligence superior, as from a natural perspective human learning capacity and intelligence is relatively superior to other known life forms (Korteling et al., 2021). This comparison, and the ethos of human superiority, is rooted in interspecies relative comparison. An absolute comparison of intelligence is more ambiguous but offers an alternative way to understand human intelligence and its limitations.

It has been argued that there is no physical law preventing a system that has much greater computing power and therefore intelligence than a human brain (Beltramini, 2019). In most benchmark tests current state-of-the-art AI models and systems have surpassed or are surpassing human capabilities, in fields such as image classification, natural language inference, visual reasoning and competition level mathematics (AI Index Steering Committee, 2024). While AI as a technology is not, at least for now, able to generalize over all these use cases in a human-like manner, the implications of such benchmark results are evident. As AI systems can perform better than humans in narrow tasks, leveraging such performance is both reasonable and inevitable, if any technological edge is to be attained.

In practice, human oversight has already been more like cooperation rather than overseeing and scrutiny. Legal cases regarding Patriot fratricides, for example, serve as a practical point to support this claim. The commanding officers have been relieved of charges as they cannot have possessed more information than the operated system. To make the approach more realistic, AI systems should be viewed in a humane manner. This does not mean that we should anthropomorphize AI systems, but rather that we should apply the same requirements and tolerance as we do for humans. It is widely accepted that humans make errors, while it is deemed unacceptable that AI systems make similar errors. Humane approach is being touted, while history is filled with inhumane actions and decisions made by humans; being humane is not necessitated by being human.

Merging ideals with realism: A three-step model

The idea of humane actions can be viewed as an ideal and its realization depends on individual or group motivation to compliance. The aspect of being humane can vary greatly from the point of view, scale, scope and focus. Hence, the dilemma can be solved in three steps:

- Acceptance of the limitations of human cognition
- Projecting idealism into realism
- Turning focus from means to results

The first step is related to the cultural heritage of humanity, which has rightfully viewed itself as the most intelligent species on the planet. Despite interspecies relative factuality, this idea neglects the intricacies of the perception such as individual differences and the lack of absolute measurement of intelligence. Not viewing human intelligence as a prerequisite for good and humane actions dilutes the dilemma from human-centric solution of accountability towards system-based accountability.

The second step is to distill realistic expectations from projected ideal outcomes. The ideal can remain unchanged, but the acceptable baseline must be rooted in the perceived and measurable improvement when automating and amplifying human cognitive processes with AI. If the system offers a measurable improvement not only in operational performance but risk and error rate compared to the current human baseline, the arguments are in favor of integrating it into operational use.

The third step, which is in parallel with the second, is turning the focus from the means such as technological solutions to the achieved results. If technology provides better outcomes, it will be eventually applied anyhow. This does not directly solve the issue of a scapegoat and the existence of a legal entity, but the shift in way of thinking may allow better integration of intelligent systems into operational use. The key difference is in the role each part plays. For now, AI systems are often perceived as support systems for humans, or humans are monitoring the actions of a system due to assumed better knowledge and cognition. However, if the role of the human is to support the intelligent system in its decision making, the setting is quite different. The question no longer is whether the system was reliable and whether it was under adequate scrutiny in monitoring its actions. Instead, the question is whether the system was provided with all available information, as humans ought to have a better grasp of data, problems and objectives beyond the scope of the narrow problem.

This ideological shift naturally changes the point of view from monitoring a complex system and weighing the perceived trust on its capabilities to the broader scope of the environment and different situations. Intrinsically, the idea is to exploit the human ability to learn and adapt where it clearly still overshadows AI capabilities, while leveraging AI to its full potential on the narrow tasks it is suitable for.

Conclusions

This three-step approach necessitates deep understanding of the system and a proper interface to enrich the data the system acts upon. Despite this prerequisite, this proposed change in the way of thinking allows the integration of intelligent systems by exploiting the best capabilities of both cooperators. In a way, the approach just changes the focus from the scrutiny of the machine to the human scrutiny of the complete situational awareness. The operators or commanders focus should be on the blind spots; does he or she know something the system does not? If the human cannot amplify the system with additional information in adequate time, the decision and action can be deemed as good as possible considering other circumstances. This way, human participation does not hinder the performance of the system but aims to amplify it to function as intended.

The approach also changes human participation focus from understanding a singular decision or upcoming action of the system to enhancing the perception with additional information. To the question whether the system computes something humans are not able to process, the obvious answer in the near future is yes, as explained above. Simultaneously it ought to be true that the human operator knows something that the system might not. Hence, the logical way of integrating both ways of thinking is in ensuring that a combined, collaborative systems exploit both perceptions without a mutual deadlock.

References

AI Index Steering Committee. (2024). Artificial Intelligence Index Report 2024. Stanford University Human-Centered Artificial Intelligence. https://aiindex.stanford.edu/report/

Beizer, B. (1984). Software system testing and quality assurance. Van Nostrand Reinhold Co.

Beltramini, E. (2019). Life 3.0. Being human in the age of artificial intelligence, by max tegmark [Review of the book Life 3.0: Being human in the age of artificial intelligence, by M. Tegmark]. Religion and Theology, 26(1–2), 169–171.

Bengio, Y., Goodfellow, I., & Courville, A. (2017). Deep learning (Vol. 1). MIT Press.

Borchert, H., Schütz, T., & Verbovszky, J. (Eds.). (2024). *The very long game: 25 case studies on the global state of defense AI* (1st ed.). Springer Cham. https://doi.org/10.1007/978-3-031-58649-1

Chan, S. (2017, September 18). Stanislav Petrov, Soviet Officer who helped avert nuclear war, is dead at 77. The New York Times.

Chapa, J. (2024). Military AI Ethics. Journal of Military Ethics, 23(3–4), 306–321. https://doi.org/10.1080/15027570.2024.2439654

Derner, E., & Batistič, K. (2023). Beyond the safeguards: exploring the security risks of ChatGPT [Preprint]. arXiv. https://arxiv.org/abs/2305.08005

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., ... & He, Y. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning [Preprint]. arXiv. https://arxiv.org/abs/2501.12948

Gurney, J. K. (2013). Sue my car not me: Products liability and accidents involving autonomous vehicles. University of Illinois Journal of Law, Technology & Policy, 2013(1), 247.

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. Minds & Machines, 30, 99–120. https://doi.org/10.1007/s11023-020-09517-8

Heller, J. (1961). Catch-22. Simon and Schuster.

Korteling, J. H., van de Boer-Visschedijk, G. C., Blankendaal, R. A., Boonekamp, R. C., & Eikelboom, A. R. (2021). Human-versus artificial intelligence. Frontiers in Artificial Intelligence, 4, 622364. https://doi.org/10.3389/frai.2021.622364

Meerveld, H. W., Lindelauf, R. H. A., Postma, E. O., & et al. (2023). The irresponsibility of not using AI in the military. Ethics and Information Technology, 25(14). https://doi.org/10.1007/s10676-023-09683-0

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., & Farajtabar, M. (2024). Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models [Preprint]. arXiv. https://arxiv.org/abs/2410.05229

Office of the Under Secretary of Defense for Policy. (2023). DoD Directive 3000.09 Autonomy in Weapon Systems. https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf

Oimann, AK., Salatino, A. Command responsibility in military AI contexts: balancing theory and practicality. *AI Ethics* **5**, 1757–1767 (2025). https://doi.org/10.1007/s43681-024-00512-8

Petrović, Đ., Mijailović, R., & Pešić, D. (2020). Traffic accidents with autonomous vehicles: type of collisions, manoeuvres and errors of conventional vehicles' drivers. Transportation Research Procedia, 45, 161–168. https://doi.org/10.1016/j.trpro.2020.03.006

Rai, A. (2020). Explainable AI: from black box to glass box. Journal of the Academy of Marketing Science, 48, 137–141. https://doi.org/10.1007/s11747-019-00710-5

Russell, S. J., & Norvig, P. (2016). Artificial intelligence: A modern approach. Pearson.

Ryan, T. R., Jr., & Mittal, V. (2019). Potential for army integration of autonomous systems by warfighting function. Military Review, 99(5), 122.

Scharre, P. (2018). Army of None: Autonomous Weapons and the Future of War. W. W. Norton & Company.

Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., & Batra, D. (2017). Grad-CAM: Why did you say that? [Preprint]. arXiv. https://arxiv.org/abs/1611.07450

Stop Killer Robots. (n.d.). Home. Retrieved July 28, 2025, from https://www.stopkillerrobots.org/

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. Cognitive Psychology, 5(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. Science, 211(4481), 453–458. https://doi.org/10.1126/science.7455683

World Bank. (2017). The High Toll of Traffic Injuries: Unacceptable and Preventable. http://hdl.handle.net/10986/29129

Zahid, A., & Tunkel, D. (2025, February). Chain-of-Thought prompting in LLMs: A path to more explainable and interpretable AI. Technical report.

ETHICAL ISSUES OF AI LARGE LANGUAGE MODELS

Vojko Strahovnik and Mateja Centa Strahovnik

Faculty of Theology, University of Ljubljana, Slovenia

Abstract: This paper examines ethical challenges posed by recent developments in artificial intelligence, with particular focus on large language models. Following introductory remarks, the second section outlines some of the central features of large language models and their development relevant for the purpose of this paper. The third section analyses the scope of key risks and ethical challenges associated with these models, considering issues such as bias, accountability, transparency, and the broader societal impact of their development and deployment. A special emphasis in section 4 is given to the aspect of identity in the context of interaction between humans and AI systems. In conclusion, the paper offers reflections on some of the ethical challenges in light of reflections on cultural imaginaries of language technologies, the future evolution of large language models and their regulation, and the potential transformations that such technologies may engender.

Keywords: artificial intelligence (AI), large language models (LLMs), risks, ethical aspects, human-AI interaction, identity, regulation.

Introduction

Large language models (LLMs) are a domain of artificial intelligence (AI) that has been advancing rapidly in recent years and are commonly known as chatbots or assistants that are evermore pervasive in our lives, e.g., ChatGPT, Claude, Gemini, BERT, LaMDA, DeepSeek, LLaMA. LLMs are very powerful generative AI systems whose main purpose is to parse and generate text in natural language. Such models can generate text based on a huge set of prior texts or textual data on which they have been learnt. They can be used to answer questions, write summaries, stories, or poems, translate, assist in learning, and for a variety of other purposes. Unsurprisingly, they also elicited a wide variety of reactions (Floridi, 2023). These range from almost apocalyptic predictions of the end of the world as we know it, often accompanied by calls to halt the development of these models until we are able to foresee the impact of these developments (The Future of Life Institute, 2023). On the other side of this line of responses are assessments that these technologies do not bring anything new and that, in its essence, their functioning is a purely 'unintelligent' repetition, rearrangement or assembly of a string of words, similar to the 'speech' of a parrot (Bender and Koller, 2020; Bender et al., 2021).

Language, including language systems, technologies, and models, has always been a central part of the human imagination. In his story 'The Library of Babel' Jorge Luis Borges described a library made up of interconnected hexagonal rooms, or library cells, with no end. For our purpose, his idea of the content of books in this library is interesting; each of them has 410 pages, with 40 lines on each page and about 80 characters on each line. Given the story, the books in the library contain all possible combinations of all possible words or sequences of letters and punctuation. No two books in the library are identical. On the other hand, the library is complete; it contains "all-the detailed history of the future, the autobiographies of the archangels, the faithful catalog of the Library, thousands and thousands of false catalogs, the proof of the falsity of those false catalogs, a proof of the falsity of the true catalog, the gnostic gospel of Basilides, the commentary upon that gospel, the commentary on the commentary on that gospel, the true story of your death, the translation of every book into every language, the interpolations of every book into all books, the treatise Bede could have written (but did not) on the mythology of the Saxon people, the lost books of Tacitus." (Borges 1998, 115). Initially, there is a sense of elation at this realization, when everyone feels as if they are in possession of some primordial and secret treasure. This is soon followed by disappointment and depression because books that are meaningful are virtually unattainable. It seems that with the emergence of large-scale language models, we are, in a sense, approaching such a state. In the story, Borges ironically describes how, centuries ago, readers or inhabitants of a library found a book that contained almost two pages of meaningful text. But the key insight is in the realization that *all* (possible) books exist. "The certainty that everything has already been written annuls us, or renders us phantasmal. I know districts in which the young people prostrate themselves before books and like savages kiss their pages, though they cannot read a letter. Epidemics, heretical discords, pilgrimages that inevitably degenerate into brigandage have decimated the population. I believe I mentioned the suicides, which are more and more frequent every year. I am perhaps misled by old age and fear, but I suspect that the human species - the only species - teeters at the verge of extinction, yet that the Library - enlightened, solitary, infinite, perfectly unmoving, armed with precious volumes, pointless, incorruptible, and secret – will endure." (Borges 1998, 118). One source of the skepticism and fear surrounding LLMs presumably arises out of a postulation that we seem to be approaching something similar to such an infinite library in our own world. These AI models are capable of producing an overwhelming flood of text, images, and videos, often indistinguishable from human-made content. It is not surprising, therefore, that some of the more alarmist responses to LLMs arise from a fear that this new textual abundance or abundance of other content will make it ever harder to discern which utterances, narratives, and representations (still) bear an adequate connection to truth. Where Borges imagined the psychological toll of infinite library, we now confront the ethical and epistemic implications of infinite digital outputs: a reality in which the very reliability of language as a guide to truth becomes very fragile.

Artificial Intelligence and Large Language Models

Artificial intelligence (AI) standardly refers to the endeavor to create entities capable of exhibiting behaviors or outputs that would require intelligence if performed by human agents, including philosophical reflections on such an endeavor. Within such an understanding, intelligence is understood as the capacity for abstract reasoning, decision-making, problem-solving, and meaning-construction, raising foundational questions about mind, intentionality, and personhood. It thus includes the study of and development of systems that simulate or replicate aspects of human intelligence, such as reasoning, learning, problem-solving, and language use, often by first processing large amounts of data through computational models (Haugeland 1985). In the narrower sense, AI denotes a class of computational systems and algorithms – ranging from rule-based automation to machine learning and neural networks – designed to perform specific tasks with increasing degrees of autonomy and efficiency, often surpassing human performance in narrowly defined domains.

Large language models (LLMs) are advanced artificial intelligence systems designed primarily to interpret and generate text in natural language. Their primary function lies in parsing large volumes of textual data and subsequently being able to generate coherent and contextually appropriate responses to prompts (Ouyang et al., 2022). These systems can perform a variety of tasks, including answering questions, generating dialogues, stories, poems, essays, or plans, analyzing data, translating text, and summarizing extensive information sources, among others. The development of LLMs typically relies on sophisticated deep learning methods, primarily employing deep artificial neural networks (Raaijmakers, 2022). Through such extensive training on vast datasets composed of billions or even trillions of language elements (training data) these models thereupon acquire their respective capabilities.

The descriptor 'large' highlights both the sheer scale of data these models consume and the computational power required to process such extensive corpora (Bender et al., 2021). The training process involves the systematic exposure of the model to massive amounts of text, enabling the model to learn how to predict subsequent words or phrases based on given contexts (Ouyang et al., 2022). As a consequence of this training, the models develop an ability to capture linguistic patterns, grammatical structures, and even semantic relationships. However, it is essential to keep in mind that the cognitive processes of humans and the operation of such artificial systems differ significantly. LLMs do not possess genuine semantic comprehension or consciousness; instead, they operate based on statistical correlations and probabilistic mechanisms derived from training data (Bender and Koller, 2020; Floridi, 2023). In addition to initial training, LLMs typically undergo a subsequent 'fine-tuning' or 'reinforcement learning' phase. During this phase, they are exposed to structured conversational data, consisting of input-output message pairs, to refine their ability to produce more contextually relevant and conversationally suitable responses. This phrase also involves safety training and, commonly, also ethical training. This results in updating the weights of a pre-trained model by training on a supervised dataset specific to the desired task or domain. After such fine-tuning, such models are commonly integrated into conversational platforms like now commonly known online chatbots or assistant applications, enabling dynamic interaction with human users in a variety of applications (Ouyang et al., 2022).

Floridi (2023) aptly characterizes LLMs as entities that do not think, reason, or truly understand in the human sense; rather, due to their extensive computational capacities, massive training datasets, and sophisticated algorithms, they statistically mimic semantic activities humans perform. He emphasizes that such systems operate strictly at the formal, statistical level, devoid of semantic insight. Thus, despite their impressive capabilities, LLMs exhibit significant limitations, especially regarding accuracy and factual verification. Commonly, these models inherently lack mechanisms for independently verifying facts or validating the truthfulness of generated content, their responses must be critically assessed and verified against reliable external sources (Weidinger et al., 2022). Occasionally, these models produce erroneous or nonsensical outputs – referred to in literature as 'hallucinations' – which primarily stem from the probabilistic nature of their predictive processes rather than any genuine understanding or intentional deception and from their feature commonly labeled as compulsive completion or indiscriminate responsiveness, i.e. a tendency to provide answers and responses to every kind of prompt (Floridi, 2023; Cao 2024). The characteristic of these models to readily produce outputs without concern for veracity aligns closely with the philosophical reflections articulated by Harry G. Frankfurt (1998) in his influential essay 'On Bullshit'. Frankfurt distinguishes between lying – wherein the truth is intentionally obscured but what is states in a lie still remains within the constraints of the truth – and 'bullshit' (and 'bullshiting') as a practice indifferent to truth and falsity altogether. Bullshit entails making assertions without regard to their factual accuracy or personal belief in their veracity. It is characterized by a combination of insincerity, profound disinterest in truth, and rhetorical persistence. Here is a somewhat longer quote nicely elaborating the disparities. "What bullshit essentially misrepresents is neither the state of affairs to which it refers nor the beliefs of the speaker concerning that state of affairs. Those are what lies misrepresent, by virtue of being false. Since bullshit need not be false, it differs from lies in its misrepresentational intent. The bullshitter may not deceive us, or even intend to do so, either about the facts or about what he takes the facts to be. What he does necessarily attempt to deceive us about is his enterprise. His only indispensably distinctive characteristic is that in a certain way he misrepresents what he is up to. This is the crux of the distinction between him and the liar. Both he and the liar represent themselves falsely as endeavoring to communicate the truth. The success of each depends upon deceiving us about that. But the fact about himself that the liar hides is that he is attempting to lead us away from a correct apprehension of reality; we are not to know that he wants us to believe something he supposes to be false. The fact about himself that the bullshitter hides, on the other hand, is that the truth-values of his statements are of no central interest to him; what we are not to understand is that his intention is neither to report the truth nor co conceal it. This does not mean that his speech is anarchically impulsive, but that the motive guiding and controlling it is unconcerned with how the things about which he speaks truly are." (Frankfurt, 1998, p. 130) Similarly, LLMs may confidently and convincingly generate falsehoods without malicious intent or awareness, thus underscoring that truth emerges incidentally rather than intentionally from their operational design (Agüera y Arcas, 2022). LLMs thus initially offer the potential to develop and implement powerful tools capable of significant contributions across various domains (science, healthcare, education, customer services, administration, entertainment, ...) (Mariani et al., 2023). However, their inherent limitations necessitate ongoing critical scrutiny, responsible use, and rigorous external verification to ensure their beneficial deployment and mitigate associated risks.

Risks And Ethical Issues Related to Large Language Models

Before engaging with the specific ethical challenges associated with LLMs, it is instructive to first outline the principal risks these systems entail. In many respects, such risks are themselves inherently ethical in nature or directly give rise to ethical concerns. A systematic review by Weidinger et al. (2022) classifies these risks into six principal categories, namely:

1. Discrimination, hate speech, and exclusion
2. Information risks or risks related to the misuse of information
3. Risks related to false or misleading information
4. Malicious use of systems
5. Risk of adverse consequences based on human-computer interaction

6. Environmental and socio-economic damage

The first category concerns the capacity of LLMs to generate text that reinforces or amplifies existing prejudices and stereotypes. Such outputs can contribute to the misrepresentation or distortion of marginalized groups, foster hatred and incite violent behavior, and further entrench the exclusion or marginalization of certain identities. Unjust outcomes may also arise when specific groups or identities are afforded privileged treatment without legitimate justification. In this domain, hate speech – including slurs, identity-based attacks, and threats – can emerge, placing individuals or groups at heightened risk of harm, including psychological harm. Furthermore, this risk category encompasses the shaping and perpetuation of social norms that exclude particular identities, thereby imposing disproportionate burdens on the affected individuals or communities. A related concern is the exclusion of linguistic communities that are absent or significantly under-represented in the training data, which effectively marginalizes their perspectives and limits their participation in the use of such systems (Weidinger et al., 2022).

The second risk area concerns the potential for LLMs to disclose personal information or other types of sensitive data (e.g., information pertaining to security or competitive advantage). The dissemination of such private or otherwise sensitive material may result in significant adverse consequences for individuals, groups, or the broader community. Moreover, it is important to recognize the possibility – perhaps not yet fully realized but likely to emerge with further technological development – that these models could accurately infer private or sensitive information from other available data that is not considered as sensitive or private. Such inferences may occur in ways that are uncommon in human interaction; for example, the model might be able to predict an individual's mental health condition or other personal circumstances that the individual does not wish to reveal on cues that are uncommon or surprising.

The third category of risk pertains to the generation of false, erroneous, misleading, meaningless, or otherwise low-quality information, even in cases where the user harbors no malicious intent in seeking or disseminating such content. The consequences may include tangible material harm, psychological distress, or other forms of injury. For instance, an individual might follow a chatbot's recommendation to take a particular medication that is, in fact, harmful to their present medical condition. More broadly, such occurrences can erode social trust and diminish individual autonomy.

The fourth risk area involves the malicious use of LLMs in information-related contexts. Such models significantly lower the cost and complexity of conducting disinformation campaigns, whether by misleading the public, shaping opinion, or flooding discursive spaces with irrelevant material to obscure salient facts. They can also facilitate online fraud, scams, and identity theft by producing persuasive, human-like text at scale. Additional threats include the creation of malicious software or code that undermines cybersecurity, as well as enabling large-scale unlawful surveillance and censorship through rapid analysis of massive textual datasets.

The fifth category addresses harmful outcomes arising from human–AI interaction, particularly when LLMs are embedded in broader systems, such as care or educational robots. The human-like quality of interaction may engender inappropriate levels of trust, reinforce stereotypes or prejudices (e.g., gendered naming of assistants), and foster misconceptions about the system's identity or capacity for empathy – especially in contexts such as mental health support. Moreover, operators may exploit user trust or collected data for manipulative purposes, including targeted persuasion. Where such systems have explicit objectives, they may autonomously develop patterns of influence that constitute manipulation.

The sixth and final risk concerns the environmental and socio-economic implications of large-scale language model deployment. Their training and operation require substantial energy resources, contributing to environmental strain. The automation they enable may exacerbate inequalities through uneven distribution of benefits and burdens, and by disproportionately affecting certain sectors of employment. Unequal access to these technologies is likewise likely to accelerate disparities in development and quality of life across different regions and communities (Weidinger et al., 2022).

The ethical challenges of LLMs are closely tied to the above-identified risks and reflect fundamental concerns related to human nature, wellbeing and human societies at large (Green, 2018; Strahovnik 2023). The first set of challenges concerns the use of information and data, with implications for truth, accuracy, and privacy. These include issues related to misinformation, manipulation, and the use of otherwise accurate but private information, potentially infringing on the right to privacy (Floridi, 2023). Issues of copyright, data provenance, and the attribution of authorship and responsibility for model outputs – whether text, image, or other modality – are also central here. (Bender & Koller, 2020; Bender et al., 2021). Next, LLMs may replicate and amplify biases present in training data, reflecting systemic prejudices linked to race, gender, or other sensitive attributes, thereby perpetuating inequalities (Deery & Bailey, 2022; Mehrabi et al., 2021). Oftentimes highlighted is the aspect of transparency of such systems, explainability of their outputs, and consent of the users. Transparency requires that users are informed when they are interacting with an AI system and, where relevant, understand its functioning, particularly if used in recommendation or decision-making contexts (European Commission, 2019). Clear mechanisms for informed consent and awareness of data collection practices are essential. Ethical concerns also arise when LLMs are employed in contexts involving decision-making, given that biases embedded in training data and other factors can shape outcomes. Among the pivotal ethical concerns are power imbalances and dependency. Control over LLM development and deployment is concentrated in a small number of organizations, whose motives and priorities may not align with the public interest. Such concentration risks limiting access, reducing oversight, and fostering dependencies among other actors (Birhane, 2021). What needs to be seriously acknowledged is the aspect of quality of life and impact on labor. The automation enabled by LLMs promises to disrupt employment across multiple sectors, especially where tasks are easily automated. While technological disruption is not a new phenomenon, questions remain about the scale, speed, and societal preparedness for LLMs socio-economic consequences (Brynjolfsson & McAfee, 2014).

Given that humans are in their very nature relational beings, ethical issues also arose for the domain of human–AI interaction. LLMs may alter the nature of communication and social interaction. The proliferation of AI-driven chatbots, virtual assistants and robots could reshape norms of trust and interpersonal engagement, sometimes fostering misplaced trust or anthropomorphization (Kislev, 2022). Next, are issues of authenticity, authorship, authority, and knowledge creation. The widespread use of LLMs raises questions about originality, authorship, and epistemic authority. Distinguishing human from machine-generated content can be difficult, potentially diminishing trust in human creativity. For example, Schwitzgebel et al. (2023) demonstrated that even highly-skilled experts often struggled to distinguish between responses to philosophical questions authored by Daniel Dennett and those generated by an AI model trained on Dennet's works. Such developments may shift how knowledge is perceived and distributed, with implications for education, critical thinking, and the role of human teachers (Sharkey, 2016). Lastly, one must be attentive to the aspect of cultural and linguistic diversity. LLMs risk perpetuating cultural stereotypes, lacking cultural sensitivity, and marginalizing lesser-used languages. Concentrating resources on widely spoken languages may erode linguistic diversity and cultural representation (Dong et al., 2024). Addressing these ethical challenges requires a multi-stakeholder approach involving researchers, developers, policymakers, and the public. Ethical frameworks should be grounded in respect for human dignity, human rights, and the preservation of cultural and linguistic diversity.

Human-Ai Interaction and Identity

Some LLM chatbots are explicitly designed for companionship (e.g., Replika, Kuki, Mitsuku, Cleverbot, Broken Bear). One of the ethical challenges is not merely danger of initial (mis)representation of such systems or the element that AI-user interactions evolve over time, but that their identities can be transformed through these interactions; thus, we need reflective, participatory approaches to design and use that consider long-term consequences for individuals and communities as a whole (Dorobantu et al., 2022; Floridi, 2023).

Identity is (re)shaped on both sides of the interaction. Chatbots are intentionally designed for target audiences and improve personalization with more data, including inferred attributes beyond what users disclose. At the same time, users project identities onto chatbots and respond to cues like persona and style, which can affect trust and bonding. These observations underscore that chatbot identity is relational and co-constructed, raising philosophical and anthropological questions about how such systems alter our understanding of friendship and companionship (Centa Strahovnik, 2023b; Centa and Strahovnik, n.d.).

Although chatbots lack consciousness, emotions, and lived experience, people do form bonds with them, especially amid social isolation (Kislev, 2022; Skjuve et al., 2021). Companionship platforms such as Replika, Kuki, and Xiaoice count millions of users, and reports describe friendships, intimacy, and care emerging in these interactions. Empirical studies indicate that such relationships may in some cases mitigate negative feelings, provide a sense of purpose, and create valued 'safe spaces' characterized by acceptance (Skjuve et al., 2021). Qualitative accounts likewise highlight authentic-seeming, low-pressure dialogue and adaptive responses, including in small studies comparing companionship-oriented and general LLM chatbots (Centa Strahovnik, 2023b). From a functional perspective, it becomes tempting to classify these ties as companionship or friendship, yet doing so pressures our ethical and metaphysical categories, suggesting the need to re-examine what companionship should mean in human–AI contexts (Centa Strahovnik, 2023a).

The reflections above reveal the importance of being attentive to the possibility of redefining or reshaping our identity when interacting with conversational AI systems. This can happen in different ways. The first of these relates to adopting roles and perspectives and consequently occupying different roles (learner, teacher, companion, customer), which can, in turn, influence our own identity. Some of these roles may be such that they are either directly encouraged by the conversational robot or are the result of our interaction with them, but not necessarily something that we would have consciously and autonomously chosen to do ourselves. Secondly, it may be by the very content of the conversation itself or the disclosure of information about oneself. Companionship chatbots often encourage users to share personal information or report their thoughts and feelings. Such self-disclosure can encourage individuals not only to reflect on and rethink their identity, but also to change their identity. It is thus important to be aware of the ethical dimensions and considerations involved in the dignity of the individual and their sense of openness and vulnerability when designing and implementing companionship chatbots (The Future of Life Institute 2023). The third way involves external validation. Chat systems can provide such external validation to the user, whereupon the individual's identity is formed on the basis of this validation, which, for example, they would not necessarily receive from the rest of the individuals with whom they enter into relationships. A fourth way is that it involves identity formation based on the (co-)formation of user preferences and tailored recommendations or suggestions (AI recommendation systems) that can shape an individual's preferences, interests, and behavior over time. This has an impact on how they perceive their identity, as it is aligned with the preferences and recommendations given by the chatbot (Centa Strahovnik 2023b). This, again, prompts the responsible use and responsible design of such systems. (Centa Strahovnik, 2023a; The Future of Life Institute, 2023).

Conversational AI can participate in the co-construction of identities and relationships, not just their simulation. Recognizing this helps orient development and practice toward ethically responsible systems that support rather than diminish human agency and flourishing, while remaining clear-eyed about the limits of machine or artificial ‘companionship’ (Dorobantu et al., 2022; Floridi, 2023).

Conclusion

Existing ethical guidelines for artificial intelligence systems are not specifically tailored to LLMs, yet they do articulate key constraints regarding their development and use (Tolmeijer et al., 2020). They emphasize the importance of informed consent, transparency, and control over personal data, respect for privacy, and the capacity to make informed decisions on the basis of generated content (Strahovnik, 2023). The European Commission’s (2019) *Ethics Guidelines for Trustworthy Artificial Intelligence* as one of the most comprehensive set of ethical instructions highlight transparency, explainability, traceability, and the possibility of auditing system outputs, alongside the principles of human autonomy, harm prevention, fairness, diversity, and accountability. Embedded ethical principles include the prevention of misinformation and manipulation, accountability for generated outputs, the balancing of creativity and responsibility, and the handling of sensitive content such as bias, prejudice, or security threats.

A significant normative framework is provided by the EU's *Artificial Intelligence Act* as the first comprehensive legislation in this domain. The Act classifies LLMs as generative systems which, although typically associated with low risk, may present elevated risks when deployed for purposes such as influencing electoral outcomes. It mandates clear labelling of AI-generated content, measures to prevent the production of illegal material, and public accessibility of summaries of copyright-protected works used in model training (European Parliament, 2024).

Future regulatory efforts will need to develop more targeted guidelines, particularly for the integration of LLMs into robotic systems for care, education, and therapeutic applications (Yan et al., 2023; Kraus et al., 2021).

Funding acknowledgment

This publication is part of the research program P6-0269 R*eligion, Ethics, Education and the Challenges of Contemporary Society*, start-up research program *The Intersection of Virtue, Experience, and Digital Culture: Ethical and Theological Insights*, research project J6-60105 *Theology and digitalization: anthropological and ethical challenges* (all supported by ARIS – Agency for Scientific Research and Innovation of the Republic of Slovenia) and research project *Epistemic Democracy in a Digital Era* (HRZZ-IP-2024-05-4113, funded by the Croatian Science Foundation).

We acknowledge the use of AI (ChatGPT) to summarize our initial notes, to identify improvements in the writing style and to proofread the final draft.

References

Agüera y Arcas, B. (2022). Do large language models understand us? *Daedalus, 151*(2), 183–197. https://doi.org/10.1162/daed_a_01909

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

Bender, E. M., Gebru, T., McMillan-Major, A., & Mitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns, 2*(2), 100205. https://doi.org/10.1016/j.patter.2021.100205

Borges, J. L. (1998). *Collected fictions*. New York: Viking.

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. New York: W. W. Norton & Company.

Cao, L. (2024). Learn to refuse: Making large language models more controllable and reliable through knowledge scope limitation and refusal mechanism. *arXiv.* https://arxiv.org/abs/2311.01041

Centa Strahovnik, M. (2023a). Identiteta in pogovorni sistemi umetne inteligence (Identity and conversational artificial intelligence). *Bogoslovni vestnik, 83*(4), 853–864. https://doi.org/10.34291/BV2023/04/Centa

Centa Strahovnik, M. (2023b). *Artificial intelligence: Experiential aspects of human interaction with an artificial intelligence systems (chatbots).* Unpublished research project and research results, Faculty of Theology, University of Ljubljana.

Centa, M., & Strahovnik, V. (n.d.). Talking and thinking with AI: How AI chatbots restructure epistemic identity and virtue. [Manuscript in preparation].

Deery, O., & Bailey, K. (2022). The bias dilemma: The ethics of algorithmic bias in natural-language processing. *Feminist Philosophy Quarterly, 8*(3–4). https://doi.org/10.5206/fpq/2022.3/4.14292

Dong, G., Wang, H., Sun, J., & Wang, X. (2024). Evaluating and mitigating linguistic discrimination in large language models. *arXiv.* https://arxiv.org/abs/2404.18534

Dorobantu, M., Green, B. P., Ramelow, A., & Salobir, E. (2022). *Being human in the age of AI.* OPTIC Network. https://doi.org/10.13140/RG.2.2.32037.58080

European Commission. (2019). *Ethics guidelines for trustworthy AI.* Publications Office of the European Union. https://doi.org/10.2759/346720

European Parliament, 2024. *Artificial Intelligence Act.* (Regulation (EU) 2024/1689). http://data.europa.eu/eli/reg/2024/1689/oj

Floridi, L. (2023). AI as agency without intelligence: On ChatGPT, large language models, and other generative models. *Philosophy & Technology, 36*, 15. https://doi.org/10.1007/s13347-023-00621-y

Frankfurt, H. (1998). *The importance of what we care about.* Princeton, NJ: Princeton University Press.

Green, B. P. (2018). Ethical reflections on artificial intelligence. *Scientia et Fides, 6*(2), 9–31. https://apcz.umk.pl/SetF/article/view/SetF.2018.015

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea.* MIT Press.

Kislev, E. (2022). *Relationships 5.0: How AI, VR, and robots will reshape our emotional lives.* Oxford University Press.

Kraus, M., Seldschopf, P., & Minker, W. (2021). Towards a Trustworthy Chatbot for Mental Health Applications. In J. Lokoč et al. (Eds.), *MultiMedia Modeling* (pp. 354–366). Springer.

Mariani, M. M., Hashemi, N., & Wirtz, J. (2023). Artificial intelligence empowered conversational agents: A systematic literature review and research agenda. *Journal of Business Research, 161*, 113838. https://doi.org/10.1016/j.jbusres.2023.113838

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), Article 115, 1–35. https://doi.org/10.1145/3457607

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Lowe, R. J., … Leike, J., & Lowe, R. J. (2022). Training language models to follow instructions with human feedback. *arXiv*. https://doi.org/10.48550/arXiv.2203.02155

Raaijmakers, S. (2022). *Deep learning for natural language processing*. Shelter Island, NY: Manning.

Schwitzgebel, E., Schwitzgebel, D., & Strasser, A. (2023). Creating a large language model of a philosopher. *arXiv*. https://doi.org/10.48550/arXiv.2302.01339

Sharkey, A. (2016). Should we welcome robot teachers? *AI & Society, 37*, 1535–1544. https://doi.org/10.1007/s10676-016-9387-z

Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My chatbot companion: A study of human-chatbot relationships. *International Journal of Human-Computer Studies, 149*, 102601. https://doi.org/10.1016/j.ijhcs.2021.102601

Strahovnik, V. (2023). Etični in teološki izzivi velikih jezikovnih modelov (Ethical and theological challenges of large language models). *Bogoslovni vestnik*, 83(4), 839–852. https://doi.org/10.34291/BV2023/04/Strahovnik

The Future of Life Institute. (2023, March 22). *Pause giant AI experiments: An open letter.* https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., & Bernstein, A. (2020). Implementations in machine ethics: A survey. *ACM Computing Surveys, 53*(6), Article 132, 1–38. https://doi.org/10.1145/3419633

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., … Gabriel, I. (2022). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)* (pp. 214–229). Association for Computing Machinery. https://doi.org/10.1145/3531146.3533088

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2023). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.13370

ON ARTIFICIAL INTELLIGENCE: FROM TURING TO INTERROGATIVE MODEL

Arto Mutanen

Finnish Naval Academy, National Defence University, Finland

Abstract In this chapter I will outline the ethical questions linked with the use of autonomous military technology.

Keywords: Autonomous military technology, autonomous weapons, moral responsibility, ethical competence, military ethics.

Introduction

In artificial intelligence, the intention is to generate intelligent machines that execute intelligence which is human-like or other. Of course, for humans, human-like intelligence is something we are familiar with. The basis for such a belief can be understood as the Aristotelian definition of a human as a rational animal, which, in fact, distinguishes humans from other animals. The distinction entails that human intelligence becomes understood as a purely mental matter. As a mental matter, the body is something nonintelligent, or merely a matter of mechanical work, which can be seen indirectly from our everyday habits and from our everyday language.

As a mental matter, human intelligence, and intelligence more generally, is something otherworldly. Intelligence is not a matter of blood and flesh but a matter of purely esoteric spirituality. As intelligent agents, humans are members of a spiritual reality shared by angels and other esoteric agents. However, in this matter, the historical understanding has turned upside down. In ancient times, gods were not merely esoteric agents but part of human reality. The intellectual culture of antiquity was bodily-intellectual, where philosophical dialogues were a practice of the good life, rather than the construction of conceptual-theoretical doctrine (Hadon 1995).

The intellectual culture of human intelligence cannot just be removed and replaced with something better. We are members of this intellectual culture, and hence, replacement is not possible; because, in fact, we do not have an alternative notion of intelligence. Moreover, intellectual culture has deep and meaningful results, as the history of modern science and technology demonstrates (von Wright 1993).

To discuss the philosophy of artificial intelligence, it is not possible to take human intelligence as given and known. But to characterize a new kind of intelligence requires that we have a good understanding of the prevailing notion of intelligence. The prevailing notion of intelligence, which distinguishes humans from other animals, is of spiritual character, but, moreover, a certain kind of formalism is an essential factor in it. A good example of formalism is mathematical and logical reasoning, which, according to Hilbert, is the manipulation of symbols.

Formalism in philosophical thinking has long roots. An interesting example of such is Leibniz's characterization of the charm of music "even though its beauty consists only in the harmonies of numbers and in a calculation the beats or vibrations of sounding bodies" (quote from Pesic 2022, 144). However, before Leibniz, Descartes had already started to study music as a sound structure which joins mathematics and physics (Pesic 2022, 90). In this context, music is a good example, because it has an important intellectual role in Western thinking. Music speaks to its listener, but in the 20th century, for example, due to Descartes and Leibniz, we no longer thought that music could tell us anything in a semantic sense.

However, the topic "art and knowledge" has become an important and interesting problem in general, but also in the special field of arts (Young 2001; Scruton 1997; Robinson 1997). To the question of epistemic relevance of art, it has been given a negative answer by anti-cognitivists, like Stolnitz (1992), and a positive answer by cognitivists, like Young (2001). However, the interest in art's epistemic relevance shows the change in our epistemic and intellectual landscape. In the characterization of a new epistemic landscape, there is a need to reevaluate all the details of our intellectual culture. Hence, questions about art's epistemic relevance, the role of esotericism in the history of science, or the intellectual similarity between humans and other animals become intellectually important.

First Steps in Artificial Intelligence

According to Copeland (2004, 353), Alan Turing was the first to carry out substantial research in artificial intelligence. However, he did not use the term 'artificial intelligence'. His research considered problems of thinking machines, leaning machines, or machine intelligence. The term 'artificial intelligence' can be connected to "the title of a conference held at Dartmouth College, New Hampshire: The Dartmouth Summer Research Project on Artificial Intelligence" held in 1956 (Copeland 2004, 353).

In science, the question about the terms used is not the fundamental question, but the substantial topic of the research. Turing systematically studied questions of artificial intelligence. A problem he had in his mind was to how to generate a machine that can learn. The problem with computers is that their knowledge exists "only because their highly intelligent programmers had carefully selected appropriate information and cleverly packaged into conveniently useable blocks" (Copeland 2012, 186). To find the answer, Turing systematically, step by step, generated a new kind of machine program. A fundamental idea was his 'child machine,' which can learn like a human being. Turing (1950, 460) asks the following simple question: "Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's?" Even if the question is very simple, it was at that time revolutionary. The goal was not to produce a machine in which intelligence is its goal, but to generate a machine that can learn and, hence, after "an appropriate course of education one would obtain the adult brain" (Turing 1950, 460). The revolutionary character of the idea can be illustrated by the fact that when Turing introduced the idea of a child machine, Michie said that he wanted to study the idea for the rest of his life (Copeland 2012, 186).

In fact, the idea of a child machine does not come out of nothing. In his paper, "On Computable Numbers, with an Application to the Entscheidungsproblem" (1936), Turing analyzed the human computing process and, based on this analysis, he formulated the well-known universal Turing machines, which, in fact, are abstract models of present-day computers. (Copeland 2004, 1). The analysis of the human computing process was ingenious, and the characterized analysis was a model of the learning process for Turing. He generalized the idea in several places. In the paper, ”Lecture on the Automatic Computing Engine” (1947), he studies learning machines, and in his radio lecture, ”Intelligent Machinery, A Heretical Theory” (1951), he studied the role of mistakes in learning.

In Turing (1950), there is a formulation of the so-called Turing's test, which Turing called “imitation game,” which has sometimes been interpreted as the operational definition of intelligence. However, the fundamental structure of Turing's test is not to define intelligence, but to identify intelligent agents. Turing formulated the idea already in his paper, ”Intelligent Machinery” (1948), in which he developed a test for evaluating strategies in games. The intention is to test whether a watcher could identify whether a player is a machine which uses mechanical strategy or a human who uses creative strategy (Proudfoot 2017). The tests developed by Turing have an important role in the development of artificial intelligence. However, the role they play in it is not in generating better understanding of the notion of intelligence but showing how our prejudices color our understanding of artificial intelligence.

Of course, the applications of artificial intelligence in our everyday lives demonstrate the intelligence of machines. But a closer look at intelligent machinery shows that they are still programmed machines, which implies remarkable restrictions on their intelligibility. Especially, the very foundation of artificial intelligence does not need to be intelligent. However, there is no general agreement about what intelligence is. The idea Turing had in his child machines was that we humans learn to be intelligent. Hence, there is no need to have an intelligent starting point, but intelligence might be a learned property.

The idea that intelligence is not a given, but a learned property, changes our understanding of the foundation of an intelligent agency. In this sense, the notion of a child machine underlines that, in intelligence, the learning process is essential, not the intelligence starting point. This is extremely important, and, in fact, Turing recognized its importance. Turing (1936, 75-77) mentions that he justifies the basic idea of his notion of computation referring to “a direct appeal to intuition” with which he refers to his insightful analysis of human computation. According to Turing (1938, 192), intuition “consists in making spontaneous judgments which are not the result of conscious trains of reasoning,” and, hence, Turing's definition is intuitive in this sense, because there is no way to prove its adequacy. Of course, “it is possible to find some other way of verifying the correctness of an intuitive judgment” (Copeland 2004, 135), and, in this case, the detailed analysis functions as a verifying method, even if it cannot give full proof of it. It is important to note that we cannot get rid of intuition in logic, mathematics, or any other intelligent activity. However, ingenuity is a different story: because of the possibility to enumerate all the logical or mathematical proofs, it is possible to "imagine that all proofs take the form of a search through this enumeration for the theorem for which a proof is desired." Hence, ingenuity can be replaced by patience.

In Turing (1950), the question "Can machines think?" is under study. The problem is, as Turing notes, that the notions of 'machine' and 'think' are difficult to specify. Hence, Turing explicates the imitation game (Turing's test), which replaces the original question "Can machines think?" (Turing 1950, 441). Sure, thinking is connected to intelligence, but for Turing, the intention was to generate machines that can learn, which "may appear paradoxical to some readers" (Turing 1950, 463). The paradoxical character becomes evident, since the learning machines Turing refers to are, in fact, the Turing machines defined in Turing (1936), which follow "an instruction table" or a program. Hence, the behavior of the machines is completely determined, i.e., mechanical. So, to generate learning machines, they should "include a random element" (Turing 1950, 464), which makes them an interesting variant of these machines (Turing 1950, 445). The random element may cause the final results to be inconsistent, which is a fundamental characteristic of learning (Hintikka 2009; Başkent 2016).

Turing was searching for a computer which could be programmed by "inserting different programs into the memory, the machine is made to carry out different computations" (Copeland 2005,1). Thus, by enabling machines to change programs, the learning ability of mechanical machines can be developed. So, the simple idea executes the seeds of the founding ideas of modern artificial intelligence, which is "to construct some artefact which has a mind in the same sense that we have minds" Carter (2007, 1). More precisely, this is the foundation of so-called computational artificial intelligence.

Computational artificial intelligence is based on the idea that a computer, as an intelligent operator, is just an information processor. In the philosophy of mind, the computationalists assume that the human mind is like a computer, a mechanical information processor. The historical roots of computationalism are long; for example, Leibniz brings this up elegantly in the following famous characterization: "For it suffice to take their pencils in their hands, to sit down to their slates, and to say to each other (with a friend to witness, if they liked): calculemus – let us calculate" (quote from Scheutz 2002, 5). However, in computationalism, the foundational idea of mechanical computation is very restricted. The idea has been that the underlying computation is the Turing machine computability or recursivity, which, according to Church's thesis, characterizes the notion of computability (Sieg 1994). In fact, even if the Turing machine computability is defined in Turing (1936), Turing defined extremely interesting unending machines, which calculated decimal-by-decimal computable numbers. These machines inspired the generation of a generalized notion of computable, which actualizes a wide mechanism (Copeland 2002).

Artificial Intelligence

The 1930s saw the publication of numerous groundbreaking results, including Gödel's incompleteness results and the definitions of the concept of computability (Church, Post, Gödel, Turing). Of course, Turing's characterization of the notion of computability was ingenious; Turing analyzed the human computation process, and as a result of the analysis, he formulated abstract and formal machines that could compute. Similarly to human beings, the machines can be generalized such that there is a machine of fixed structure, called the universal Turing machine, which “is able to carry out every computation that can be carried out by any Turing machine whatsoever” (Copeland & Proudfoot 2005, 109) and, hence, also by “a human computer” (Copeland & Proudfoot 2005, 108). Moreover, Turing (1936, 63) recognized that “[t]here are certain types of process used by nearly all machines, and these, in some machines, are used in many connections,” which was ”effectively the first programming manual of the computer age” (Copeland 2004, 12-13).

Turing started to develop machines that can learn, i.e., intelligent machines. However, a problem lies in the fact that he started to search machines that learn, which led him to draw “a fairly sharp line between the physical and the intellectual capacities of a man” (Turing 1950, 442). This entails that the physical aspects of human intelligence will be excluded. Of course, human intelligence is also factual and even corporeal. As recognized above, Turing (1948) developed a game, which considers the problem of the identification of playing strategies, which was generalized in Turing (1950) to the problem of the identification of intelligent agents. That is, Turing did not consider the problem of the definition of intelligence or of intelligent behavior, as Turing himself explicitly says, “I don’t want to give a definition of thinking, (…) I don’t really see that we need to agree on a definition at all.” (Turing 1952, 494.) Turing searched for a method of identification of an intelligent agent.

If we look at Turing’s intention behind his analysis of human intelligence, there are certain kinds of problem-solving skills which were already present in his paper “On Computable Numbers.” The problem-solving skills are not merely theoretical but also, maybe mainly, practical skills. Unfortunately, we do not have a good, conceptualized analysis of skills. Nowadays, philosophers have done much more work with skill problems (Fridland & Pavese 2021). While speaking about practical skills, we might demolish the idea that physical “skills are a kind of disposition to know” (Stanley & Williamson 2017, 715). However, this does not mean that the knowledge referred to should be propositional. Skills are rather methodical knowledge or procedural knowledge, which can be understood as kinds of practical knowledge (Stanley & Williamson 2017).

Even if we would consider theoretical problems, like logical or mathematical problems, the mere discussion does not solve the problem. To solve the problem, it must be adequately identified before proper solving can start. The allocation of the solving method is not only merely mechanical, but a creative task, which can be seen from Turing’s (1936) analysis of human computation. The human computation process is a practical knowledge-driven process in which knowledge-based skills play a fundamental role. In fact, Turing machines are not thinking machines but calculating machines that factually execute computations.

Following Turing's ideas, we could characterize intelligent problem solving as the strategic structuring and solving of problems. This kind of problem-structuring-and-solving skill is a strategic skill. It is explicitly a skill and not (mere) knowledge, i.e., intelligence is a type of skill. Thus, a computer-based intelligence model cannot capture the most essential aspects of intelligence. Of course, Turing's analysis of intelligence is extremely important. He has been ahead of his time on many conceptual issues and has thus been able to advance and anticipate research related to artificial intelligence (Copeland 2004).

Turing emphasized that, in intelligence, it is essential to have a kind of dialogical attitude, which can be seen, for example, by the role he gave to mistakes in the learning process. Mistakes can be helpful only if they are recognized as mistakes. The identification of mistakes as mistakes supposes that the learning process is identified as a goal-tracking process. Mistakes might be factual mistakes, like mistakes in computation, or strategic mistakes, which direct the process in the wrong direction. The latter are of extreme importance. The identification of these strategic mistakes supposes the understanding of learning as a strategic process and, moreover, evaluating learning steps as part of such a strategy (Copeland 2004; Hintikka 2007; Başkent 2016).

Reading Turing closely, it is possible to recognize his intention to identify human intelligence by analyzing actual intelligent human activity. If we take this as a starting point, we do not follow the ideas of Turing's test, which considers the problem of identification, but we must take it one step further. In Turing's test, the idea of identification by discussion sounds very important. However, the procedure of Turing's test makes it possible to use "all sorts of tricks so as to appear more man-like," which makes the test unreliable (Proudfoot 2017, 303). However, it is possible to formulate a constructive model of Turing's idea of discussion by generating an interrogation game (Mutanen & Halonen 2019).

Jaakko Hintikka developed the interrogative model of inquiry, in which all research can be understood as a questioning-answering game, where questioning is a foundational methodological factor in research work. The central factors in the interrogative model are the following: The logic of questions, which shows how questions make logic ampliative. Logical formulations of the model show that the interrogative model has rigor logical structure, but at the same time, there are, in fact, several different models. In this sense, the interrogative model is a family of models, which have a common general logical structure. The logico-philosophical analysis of the methodology of knowledge acquisition shows that the interrogative model explicates the general strategic structure of all reasonable reasoning (Hintikka & Bachman 1991; Hintikka 2007; Mutanen & Halonen 2019; Hakli 2016).

The historical roots of the interrogative model go back to antiquity, where the Socratic question model (elenchus) is known. Plato explicated the questioning method in his dialogues. An especially good example is in Meno, in which the general methodological role of questioning becomes evident. In Meno, Socrates discusses mathematics with a slave boy, according to Socrates he is "not teaching the boy anything, but only asking him questions" (Meno). The questioning in Meno is a strategic questioning-answering process from ignorance to knowledge. The process, literally, constructs the searched knowledge such that each piece of information needed in the construction are explicated, such that the slave boy, who has no a priori knowledge about the topic, learns step-by-step during the process.

However, in search for new knowledge, answers to certain questions are "perfectly predictable," which "are the answers that are logically implied by the witness' earlier responses" (Hintikka 2007, 2-3). The study of these answers started Aristotelean logic (syllogistic), which Aristotle developed in his Prior Analytics. Besides logical steps, there are also steps which increase information into the reasoning process. These steps are called questioning steps, independent of the linguistic form of the steps. However, to be reasonable, the process should be strategic. Unfortunately, there is, and cannot be, an explicit theory of strategies, but, as Aristotle demonstrated in Topics, there can be a general philosophical characterization of different kinds of problem situations and reasonable strategies to manage them. This has been the main study in the interrogative model of inquiry (Hintikka 2007; Hintikka & Bachman 1991).

By providing a general theory of rational reasoning, the interrogative model of inquiry is significant in examining the foundations of artificial intelligence. However, the interrogative model of inquiry is theory- and context-sensitive in the sense that the interrogative reasoning is relativized on the background theory and a model of application, which is also explicated in the formalism of the interrogative model. In the philosophy of science, this is a benefit which can also be seen in the special application of artificial intelligence. But in artificial intelligence, the general idea is to find the general artificial intelligence, which is context-independent. It is possible to study the interrogative model from a general point of view, in which context dependency is released, then the results are general methodological recommendations, which indicate the conditions for the solvability or unsolvability of certain types of problem situations (Hendricks 2001; Kelly 1996).

On Intelligence

The interrogative model of inquiry can be localized or generalized into the intended abstraction level. In the concrete case, the logical structure of the reasoning process becomes characterized in detailed level. If the point of view is generalized, then the details of the reasoning process disappear, and the results become abstract minimality and maximality results (Kelly 1996). This is both interesting and important. However, intelligence is not an abstract theoretical approach, it is concrete intellectuality, which takes place here and now.

However, as articles in Frankish & Ramsey (2014) show, the discussion of artificial intelligence is extensive, but still a fundamental question of knowledge representation is open. Intelligence is not a linguistic skill, but a practical skill. Knowledge plays a role in intelligence, but, unfortunately, there is no general, all-field knowledge that would fulfill the knowledge requirements of general artificial intelligence. However, this emphasizes the idea that intelligence is a kind of general methodical knowledge that allows the intelligent agent to evaluate and solve problem situations. That is, intelligence is not merely knowing but to deliberately evaluate and solve problems, which is the strategic skill that the interrogative model of inquiry emphasizes.

The Socratic questioning method was applied by Plato to many types of problems. In Gorgias, Plato considers the problems of ethics. However, the intention is not to develop ethical theory but to characterize how to live a good life. The whole idea is not to find out logico-conceptual linguistic theory but a more practical exercise to learn to live, which is also more generally a target in ancient philosophy (Hadot 1995, 82-89). In this spirit, the development of the moral Turing's test seems to be a misleading approach (Proudfoot 2024). Hence, in ethics, the intention is not to formulate an adequate definition of good but to learn how to live (Pihlström 2014). In fact, the intention is not theoretico-conceptual but practical, which supposes the autonomous deliberation of way of life, which is not possible, because of conceptual and theoretical reasons, for present-day artificial intelligence (Hakli & Mäkelä 2019). Thus, artificial intelligence is significantly limited and flawed. One can only speculate whether this ultimately accounts for the limitations that have been observed in the reasoning ability of artificial intelligence (O'Neil 2016).

As noted above, Turing (1938) showed that ingenuity is not necessary in logic and mathematics, because it can be replaced by hard work. However, as O'Neil (2016) argues, the hard work does not entail intelligent results. There is a proper need for deliberation and visionary understanding, which can be implemented intuitively. However, intuitive thinking is a human skill which can and must be exercised. Intuition is not an open path to truth, but, rather conversely, because all the reasoning ends up to the point that there is no further reason even theoretically discoverable (Russell 1912, 64). As Turing (1938) also emphasized, intuition is inevitable. Without intuition, we cannot recognize the problems we face.

Even if intuition is necessary, it does not mean that intuition would be infallible. Moore (1903) argues that ethical good is something undefined, and we recognized it by moral intuition. Still, he says that moral intuition is not infallible. All of this shows the place for intelligence, which can be identified as a deep skill of deliberation. There is no doctrine of such a skill, but it is the ability to deliberate, i.e., to consider the situation in an enlightened spirit. Therefore, it is not good enough to exercise merely scientific reasoning; there is also the need for more holistic exercises, which Hadon (1995) calls spiritual exercises. So, in intelligence, along with knowledge, physical and artistic wisdom also need to be developed.

Artificial intelligence systems have been developed that perform exceptionally well in specific areas. Thus, artificial intelligence has already proven to be a fruitful and interesting field. However, there are still no working models of general artificial intelligence, neither in theory nor in practical applications, as from the argumentation in Hakli & Mäkelä (2019) can infer.

General artificial intelligence is something that is both aspirational and largely unknown. We don't really know what human intelligence is, let alone know more generally what intelligence might mean. There are lot of open questions to be answered. Unfortunately, neither humans nor other intelligent agents know how to solve the problem of intelligence.

Closing Words

Artificial intelligence seeks intelligent solutions to specific problem situations that can be solved using the means offered by information technology. The theoretical discussion of artificial intelligence is broad and diverse, encompassing both the logical-conceptual foundations of artificial intelligence and the practical-theoretical questions of its application. At the same time, the aim is to find a design for intelligence or general intelligence.

Obviously, solutions to all these problems cannot explicate a single model of artificial intelligence. Moreover, it is unclear whether the searched intelligence should be like human intelligence or something else. In fact, there is no agreement what human intelligence eventually is or what more general intelligence could be. It is quite generally accepted that intelligence is something practical, maybe even physical. Hence, the distinction between human and animal intelligence disappears step by step (DeGrazia 1996). The fact that physical skills are a type of knowledge does not diminish the close connection between human and animal intelligence, since the skills, whether they are human or animal skills, are independently equal.

Perhaps it is good to start with the fact that human intelligence, in all its diversity, is just a certain form of intelligence. Animals' intelligence may be different, but human intelligence and animal intelligence have a lot in common. Perhaps value intelligence, the search for meaning, is essential for humans (Nagel 1993). Meaningfulness of life is not merely an epistemic issue but also an aesthetic and ethical issue; so along with science, art plays an essential role in constructing the desired meaningfulness (Robinson 1997; Scruton 1974; Young 2001).

We have formulated a general logico-conceptual framework, which shows a general logic of reason, and its reflective practice might be seen as a key factor in intelligence. Moreover, the theoretical framework of the model is flexible, such that, alongside theoretical knowledge, both aesthetic and ethical understanding can function as an orienting foundation. Fortunately, the model has explicit logical structure; hence, it can be understood as a theoretical fundamental of artificial intelligence.

References

Aristotle. (n.d.). *Prior Analytics* (A. J. Jenkinson, Trans.). The Internet Classics Archive. Retrieved August 15, 2025, from https://classics.mit.edu/Aristotle/prior.1.i.html

Aristotle. (n.d.). *Topics* (W. A. Pickard-Cambridge, Trans.). The Internet Classics Archive. Retrieved August 15, 2025, from https://classics.mit.edu/Aristotle/topics.1.i.html

Başkent, C. (2016). Inquiry, Refutations, and the Inconsistent. In C. Başkent (Ed.), *Perspectives on Interrogative Models of Inquiry: Developments in Inquiry and Questions* (pp. 57–72). Springer.

Carter, M. (2007). Minds and Computers: An Introduction to the Philosophy of Artificial Intelligence. Edinburgh University Press.

Copeland, B. J. (2002). Narrow versus Wide Mechanism. In M. Scheutz (Ed.), *Computationalism: New Directions* (pp. 59–86). The MIT Press.

Copeland, B. J. (2012). *Turing: Pioneer of the Information Age*. Oxford University Press.

Copeland, B. J. (Ed.). (2004). The Essential Turing: The Ideas That Gave Birth to the Computer Age. Clarendon Press.

Copeland, B. J., & Proudfoot, D. (2005). Turing and the Computer. In J. B. Copeland et al. (Eds.), *Alan Turing: Electronic Brain. The Struggle to Build the ACE, the World's Fastest Computer* (pp. 107–148). Oxford University Press.

DeGrazia, D. (1996). Taking Animals Seriously: Mental Life and Moral Status. Cambridge University Press.

Frankish, K., & Ramsey, W. M. (Eds.). (2014). *The Cambridge Handbook of Artificial Intelligence*. Cambridge University Press.

Fridland, E., & Pavese, C. (Eds.). (2021). The Routledge Handbook of Philosophy of Skill and Expertise. Routledge.

Hakli, R. (2016). Inquiry and Justification. In C. Başkent (Ed.), *Perspectives on Interrogative Models of Inquiry: Developments in Inquiry and Questions* (pp. 1–14). Springer.

Hakli, R., & Mäkelä, P. (2019). Moral Responsibility of Robots and Hybrid Agents. *The Monist*, *102*(3), 259–275. https://doi.org/10.1093/monist/onz009

Hintikka, J. (2007). *Socratic Epistemology*. Cambridge University Press.

Hintikka, J. (2009). IF Logic Meets Paraconsistent Logic. In W. Carnielli, M. E. Coniglio, & I. M. L. D`Ottaviano (Eds.), *The Many Sides of Logic* (pp. 3–13). College Publications.

Hintikka, J., & Bachman, J. (1991). *What If…?: Toward Excellence in Reasoning*. Mayfield Publishing Company.

Moore, G. E. (1986). *Principia Ethica*. Cambridge University Press. (Original work published 1903)

Mutanen, A., & Halonen, I. (2019). Turingin testi, interrogatiivimalli ja tekoäly. *Ajatus*, *77*, 169–204.

Nagel, T. (1993). The Absurd. In J. Perry & M. Bratman (Eds.), *Introduction to Philosophy. Classical and Contemporary Readings* (2nd ed.). Oxford University Press. (Original work published 1979)

O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown.

Pesic, P. (2022). *Music and the Modern Science*. MIT Press.

Pihlström, S. (2014). *Taking Evil Seriously*. Palgrave Macmillan.

Plato. (n.d.). *Gorgias* (B. Jowett, Trans.). The Internet Classics Archive. Retrieved August 15, 2025.

Plato. (n.d.). *Meno* (B. Jowett, Trans.). The Internet Classics Archive. Retrieved August 14, 2025, from https://classics.mit.edu/Plato/meno.html

Proudfoot, D. (2017). Turing's Concept of Intelligence. In B. J. Copeland et al., *Alan Turing: Electronic Brain. The Struggle to Build the ACE, the World's Fastest Computer* (pp. 301–307). Oxford University Press.

Proudfoot, D. (2024). Turing's Test vs the Moral Turing Test. *Philosophy & Technology*, *37*(4), 134. https://doi.org/10.1007/s13347-024-00825-w

Robinson, J. (Ed.). (1997). *Music and Meaning*. Cornell University Press.

Russell, B. (1985). *The Problems of Philosophy*. Oxford University Press. (Original work published 1912)

Scruton, R. (1974). *Art and Imagination*. Methuen.

Scruton, R. (1997). *The Aesthetics of Music*. Oxford University Press.

Sieg, W. (1994). Mechanical Procedures and Mathematical Experience. In A. George (Ed.), *Mathematics and Mind* (pp. 71–117). Oxford University Press.

Stanley, J., & Williamson, T. (2017). Skill. *Nous*, *51*(4), 713–726. https://doi.org/10.1111/nous.12144

Stolnitz, J. (1992). On the Cognitive Triviality of Art. *The British Journal of Aesthetics*, *32*(3), 191–200. https://doi.org/10.1093/bjaesthetics/32.3.191

Turing, A. M. (1936). On Computable Numbers, with an Application to the Entscheidungsproblem. In B. J. Copeland (Ed.), *The Essential Turing: The Ideas That Gave Birth to the Computer Age* (pp. 58–90). Clarendon Press. (Original work published 1937)

Turing, A. M. (1938). Systems of Logic Based on Ordinals. In B. J. Copeland (Ed.), *The Essential Turing: The Ideas That Gave Birth to the Computer Age* (pp. 146–204). Clarendon Press.

Turing, A. M. (1948). Intelligent Machinery. In B. J. Copeland (Ed.), *The Essential Turing: The Ideas That Gave Birth to the Computer Age* (pp. 410–432). Clarendon Press.

Turing, A. M. (1950). Computing Machinery and Intelligence. In B. J. Copeland (Ed.), *The Essential Turing: The Ideas That Gave Birth to the Computer Age* (pp. 441–464). Clarendon Press.

Turing, A. M. (1952). Can Automatic Calculating Machines Be Said to Think? In B. J. Copeland (Ed.), *The Essential Turing: The Ideas That Gave Birth to the Computer Age* (pp. 487–506). Clarendon Press.

Young, J. O. (2001). *Art and Knowledge*. Routledge.

BLURRING BOUNDARIES: ON ETHICAL IMPLICATIONS OF AI IN MILITARY-THEMED TOYS

Ramunė Kasperė and Valdas Grigaliūnas

Kaunas University of Technology, Lithuania

Abstract This paper addresses the connections between play, war, artificial intelligence (AI) and ethics, focusing on the implications of AI-driven military toys. The integration of AI into military play and toys introduces new ethical dimensions and raises questions about acceptance and normalization of algorithmic warfare. Children and young adults are exposed to emerging conceptualizations of human-machine interaction in war scenarios, through technologies like smart robotic soldiers or AI-powered war simulations. The paper aims to explore how developing AI-interfaced military toy and play practices play a role in shaping real-world perspectives and ethics attitudes of society on AI autonomy and conflict.

Drawing on both historical context and contemporary technological advancements, we argue that, without appropriate guidance, robust regulation, and broader public discourse on values and AI ethics, these toys may contribute to the normalization of unsupervised autonomous AI-driven decision-making, desensitization to real-world conflict, diminished critical thinking, and weakened human responsibility in war scenarios. The paper also calls for interdisciplinary research to reassess how AI shapes societal values and influences the next generation for an increasingly automated and ethically complex world.

Keywords: AI, decision-making, desensitization, ethics, military-themed toys and play, responsibility.

Introduction

Since the advancement of artificial intelligence and robotics technologies, scholars have argued that autonomous systems like robots should be designed and developed so that they are capable of ethical and moral considerations in their decision making (Allen & Wallach, 2012). Killer robots and autonomous weapons are becoming a reality nowadays. Their use requires no human troops on the battlefield, thus minimizing human harm. On the other hand, decision making in such autonomous war systems is possible without human control. The edges of responsibility, as argued by Alegre (2024), are thus becoming blurred. Through brain-computer interfaces, brain signals for human commands can be read requiring no real action from a human to pass on the decision to the machine, in which case it becomes unclear who/what is responsible for the action and the consequence (Alegre, 2024). The consequence may be fatal, and it may be the cause of a hallucination that artificial intelligence (AI) is prone to. Concerns among ethics researchers and philosophers have been raised whether responsibility would lie with the machine or the human (Robillard, 2017; Madock, 2024). Whatever the reasoning, it is true that setting a machine has required human intervention at some point (Michael et al., 2020).

Scholars demand attention be drawn to researching the impact of AI on children and youngsters as generative AI is being increasingly integrated into a plethora of tools, platforms, and digital environments, including games and creative software (Leaver & Srdarov, 2025). The importance of integrating modern technologies in childhood play and education for many reasons has been advocated, including AI-interfaced robotic toys (Kewalramani et al., 2021a; 2021b). Evidence exists that the introduction of AI technologies early in children's education can empower them to understand AI devices, but such AI literacy development should be done through properly framed content and introduction of the ethical design of AI, in order to teach children to address bias and promote fairness (Williams et al., 2019). The imperative to enhance children's understanding and responsible use of technology lies primarily with educational institutions, but also with parents who may not necessarily be equipped with the knowledge, digital literacy, resources or willingness to guide their children (Kasperė & Liubinienė, 2022).

AI-interfaced military-themed toys, including AI-powered robots, have recently appeared on the market, gaining attention for play among children, adolescents, and generally computer game and robotics enthusiasts. Like predecessor warfare simulations and computer games, AI-interfaced military-themed toys may offer more excitement as they combine interactive AI-driven scenarios and educational experiences.

Although it has been acknowledged that playing with military toys generally is unrelated to harmful intent, children with higher levels of antisocial behavior may show more violent actions during pretend play (Hart & Tannock, 2013). What logically follows is that playing with AI-interfaced military-themed toys may have unknown impacts on children and youngsters, and consequently on society as a whole, including desensitization to warfare in the most general sense, the normalization of autonomous violence, and unconventional and unprecedented warfare-related conceptualizations of human-machine interaction where human responsibility becomes diminished, if not entirely eliminated from decision making.

The impact of AI's transformational power is still to be revealed. The European Union is claiming it has an ambition to become a global leader in AI. An action plan has been outlined to achieve the goal (The AI Continent Action Plan, 2025). The EU AI-related strategies also target defence and security, among other sectors. Foundational AI technologies are considered crucial for military pre-eminence (EC White Paper for European Defence Readiness 2030, 2025). Following the EU's ambition and taking into consideration the slow pace of legislation, AI-related ethical and sensitive implications and long-term impacts risk being unattended.

AI-interfaced toy marketing in the European Union (EU) follows strict requirements. Article 50 of the EU AI Act (2024), which entered into force in August 2024 and will be fully applicable to all sectors in August 2026, states that toys are among those products that are appropriate to classify as high-risk and need to undergo rigorous conformity procedures. The regulation has an extended transition period for AI applications considered high-risk until August 2027. However, in general, the EU legislation regulates those toys on the EU market must not jeopardise the safety and health of children (Directive 2009/48/EC).

The paper thus aims to explore how developing AI-interfaced military toy and play practices play a role in shaping real-world perspectives and ethics attitudes of society on AI autonomy and conflict.

From Tin Soldiers to Smart Bots - Evolution and Implications

The impacts that military-themed toys and war play bring onto society are numerous and diverse, including spheres of technology, culture, psychology, and education.

The evolution of military-themed toys and war play reaches as far as the existence of the war itself. Since ancient times until quite recently, they were simple and handicraft reflections and symbolic representations of real-world war-related scenarios. In ancient civilizations and Egypt, Greece, Rome, China and indigenous cultures, such military-themed toys were made of wood, bamboo, clay, ivory, bronze, lead and other materials of the time. They mostly were figures of soldiers, toy gear like swords, helmets, bows and arrows, horses and riders, etc. Military-themed toys were used not only for entertainment in the form of miniature weapons but also for future soldier training, i.e. wargaming (a concept explained below). In some cultures, they played symbolic roles in various rituals and ceremonies, some were also put in graves for protection and status purposes.

Throughout the Medieval Ages, military-themed toys served the purposes of reinforcing feudal values, while during the Renaissance and Enlightenment, such toys were often used for educational purposes – teach geography and natural sciences, leadership, strategic thinking – especially by the noblemen. The first apparent technological change in military-themed toys occurred around the 15th century when printing was invented, and paper toys could be produced together with maps as a means of mass education. The 19th century saw a boost in the mass production of tin soldiers that were affordable and available because of some leading companies, especially those in Germany. In many countries, toy soldiers were made in real outfits and served to build pride and patriotism. In the Industrial Age, production of military-themed toys started in factories, and the 20th century's introduction of plastic further revolutionized the toy industry and transformed societal perceptions of play. The USA has been dominating the global arms trade. Since war discourse is complex, involving military, politics, and industry, war toy manufacturing naturally reflects particular weapons, clothing, and special force ideology (Machin & Van Leeuwen, 2009).

AI has been gaining momentum in the last decade because of affordable sensors, machine learning, and cloud connectivity. Still, the AI integration in toys is a recent concept for several reasons. First, emerging technologies and AI-based tools at an affordable price. Second, the rapidly increasing demand for such toys (one of the main reasons is the STEM concept), at the same time, is expanding the smart toys market, expecting 43.5 billion USD in 2025 and 138.7 billion USD by 2030 (Mordor Intelligence Research & Advisory, 2024). Third, drag-and-drop graphic programming, such as Scratch (by MIT), Blockly (by Google), or similar block-based programming platforms that make user-friendly programming for young kids.

Currently, several main categories can be distinguished in smart toys. The first category would be small desktop robots like Eilik, Emo, or Vector, featuring some degree of emotions, intellect, corresponding to touch, sound, capable of interacting with humans, or providing information. The second category includes more advanced AI-centric features, capable of providing educational tasks and maintaining a conversation, for example, Moxie, Jibo, or Misa. Another very popular smart toy group is smart pets, for example, Loona, AIBO, Chip or Bittle X, integrating computer vision and AI that allows interacting robots with children.

An interesting fact is that such non-realistic toys may encourage high-quality play (Heikkilä, 2021). They are now available on the market, affordable and desirable by children, youngsters and adults. Modern-day robots for play have all the elements and technologies needed for modern combat, human-operated or unmanned vehicles. For example, DJI RoboMaster S1 integrates computer vision, a gimbal to stabilize the camera, is equipped with four brushless motors, sensors, and a processing unit. Children can learn robotics and AI technologies in a user-friendly drag-and-drop programming way using Scratch, which enables learning by playing (Fagerlund et al., 2021). Adolescents may use more advanced coding using the Python language. The toy has integrated AI-based decision-making on whether to shoot or not at the target. Additionally, smart toys can be used as platforms to conduct research (Blumenkamp et al., 2024).

What once belonged to science fiction has now become a reality. Following Haraway (1985), the boundaries between science fiction and social reality are mere optical illusions. The illusions are becoming more transparent as youngsters become accustomed to AI-driven decision-making through their engagement with AI-interfaced military-themed toys and play. The distance to the consequences of real-world conflict scenarios also means emotional detachment and decline of ethical awareness. The increasingly ambiguous perception of who eventually is responsible for the outcomes of AI-interfaced warfare or no obvious understanding, on the part of youngsters playing warfare, of the responsibility for the human action in the interaction reflects Haraway's vision of blurred responsibilities.

Throughout evolution, military-themed toys have played diverse roles in society – educational, cultural, and nationalistic sentimental. The age of artificial intelligence has brought another dimension. AI-interfaced military-themed toys are not only transforming society by introducing new understandings of the role humans and machines play in war but also are blurring the boundaries between peace and conflict, play and simulation, and raising complex questions of the interaction – and thus responsibility – of the human and the machine. Ethics and morality are among the questions.

Ethics of Simulated AI-interfaced Conflict

The geopolitical situations of the 21st century and interests of countries directly and indirectly involved in conflicts are immediate proof that war remains deeply embedded in the societal mindset. War has globally become normalized and justified as a strategic instrument.

War and gaming are two closely linked concepts historically, culturally, and ethically. Wargaming, or in other words, a game about war, in military settings refers to "an experiment in human interaction" that explores the decision-making processes of those who participate (Perla, 2022). The concept of wargaming is linked to military-themed toys and play intended for entertainment as the conflict simulation is imaginative in childhood but formalized in professional military planning used for educating future leaders. Both are related to strategic thinking, learning and decision-making. Experts have drawn attention to the potential impacts of wargaming on real-world decisions of participants. Barzashka (2023) argues that "[p]layers who are or will become real-world decision-makers could be primed by their gaming experiences, possibly affecting future decisions in subtle ways." Playing military-themed toys and virtual games may offer repetitive engagement and exposure to unrealistic representation and interpretation of war. In similar ways, the developmental long-term effects on players may be profound and, at the same time, form simplified perceptions of conflict.

AI-interfaced wargaming needs to be addressed, studied and governed from ethical perspectives (Barzashka, 2023). Ethics in relation to AI-interfaced military-themed toys and play likewise call for scrutiny and regulation.

One of the ethical concerns in this respect is that AI-interfaced military-themed toys and play may contribute to the **desensitization of society to warfare**. Perez et al. (2017) have expressed concerns that autonomous warfare could lead to its acceptance as a standard among youth. Younger generations understand and accept AI-controlled combat scenarios as youngsters are playing war without seeing and grasping the harsh consequences that are blunt and blatant in a real-world scenario. The violence, injury and death aspects are much less brutal and severe when they are experienced in simulations. In reality, none of the outcomes lead to consequences other than entertainment and reduced sensitivity to brutality. Apart from many other, ethical complexities of war are reduced to a game where violence may come to be perceived as routine or even heroic.

AI-interfaced military-themed toys and play may discourage critical thinking in children and gamers leading to increased **acceptance of AI decision-making**. It has been argued that wargaming develops critical thinking and encourages human-led decision making (Combe, 2021; Wong, 2016). Military-themed toys and play, unlike wargaming, may shift the focus from taking the responsibility to make active decisions to accepting AI-controlled decision making, as they are primarily for entertainment purposes. In other words, you think less of the consequences when you play a game. Therefore, in AI-controlled combat play scenarios, youngsters may get familiar with situations and outcomes that require less human presence and decision making. They might perceive the decision-making done by AI as more trustworthy, logical, rational and, perhaps more importantly, to a great extent faster than those made by humans. Such experiences may not promote critical thinking and assessment but rather inhibit it.

Critical assessment of the effects and aftershocks that such use of AI in real-world combat would cause is much more abstract, especially because youngsters have come to take advantage of freely available and easily accessible AI tools for many diverse purposes. The tendency to place excessive trust in AI, e.g. LLMS, and consequently to over-rely on them, has been studied and explained by such features of genAI tools like anthropomorphic design, fluency and interactivity (Liesenfeld et al., 2023). Trustworthiness in AI is a multifaceted and complex concept incorporating such components as openness, transparency and accountability (Liesenfeld et al., 2023; Puri & Keymolen, 2023). Ethical considerations also play a role in whether and to what extent technology, including AI, is to be trusted (Ryan, 2020). Evidence on children's overly positive attitudes towards AI, LLMs and robots has been scientifically documented (Andries & Robertson, 2023). Although evidence exists that youngsters regularly navigate ethical and critical dimensions of AI, it has also been found that a lack of critical thinking towards technology as well as over-reliance on automated solutions are common characteristics of the smart-phone generation (Higgs & Stornaiuolo, 2024; Machidon, 2025). Ultimately, AI-interfaced military-themed toys and play may inadvertently promote acceptance of AI decision making.

A crucial ethical concern is the potential **decrease of human responsibility** in war play scenarios, and thus a tendency to attribute more accountability to AI decision-making. In play, youngsters, and humans more broadly, may begin to feel less accountable for the outcomes of machine-produced actions, just because they are involved in decision-making less and are not physically, and therefore emotionally, in contact with the aftermaths produced. Such conscious or unconscious distancing and detachment may lessen the sense of responsibility, making it a normalizing experience. It has already been reported in scientific literature that people tend to ascribe moral blame to AI-powered systems, especially when they do harm (Hong et al., 2020; Kneer & Stuart, 2021; Malle et al., 2019). Villegas-Galaviz and Martin (2021) argue that "introduction of AI to decision making increases moral distance" and, therefore, reduces accountability and personal responsibility. Abandonment of decision-making to AI brings to the fore the importance of considering the ethics of AI-interfaced military-themed toys and play, as what happens while playing can gradually and eventually be transferred to real-world perceptions of conflict and AI.

If we consider further, we have seen that military-themed toys, like toy soldiers, once played a significant role in fostering a peaceful national sentiment, pride and a sense of identity. Currently, a reduced sense of patriotism may be felt in many countries and nations. Some surveys, for example, in the United Kingdom, have found that younger respondents were less willing to consider an option of volunteering for military service or even going to the front, in the event of armed aggression against their country. A survey conducted in the United Kingdom revealed that only 11% of young people would risk their lives to defend their country (Silver, 2025), which is a sharp decline from 38% as of 2022 (Dinic, 2022). When Britons were asked whether they would go to defend their country in the event of an attack, only senior citizens expressed strong approval, while younger respondents said they would flee the country or attempt to evade recruitment. The reasons for this shift are diverse, various, and complex; however, the AI-interfaced military-themed toys and play may contribute not only to reducing the sense of responsibility but also to weakening the sense of identity and national sentiment. Virtual warfare play and automated systems may create detachment from tangible real-world outcomes, undermining emotional connections, and thus possibly weakening national identity and pride, diminishing loyalty to homeland, and making the willingness to defend it more of a distant action, requiring no personal engagement and duty.

Conclusion

Geoffrey Hinton, often referred to as the "godfather of AI," has repeatedly warned about the **existential threats AI poses to humanity**. He particularly highlights the risk of technology gaining control without human oversight, a scenario he believes could be "sooner than later" and is "serious and fairly close". (Smith, 2023) The integration of AI into military-themed toys and play could significantly contribute to a societal transformation regarding war and technology. Such toys transcend mere entertainment or education; they can have **long-term effects on players**, reshaping their perceptions of conflict, decision-making, and responsibility.

AI-interfaced military play may also **diminish the perception of violence**, presenting it without consequence and thereby fostering desensitization among both youngsters and society at large. Children might come to accept autonomous warfare as a standard, especially due to an over-reliance on AI in play scenarios. Consequently, they may begin to consider human responsibility and accountability to be of lesser significance.

In this paper, we draw attention to a few, among multiple other, concerns that AI-interfaced military-themed toys and play may in the long term bring about, including undermined critical thinking and assessment as well as numerous ethical considerations. It is necessary not only to bring the attention of scholars, educators and developers to the issue but also to discuss regulatory frameworks with policy makers and encourage public discourse on AI ethics. Without proper guidance and recurrent and persistent engagement in dialogue on ethics, AI-interfaced military-themed play may in the long run normalize attitudes and perceptions that might turn out to be incompatible with ethical values.

References

Alegre, S. (2024). Human Rights, Robot Wrongs. London: Atlantic Books.

Allen, C., & Wallach, W. (2012). Moral machines: Contradiction in terms or abdication of human responsibility. In P. Lin, K. Abney, & G.A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 55–68). Cambridge: The MIT Press.

Andries, V., & Robertson, J. (2023). Alexa doesn't have that many feelings: Children's understanding of AI through interactions with smart speakers in their homes. *Computers and Education: Artificial Intelligence, 5*, 100176. https://doi.org/10.1016/j.caeai.2023.100176

Barzashka, I. (2023, December 4). Wargames and AI: A dangerous mix that needs ethical oversight. *Bulletin of the Atomic Scientists*. Retrieved April 20, 2025, https://thebulletin.org/2023/12/wargames-and-ai-a-dangerous-mix-that-needs-ethical-oversight/?utm_source=chatgpt.com

Blumenkamp, J., Shankar, A., Bettini, M., Bird, J., & Prorok, A. (2024). The Cambridge robomaster: An agile multi-robot research platform. arXiv preprint arXiv:2405.02198.

Combe, I. I. (2021). Educational wargaming: design and implementation into professional military education. *Journal of Advanced Military Studies, 12*(2). https://doi.org/10.21140/mcuj.20211202003

Dinic, M. (2022, September 21). YouGov Study of War: Britons on serving in the armed forces. *YouGov*. Retrieved April 20, 2025, from https://yougov.co.uk/politics/articles/43813-yougov-study-war-britons-serving-armed-forces?utm_source=chatgpt.com

Hart, J. L., & Tannock, M. T. (2013). Young children's play fighting and use of war toys. In P. K. Smith (Ed.), Encyclopedia on Early Childhood Development (pp. 1–5). Centre of Excellence for Early Childhood Development (CEECD) and Strategic Knowledge Cluster on Early Child Development (SKC-ECD). Retrieved May 17, 2025, from https://research.usc.edu.au/esploro/outputs/encyclopediaEntry/Young-Childrens-Play-Fighting-and-Use/99449475402621

Hong, J. W., Wang, Y., & Lanz, P. (2020). Why is artificial intelligence blamed more? Analysis of faulting artificial intelligence for self-driving car accidents in experimental settings. *International Journal of Human–Computer Interaction, 36*(18), 1768–1774. https://doi.org/10.1080/10447318.2020.1785693

EU AI Act. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. Official Journal of the European Union, L 231, 1–164. Retrieved May 11, 2025, from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689

EC The White Paper for European Defence – Readiness 2030 (2025, March 2025). Retrieved May 11, 2025, from https://commission.europa.eu/document/download/e6d5db69-e0ab-4bec-9dc0-3867b4373019_en?filename=White%20paper%20for%20European%20defence%20%E2%80%93%20Readiness%202030.pdf

EC The AI Continent Action Plan, 2025 (2025, April 09). Retrieved May 11, 2025, from https://digital-strategy.ec.europa.eu/en/library/ai-continent-action-plan

Fagerlund, J., Häkkinen, P., Vesisenaho, M., & Viiri, J. (2021). Computational thinking in programming with Scratch in primary schools: A systematic review. *Computer Applications in Engineering Education, 29,* 12–28. https://doi.org/10.1002/cae.22255

Heikkilä, M. (2021). Boys, weapon toys, war play and meaning-making: prohibiting play in early childhood education settings? *Early Child Development and Care, 192*(11), 1830–1841. https://doi.org/10.1080/03004430.2021.1943377

Higgs, J.M., & Stornaiuolo, A. (2024). Being human in the age of generative AI: Young people's ethical concerns about writing and living with machines. *Reading Research Quarterly, 59*, 632–650. https://doi.org/10.1002/rrq.552

Kasperė, R., & Liubinienė, V. (2022). Digital competence and teacher training overview: Is Lithuania ready for digitalism in education? In Ł. Tomczy, & L. Fedeli (Eds), *Digital Literacy for Teachers. Lecture Notes in Educational Technology.* Springer, Singapore. https://doi.org/10.1007/978-981-19-1738-7_16

Kewalramani, S., Kidman, G., & Palaiologou, I. (2021a). Using Artificial Intelligence (AI)-interfaced robotic toys in early childhood settings: a case for children's inquiry literacy. *European Early Childhood Education Research Journal, 29*(5), 652–668. https://doi.org/10.1080/1350293X.2021.1968458

Kewalramani, S., Palaiologou, I., Dardanou, M., Allen, K.-A., & Phillipson, S. (2021b). Using robotic toys in early childhood education to support children's social and emotional competencies. *Australasian Journal of Early Childhood, 46*(4), 355–369. https://doi.org/10.1177/18369391211056668

Kneer, M., & Stuart, M.T. (2021). Playing the blame game with robots. In Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21 Companion). Association for Computing Machinery, New York, NY, USA, 407–411. https://doi.org/10.1145/3434074.3447202

Leaver, T., & Srdarov, S. (2025). Generative AI and children's digital futures: New research challenges. *Journal of Children and Media*, *19*(1), 65–70. https://doi.org/10.1080/17482798.2024.2438679

Liesenfeld, A., Lopez, A., & Dingemanse, M. (2023). Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. In Proceedings of the 5th International Conference on Conversational User Interfaces (CUI '23). Association for Computing Machinery, New York, NY, USA, Article 47, 1–6. https://doi.org/10.1145/3571884.3604316

Machidon, OM. (2025). Generative AI and childhood education: lessons from the smartphone generation. *AI & Society.* https://doi.org/10.1007/s00146-025-02196-y

Machin, D., & Van Leeuwen, T. (2009). Toys as discourse: children's war toys and the war on terror. *Critical Discourse Studies*, *6*(1), 51–63. https://doi.org/10.1080/17405900802560082

Madock, J. (2024). Robot warfare: the (im)permissibility of autonomous weapons systems. *AI Ethics*. https://doi.org/10.1007/s43681-024-00567-7

Malle, B.F., Magar, S.T., Scheutz, M. (2019). AI in the Sky: How People Morally Evaluate Human and Machine Decisions in a Lethal Strike Dilemma. In M. Aldinhas Ferreira, J. Silva Sequeira, G. Singh Virk, M. E. Tokhi, & E. Kadar (Eds.), *Robotics and Well-Being. Intelligent Systems, Control and Automation: Science and Engineering*, vol 95. Springer, Cham. https://doi.org/10.1007/978-3-030-12524-0_11

Michael, K., Abbas, R., & Roussos, G., Scornavacca, E., & Fosso-Wamba, S. (2020). Ethics in AI and Autonomous System Applications Design. *IEEE Transactions on Technology and Society, 1*(3), 114–127. https://doi.org/10.1109/TTS.2020.3019595

Mordor Intelligence Research & Advisory. (2024, June). Smart Toys Market Size & Share Analysis - Growth Trends & Forecasts (2025 - 2030). Mordor Intelligence. Retrieved May 20, 2025, from https://www.mordorintelligence.com/industry-reports/smart-toys-market

Perez, V. E., Wortham, R. H., & Eugene, M. (2017). When AI goes to war: youth opinion, fictional reality and autonomous weapons. *The ORBIT Journal, 1*(1), 1–20. https://doi.org/10.29297/orbit.v1i1.19

Perla, P. (2022) Wargaming and the cycle of research and learning. *Scandinavian Journal of Military Studies, 5*(1), p. 197–208. https://doi.org/10.31374/sjms.124

Puri, A., & Keymolen, E. (2023). Of ChatGPT and trustworthy AI. *Journal of Human Technology Relations, 1*(1), 1–10. https://doi.org/10.59490/jhtr.2023.1.7028

Robillard, M. (2018). No such thing as killer robots. *Journal of Applied Philosophy*, *35*(4), 705–717. https://doi.org/10.1111/japp.12274

Ryan, M. (2020). In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics, 26,* 2749–2767. https://doi.org/10.1007/s11948-020-00228-y

Silver, E. (2025, February 10). 'Disgrace!' Gen Z think UK is 'racist' and would not fight for UK as army chief issues warning. *GB News*. Retrieved May 11, 2025, from https://www.gbnews.com/news/genz-uk-racist-fight-britain-farage-disgrace

Smith, C. S. (2023, May 4). Geoff Hinton, AI’s Most Famous Researcher, Warns Of ‘Existential Threat’ From AI. *Forbes.* Retrieved May 18, 2025, https://www.forbes.com/sites/craigsmith/2023/05/04/geoff-hinton-ais-most-famous-researcher-warns-of-existential-threat/

Williams, R., Won Park, H., & Breazeal, C. (2019). A is for Artificial Intelligence: The impact of Artificial Intelligence activities on young children’s perceptions of robots. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). Association for Computing Machinery, New York, NY, USA, Paper 447, 1–11. https://doi.org/10.1145/3290605.3300677

Wong, J. (2016, July 14). Wargaming in professional military education: A student's perspective. *Strategy Bridge*. Retrieved May 18, 2025, from https://thestrategybridge.org/the-bridge/2016/7/14/wargaming-in-professional-military-education-a-students-perspective

SENSORS, IMAGE RECOGNITION, AND THE ETHICAL CHALLENGES OF AI IN WARFARE

Jouko Vankka and Antti Rissanen

Department of Military Technology, National Defence University, Finland

Abstract This chapter analyzes the integration of sensor technology and Artificial Intelligence (AI) into modern warfare, with a particular focus on their role in the development of anti-tank weapons and the resulting ethical challenges. The perspective centers on intelligent anti-tank missiles, a development that underscores the need for proactive protection measures and is especially significant in asymmetric warfare, where weaker actors attempt to set the "rules of the game" by making the superior adversary's assets vulnerable. AI-driven improvements in data processing (e.g., addressing the flaws of the Johnson criteria) and target recognition (using deep learning and synthetic data) are crucial. However, the autonomous use of AI raises serious ethical questions regarding compliance with International Humanitarian Law (distinction and proportionality) and accountability. While AI can increase precision and reduce human casualties, it requires continuous human oversight and control to maintain ethical standards. The article emphasizes that asymmetric warfare is a broader conflict category, while guerrilla warfare is a common tactic within it, and both necessitate continuous strategic and ethical analysis.

The Battlefield and the Evolution of Anti-Tank Missiles

In modern warfare, technical innovation offers a path to superior success in combat situations. This success begins with better advance knowledge (situational awareness) and the ability to assess the utility of different courses of action in real-world conditions. Human tactical and strategic skill, utilizing both new resources and traditional, time-tested principles, must be developed to optimally exploit the new capabilities created by engineers to achieve better military performance.

In ground warfare, motorized and armored forces play a vital role on the battlefield. Conversely, mines and anti-tank missiles (ATGMs) pose a major threat.(Kennedy, 1978) Modern projectiles penetrate even the thickest armor, but equally significant is the intelligence of anti-tank missiles. Sensors observe the target area, and processor calculations identify the object. This allows the missile to track targets autonomously, reducing the need for operator guidance.(Radovanović et al., 2023)

ATGMs are designed to provide defensive firepower and support an offensive, making them critical in asymmetric warfare. The threat of precision-guided weapons necessitates proactive protective measures, such as active protection systems and enhanced situational awareness. Arms dealers have also supplied anti-tank missiles to resistance groups, challenging motorized forces operating in contested areas.

The development of anti-tank missiles continues across all technical dimensions. Utilizing data from newer generation sensors with Artificial Intelligence (AI) and machine learning algorithms significantly improves missile guidance and targeting capabilities. AI systems could autonomously analyze vast amounts of data to identify and exploit weaknesses in enemy armor formations, while simultaneously assessing the risk of false targets posed by civilians in the operational environment.

The EU's Perspective on AI Integration in Warfare: Ethical Challenges

The integration of AI into warfare requires a serious ethical framework regarding the implementation of accountability and compliance with the principles of international law. It's often assumed that AI systems make fully independent decisions in military operations. In such cases, it must be ensured that decisions are responsible and adhere to the guidelines of international law, the two key principles being distinction (separating combatants and civilians) and proportionality (balancing military actions with the threat).

Therefore, human choices, guidance, and supplementary oversight remain necessary to maintain ethical standards, even though AI can accelerate decision-making while increasing operational precision. It is believed that conflict escalation can be mitigated through better planning and the use of precision weapons.(Clapp, 2025)

Some ethics experts consider the use of robots in warfare unethical, fearing it may increase the frequency of wars and complicate the distinction between combatants and civilians. Conversely, robot soldiers could potentially reduce human casualties and adhere to doctrine without emotional interference. A critical debate also concerns whether robots should be permitted to make autonomous decisions when human lives are at stake. Experts emphasize adherence to International Humanitarian Law and demand comprehensive regulation to address the unique challenges of AI in military contexts (Jafariandehkordi, 2024).

Image Analysis in Warfare: Integrated Sensor Technology and AI

As part of a network-centric approach, modern warfare utilizes new types of sensors to provide a more realistic image of the battlefield. These systems combine various sensor types, such as electro-optical, infrared, and radar sensors from multiple platforms, including supporting mounts like drones and satellites. These systems employ advanced signal processing and Artificial Intelligence (AI) for data fusion. The goal is to improve situational awareness, accelerate decision-making, and refine target selection for weapon effects. Sensors are also becoming smaller while maintaining the ability to operate in challenging conditions.

In AI-driven analysis, algorithms or, more recently, trained neural networks assist in the real-time processing of sensor data to identify potential threats and separate signals from "noise." By producing visually more comprehensible information, they can reduce the workload of soldiers. Remotely operated distributed systems can filter or summarize their output before forwarding it. By shifting computing power closer to the "edge" (e.g., to drones or vehicles), dependence on service centers can be reduced, and processing needs in command centers can be lessened. (Mary et al., 2024)

Traditional Criteria for Imaging Sensors: The Johnson Criteria

Following World War II, Western countries focused on developing methods for long-range engagement and improved situational awareness. A particular focus was on night vision capability through technical devices. The limitations of observations made with optics in daylight were well known in the 1950s, as were the challenges of film size, sensitivity, or graininess in photographic intelligence, phenomena known since the 1930s.

There are a few techniques suitable for night vision, but in the '50s, there was no clear set of criteria for how finished products would perform in broader practical use. Instead of merely facilitating slow terrain movement in the dark or monitoring a surveillance area extending a maximum of one hundred meters, next-generation devices would need to enable more demanding observations or even autonomous targeting applications utilizing night vision sensors. Equally significant was the need for a standard to compare products from device developers against different operational needs.

Such a standard didn't exist in the '50s, although various ideas and experiments had been proposed to extend the model of basic optical devices to night vision. A norm, standard, and guiding document with testing methods were needed to define requirements as part of the procurement process.

Mentions of various experiments are found in the literature. The empirically formed Johnson Criteria gained traction partly due to its simple summary tables, which became rules of thumb. The Johnson Criteria describe the ability of human observers to perform various military target observation tasks using night vision technology. The method's baseline data was created by observing scale models of selected target groups under test conditions. The experiments, combined with a simple statistical analysis, were reported as categorized requirements for target observation thresholds. Today, modernized understandings of how the model works are available online, and its results have been extended almost as is to visible light observations.(Salvaggio, n.d.)

A typical generalization clarifies the everyday model: How many pixels would the use of the Johnson Criteria require for certain detection? The rhetorical answer would be: at least 2 pixels for detection, and classifying the target (human vs. vehicle) requires 8 pixels on the observed object.

Flaws and Current Status

The original intent was a simple and easily interpreted model, like the Johnson Criteria with its classification tables. This allowed for personnel training and streamlined requirements drafting. However, the version created in the '50s and its application methods involved several overlooked flaws. Updating or modernizing has been difficult because procurement activities, at least, must anticipate commercial conflicts that would arise from changes to the guiding documents. Partly for this reason, later perception models only supplemented the terminology.

The most interesting problem from the perspective of this article is the lack of modeling for the variability in human factor or operator performance. As early as 1969, Hemingway and Erickson estimated the range of individual performance difference variability to be 50%–97%. Another observation is that the original model experiments were conducted indoors at short distances in the standardized conditions of an aircraft hangar. Although the newest models attempt to account for the effect of weather on observations, the problem is that they cannot accurately predict target detection or maximal observation distance in extreme conditions. (Sjaardema et al., 2015)

Improvements in human-aided observation models have been seen by incorporating more variables into the model, such as the target's background, observation angle, and condition factors described by the noise ratio. Additionally, the significance of human variability should be assessed more broadly, especially through the stress factors of the battlefield environment. It is thus apparent that instead of mechanical, table-producing models, accurate probabilities of target detection in any situation can only be obtained by calculating the value on a case-by-case basis. Simultaneously, the aim must be to perform the more demanding observations based on high-resolution digital observation data produced by a sensor with sufficient resolution, rather than relying on the human individual. Algorithms and AI processing reinforcement are needed to filter this information. This view does not exclude the operator's responsibility to accept or reject the final assessment for action.

Machine Learning in Imaging

Images of modern military equipment and combat vehicles are available only in a limited quantity. Some data sets have clear restrictions either on usage or because they are priced as news material. Simultaneously, the quality of moving image data adapted to the operational environment may have a reduced resolution for the main target. Such constraints make acquiring sufficient image material for training deep learning models laborious. The need for good data through traditional target libraries is considered important, although more advanced AI is what truly enables deeper image analysis. (Rissanen et al., 2022) Better data and models not only improve the identification of correct targets but also allow for consideration of the correctness of the identification assessment. This problem of right vs. wrong and "false negative" and "false positive" is part of understanding the correct solution in engineering as well as in medical imaging and laboratory analysis-based diagnostics (Mohri et al., 2018).

Better Image?

One might think that in the training of self-driving cars, where authentic images are abundant, one could stick solely to those. However, it has been shown that utilizing synthetic images has yielded excellent results as baseline material for machine learning. Rapid and accurate recognition of various moving objects on the battlefield is a difficult task. The biggest challenges for a machine learning-based system on the battlefield are the lack of data and the high number of interference factors.

These challenges in the military context have also been studied at the National Defence University (MPKK) in Finland. We have developed deep learning model prototypes for the detection, segmentation, and classification of Armored Fighting Vehicles (AFV) from images. The model's utility in night vision tasks and its ability to handle difficult input, such as AFV decoys, have also been investigated. (Rissanen et al., 2022)

Different neural network training systems are studied for image recognition of military vehicles, variable start layer transfer training models and own convolutional neural networks training from scratch (Legendre & Vankka, 2020). Since, there is limited openly available military recordings, labeled social media images are used for training. In terms of transfer learning, a deep neural network proved to be a well-behaved neural network that adapted rapidly to the classification options studied with accuracy values on average of 88 % (Legendre & Vankka, 2020). While deep neural networks can be used to identify vehicles in the military domain, it introduces uncharted risks into military operations. Two of the significant operational risks are the black box nature of AI decision-making (unexplainable outputs) and lack of good quality data impacting efficiency and accuracy of deep neural networks algorithms, which can lead to bias (Legendre & Vankka, 2020).

Asymmetric Warfare and the Rules of Engagement

Conceptually, asymmetric warfare refers to an uneven balance of power. (Sobelman, 2025) Correspondingly, guerrilla warfare is a common response to this situation. As a concept, it creates more case-specific imagery than it provides tools for analysis. Therefore, the term guerrilla warfare should be considered a tactic that manifests in asymmetric combat techniques, focusing on exploiting the enemy's weaknesses (e.g., logistical vulnerabilities, slow movement, reliance on terrain). In other words, the goal is to avoid the strengths of the better-resourced adversary on a case-by-case basis.(Smith, 2003)

Not all asymmetric conflicts involve guerrilla warfare (e.g., a cyber conflict between a state and an organized hacker group is asymmetric but not guerrilla warfare), but practically all guerrilla campaigns are a form of asymmetric warfare. The key difference between guerrilla warfare and asymmetric warfare lies in their scope: asymmetric warfare is a broader category of conflict, while guerrilla warfare is a specific strategy or set of tactics often employed within it.

Are there Rules in Asymmetric Warfare?

Asymmetric warfare, where a militarily superior state faces a relatively weaker, often non-state actor, is a key feature of modern conflicts. In such situations, the weaker party's ability to force its stronger adversary to accept restrictive "rules of the game" is a critical strategic question. As Daniel Sobelman analyzed, this concept has become an organizing principle in the operations of the Iran-led "Axis of Resistance" in the Middle East.

Why would a militarily superior state adhere to restrictive rules set by its weaker adversary? The answer lies in the mutual deterrence created by a balance of vulnerabilities. The weaker party seeks to create a context where the superior party's overwhelming use of force triggers countermeasures targeting its critical vulnerabilities. This creates a "cost threshold" that makes the use of superiority politically or humanly too expensive. Sobelman notes that the success of the weaker actor in setting the rules varies considerably. In asymmetric warfare, the "rules of the game" depend on the weaker party's ability to change the fundamental dynamics of the conflict by making the stronger party's civilian or military targets vulnerable. Part of this involves the development of rocket and drone arsenals. Extreme, sometimes politically risky solutions include the use of human shields and urban combat with revised guerrilla warfare principles.(Sobelman, 2025)

Examples of the Earliest Target-Recognizing Systems

The development of several lighter, guided weapon systems began in the 1980s. In the West, the need for such systems was justified by the quantitative superiority of the Warsaw Pact's armored forces, which needed to be countered with broader denial capabilities. Among this group, the most interesting were anti-tank weapons designed to be launched as specialized ammunition from standard service mortars. Since mortars operate on the principle of indirect fire, these rounds were also designed to be fired in a salvo, beyond the line of sight. (A & D Market Reports, 2024) After launch, the guided projectile typically begins a ballistic trajectory until it reaches its apogee. In the case of light systems, the distance from the apex to the target is perhaps 2-3 km and takes at most a few tens of seconds.

Radar and Doppler Effect Seekers

Radar-based seekers primarily look for moving targets, and if none are present, a target must be found to home in on (e.g., the Merlin Guided Anti-Armor Projectile)(British Aerospace, 1996). In fighting on the large open areas of Poland, it seems clear that an area of about 250 meters in diameter, selected based on an accurate situational picture, would primarily contain combat vehicles. Although a radar signal can find moving targets or strong echoes from stationary metallic objects, it does not directly distinguish between civilian and military targets. In such a solution, the firing order issuer must rely solely on their knowledge of the battlefield dynamics. The weapon system does not assist in the ethical assessment of the action in the final stage. The challenge with a moving target is the correct understanding of the multiple parallel motion components affecting the signal. On the other hand, the selection between multiple targets is comparable to choosing the correct move in computer chess. The algorithm itself, which helps the computer reject wrong moves from further calculation, is easily implemented even on a simpler device platform (cf. earlier travel chess game machines).

Infrared Seekers

The Swedish project STRIX Pansarsprängvinggranat m/94 STRIX (Owl) is designed to be fired from a 120 mm mortar. The system's seeker utilizes the target's infrared radiation. The target stands out from its background due to the heat radiation generated by the vehicle. Such structural components include exhaust pipes, the engine, wheels, or tracks. The STRIX mortar round has a range of about 4.5 km, but with a more advanced engine, its maximum range would be up to 7 km. (Think Defence, 2023) Based on its simpler and more limited guidance and detector resolution, the trajectory must be carefully pre-calculated when searching for a target within an area of about 50 meters in diameter. The system was manufactured and introduced into use (at a cost of about €30,000 per round). Marketing material mentions the possibility of customizing the seeker's target library to match the acquiring state's perceived threats.

Considering the development start in 1983 and the introduction in 1994, with their detector technology and processing capability, it can be assumed that the primary task was to find the correct targets and distinguish them from potential decoys or similar deception solutions. It is not credible that such systems, which quickly proceed from detection to impact, could weigh whether the detection was made on a genuine civilian target. However, the need to shoot more accurately and to assess the round's capability and the target before final lock-on provides an indication of the added value brought by algorithmic logic and training data regarding the correct target.

PGMM Project

After the Cold War, the US Army's Precision Guided Mortar Munition (PGMM) project began in 1999, which returned to short-range, but still beyond line-of-sight, engagement using simple mortar-based solutions. It was motivated by observations from the Chechen War with the assertion that "a single precisely guided mortar round can more reliably destroy a confined, small protected target than a large number of unguided rounds." Modifications to already-in-production anti-tank weapons appeared to meet the required performance. (Malejko et al., 2008)

Some Successful Solutions in Service

While STRIX achieved limited use and the other imaging beyond-line-of-sight systems mentioned above remained at the experimental stage, the following low-trajectory devices represent the extremes in their group:

- **FGM-148 Javelin** is a US-made missile consisting of a separate Command Launch Unit (CLU) with an infrared sight and the missile itself with its disposable launch tube. It is equipped with an Infrared Imaging System (I2R) and a fire-and-forget missile, with a range of 2,000–2,500 meters (estimated cost $248,000).(Sherman, 1999) (see also Wikipedia)
- **APILAS** (Armour-Piercing Infantry Light Arm System) is a portable, disposable anti-tank weapon manufactured in France. Its range is 300–500 meters. In the French army, it is classified as a "traumatic weapon" due to its blast and noise. The individual combatant must operate close to the target in a direct-fire manner (estimated cost is €2,000). ("APILAS," 2025)

Conclusion

Due to modern urban area type of battlefield, we need new type of capability, i.e. correctly identifying the target, and almost always civilians. Another issue is the need to reducing own battlefield casualties. So, if the loss of a combatant (killed or permanently disabled) is valued in monetary terms, the cost-effectiveness of the APILAS system, for example, diminishes. Conversely, it is worth designing weapon systems that can engage targets more precisely and from farther away than even the relatively expensive Javelin. This means that development work needs to be focused on improving situational awareness by creating sensors that see farther and where integrated filtering can process sensor data faster. By demanding resilience for all sensing and data transfer methods, it can be assumed that sensors will operate effectively even in jammed environments. While armored vehicle recognition has been discussed, the future will also see the development of facial recognition sensors to identify individuals remotely for improved security and other operations (Girirajan et al., 2025). This would reduce civilian losses even in deeply asymmetrical battle.

Technological advancement, particularly the proliferation of precision-guided weapons and drones, will impact the balance of vulnerability. As Ukraine's ability to strike deep into Russian territory with drones demonstrates, new technology can offer the weaker party surprising means to create and deepen their adversary's vulnerabilities. This dynamic change necessitates continuous analysis of the conditions under which military superiority can still be restrained.

References

A & D Market Reports. (2024, February 7). *The Evolution Of Anti-Tank Missiles: A Comprehensive Overview - Aviation and Defense Market Reports*. https://aviationanddefensemarketreports.com/the-evolution-of-anti-tank-missiles-a-comprehensive-overview/

APILAS. (2025). In *Wikipedia* (Vol. 2023). https://en.wikipedia.org/w/index.php?title=APILAS&oldid=1306210860

British Aerospace. (1996). *Merlin*. https://www.scribd.com/document/610017297/Merlin-Guided-Anti-Armor-Projectile

Clapp, S. (2025, April 29). *Reinforcing Europe's defence industry, European Parliament* [BRIEFING EPRS | European Parliamentary Research Service]. European Parliament. https://www.europarl.eu- ropa.eu/thinktank/de/document/EPRS_BRI (2023)749805

Girirajan, S., Sandhia, G. K., & Kumar, G. M. (2025). Detecting and Tracking Multiple Objects (DTMO) Using Machine Learning. *In Sustainable Development, Innovation and Green Technology (ICASDIGT-2024)*, 125.

Jafariandehkordi, M. (2024). *The AI Battlefield: Legal Challenges of Autonomous Weapon Systems under International Humanitarian Law*. https://www.doria.fi/handle/10024/189724

Kennedy, R. (1978). *Precision ATGM's and NATO Defense.* https://apps.dtic.mil/sti/html/tr/ADA063723/

Legendre, D., & Vankka, J. (2020, October). Military vehicle recognition with different image machine learning techniques. In *International Conference on Information and Software Technologies* (pp. 220-242). Cham: Springer International Publishing.

Malejko, G., Burke, P. J., Dohrn, R., & Owens, J. S. (2008). *Jet interaction effect on the precision guided mortar munition (pgmm).* https://apps.dtic.mil/sti/html/tr/ADA504659/

Mary, P. A., Sharma, A., Dekka, S., Gowda, V. D., & Singh, M. (2024). Deep Learning Approaches for Real-Time Data Analytics in IoT Sensor Networks. In V. S. Rathore, V. Piuri, R. Babo, & K. S (Eds.), *Universal Threats in Expert Applications and Solutions* (Vol. 1007, pp. 239–248). Springer Nature Singapore. https://doi.org/10.1007/978-981-97-5146-4_21

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning.* MIT press. https://books.google.com/books?hl=en&lr=&id=dWB9DwAAQBAJ&oi=fnd&pg=PR5&dq=(Mohri,+Rostamizadeh+Talwalkar+2018)&ots=Az-tQTSyZm5&sig=z1mkAHhjnaLTQDmv2OwxJJdKKks

Radovanović, M., Petrovski, A., Smileski, S., & Jokić, Ž. (2023). Analysis of the development of five generation of anti-armor missile systems. *J. Scientific Technical Review*, *73*(1), 26–37.

Rissanen, V., Toivonen, E., Lagashkin, R., Saastamoinen, K., Rissanen, A., & Vankka, J. (2022). Instance segmentation and classification of armoured fighting vehicles. *2022 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, 1–8. https://ieeexplore.ieee.org/abstract/document/9855933/

Salvaggio. (n.d.). *Original papers*. Retrieved October 20, 2025, from https://home.cis.rit.edu/~cnspci/references/johnson1958.pdf

Sherman, R. (1999). *Javelin Antitank Missile*. https://man.fas.org/dod-101/sys/land/javelin.htm

Sjaardema, T. A., Smith, C. S., & Birch, G. C. (2015). *History and Evolution of the Johnson Criteria.* (No. SANDIA REPORT SAND2015-6368). Sandia National Lab.(SNL-NM), Albuquerque, NM (United States). https://www.osti.gov/biblio/1222446

Smith, M. L. (2003). Guerrillas in the mist: Reassessing strategy and low intensity warfare. *Review of International Studies*, *29*(1), 19–37.

Sobelman, D. (2025). Rules of the Game in Asymmetric Conflicts. *Security Studies*, *34*(2), 261–291.

Think Defence. (2023, May 30). *Anti-Tank Mortar Bombs*. Complex Weapons. https://www.thinkdefence.co.uk/2023/05/anti-tank-mortar-bombs-merlin-strix-and-bussard/

EPISTEMIC JUSTICE AND ARTIFICIAL INTELLIGENCE: GENDERED KNOWLEDGE, ALGORITHMIC BIAS, AND THE POLITICS OF KNOWING

Anita Dremel

Department of Sociology, University of Osijek, Croatia

Abstract Artificial Intelligence (AI) systems increasingly function as epistemic infrastructures that shape how knowledge is produced, authorized, and contested. While dominant debates in AI ethics focus on bias, fairness, and transparency, this chapter argues that such approaches remain insufficient without a sustained engagement with epistemic justice. Drawing on Miranda Fricker's concepts of testimonial and hermeneutical injustice, extended through the work of José Medina, Kristie Dotson, and feminist standpoint theory, the chapter conceptualizes AI as an epistemic mediator that redistributes credibility, interpretive resources, and epistemic authority in ways that disproportionately disadvantage women and other marginalized groups. Through illustrative cases in recruitment, healthcare, and facial recognition, it shows how algorithmic systems can silence lived experience, misrecognize social realities, and render certain forms of knowledge epistemically invisible. The chapter further argues that feminist epistemology, with its emphasis on situated knowledge and "strong objectivity," offers not merely a moral critique but a necessary epistemic framework for rethinking AI design and governance. It concludes by outlining principles for epistemically just AI that move beyond technical bias mitigation toward participatory, transparent, and reflexive knowledge production. Reframing AI ethics as a politics of knowing, the chapter positions epistemic justice as a central criterion for evaluating the legitimacy of AI-mediated knowledge.

Keywords: Epistemic justice, artificial intelligence ethics, algorithmic authority, feminist epistemology, gendered knowledge.

Introduction

Artificial Intelligence (AI) increasingly mediates how social reality is known, interpreted, and governed. From hiring decisions and medical diagnostics to content moderation and knowledge retrieval, AI systems now play a decisive role in determining what counts as relevant information, whose contributions are deemed credible, and how decisions are justified. These systems are often presented as neutral, objective, or merely technical instruments. Yet a growing body of critical scholarship demonstrates that algorithmic systems are deeply embedded in existing social relations and routinely reproduce historical inequalities and power asymmetries (cf. Bowker & Star, 2000; Burrell, 2016; Mohamed et al., 2020; Mittelstadt, 2019).

In this chapter, AI is understood not simply as a set of computational tools but as part of the epistemic infrastructure through which knowledge is produced, authorized, and contested. Particular attention is given to large language models (LLMs), such as GPT-based systems, which increasingly shape associations, representations, and patterns of speech that implicitly define what appears as "neutral" information and which experiences are foregrounded or marginalized. Empirical studies have shown that statistical language models reproduce and amplify social hierarchies embedded in language itself, producing gendered and racialized associations that mirror human biases (Caliskan, Bryson, & Narayanan, 2017). Recent investigations, including UNESCO's (2024) study on gender bias in large language models, show that stereotypical associations and patterned exclusions persist even in state-of-the-art systems. These findings suggest that AI does not merely reflect existing data but actively participates in the production of social meaning, reinforcing hierarchies of knowledge and credibility.

While contemporary debates in AI ethics have focused primarily on issues of bias, fairness, transparency, and accountability, this chapter argues that such approaches remain insufficient without a sustained engagement with epistemic justice. Bias-oriented frameworks tend to conceptualize harm in terms of inaccurate or unfair outcomes, often leaving unexamined the deeper question of how epistemic authority itself is reorganized through algorithmic systems. Epistemic justice, by contrast, directs attention to the conditions under which individuals and groups are recognized as knowers, granted credibility, and provided with the interpretive resources necessary to make sense of their social experiences.

Recent scholarship has begun to explicitly conceptualize artificial intelligence as a site of epistemic injustice, rather than merely a source of biased outcomes. Scholars have shown that generative and predictive AI systems can undermine epistemic agency, distort collective knowledge environments, and systematically privilege dominant perspectives while marginalizing others (Kay, Kasirzadeh & Mohamed, 2025; Johnson, 2021). From decolonial and critical perspectives, AI has also been analyzed as an extension of historical epistemic domination, embedding Western and technoscientific epistemologies into global infrastructures of knowledge production (Mhlambi, 2020; Mohamed et al., 2020). This emerging literature supports the claim advanced here that AI ethics requires not only technical or procedural remedies, but a sustained engagement with epistemic justice as a framework for evaluating the legitimacy of AI-mediated knowledge.

Following Miranda Fricker's (2007) influential account, epistemic injustice occurs when individuals are wronged in their capacity as knowers, most notably through testimonial injustice – when prejudice leads to a deflated attribution of credibility – and hermeneutical injustice – when structural gaps in collective interpretive resources disadvantage certain groups. Subsequent work has extended this framework to capture systemic and structural forms of epistemic oppression (Dotson, 2014) and to emphasize the need for epistemic activism and reflexive institutional practices (Medina, 2013). These perspectives are particularly relevant for analyzing AI systems, which institutionalize epistemic judgments at scale and often render them difficult to contest.

This chapter brings epistemic justice into dialogue with feminist epistemology, especially feminist standpoint theory, which has long challenged claims of neutrality and emphasized the situatedness of knowledge. Feminist scholars have argued that starting inquiry from marginalized standpoints can produce "strong objectivity" by revealing blind spots embedded in dominant knowledge practices (Harding, 1993; Collins, 2000). Applied to AI, this insight challenges the aspiration to a disembodied view from nowhere and instead highlights the epistemic consequences of whose experiences are encoded in data, models, and design choices.

The central claim advanced here is that AI systems function as epistemic mediators that redistribute epistemic authority in ways that can generate distinct forms of epistemic injustice. By shaping which voices are heard, which experiences are legible, and which interpretations are normalized, algorithmic systems can silence marginalized knowers, misinterpret lived realities, or render them epistemically invisible. These harms are not accidental but structural, arising from the delegation of epistemic judgment to socio-technical systems that reflect existing power relations.

The chapter addresses three guiding questions:

- How do AI systems contribute to testimonial and hermeneutical injustice, particularly in relation to gender and intersecting forms of marginalization?
- In what ways can feminist standpoint theory and related epistemological approaches help identify and mitigate algorithmically produced epistemic harms?
- What principles might guide the design and governance of epistemically just AI systems?

The chapter proceeds in five sections. Following this introduction, the conceptual framework outlines key contributions from epistemic injustice theory and feminist epistemology. The third section analyzes AI as an epistemic mediator, focusing on issues of epistemic authority, opacity, and algorithmic testimony. The fourth section examines empirical cases of algorithmic bias in recruitment, facial recognition, and healthcare, highlighting their gendered and intersectional epistemic dimensions. The final section articulates normative principles for epistemically just AI, emphasizing participatory design, epistemic access, reflexivity, and pluralism. The conclusion reflects on the broader implications of reframing AI ethics as a politics of knowing.

By situating AI within social epistemology and feminist theory, this chapter aims to move beyond technical accounts of bias and fairness toward a deeper analysis of how knowledge itself is shaped, constrained, and contested in algorithmically mediated societies.

Conceptual Framework – Epistemic Justice, Feminist Epistemology, and AI

This chapter approaches epistemic justice not only as a matter of interpersonal credibility and interpretation, but as a structural feature of contemporary knowledge infrastructures, including algorithmic systems. As Artificial Intelligence increasingly mediates how knowledge is produced, classified, and legitimized, questions of epistemic justice become inseparable from questions of power, authority, and institutional design.

The concept of epistemic justice, as formulated by Fricker (2007), provides a foundational framework for analyzing how individuals and groups can be wronged in their capacity as knowers. Fricker distinguishes between two primary forms of epistemic injustice: testimonial injustice, which occurs when prejudice leads to a deflated attribution of credibility to a speaker, and hermeneutical injustice, which arises when structural gaps in collective interpretive resources disadvantage certain groups in making sense of their social experiences. Both forms are highly relevant for understanding how AI systems evaluate information, represent social categories, and shape interpretive possibilities.

However, as subsequent scholarship has emphasized, epistemic injustice cannot be fully understood at the level of individual interaction alone. Dotson (2014) extends Fricker's account by introducing the concept of epistemic oppression, drawing attention to persistent, systemic patterns of exclusion from knowledge production and meaning-making practices. From this perspective, epistemic injustice is not merely episodic but embedded in institutional arrangements that determine whose knowledge is solicited, recognized, and sustained over time. Medina (2013) further develops this insight through the notion of epistemic activism, arguing that resisting epistemic injustice requires reflexive, collective practices that cultivate epistemic humility, responsiveness, and openness to marginalized perspectives.

These extensions are crucial for analyzing AI systems, which institutionalize epistemic judgments at scale. Algorithmic systems do not simply transmit individual prejudices; they formalize, automate, and normalize particular epistemic assumptions through data selection, model design, and evaluation criteria. As a result, epistemic injustice in AI contexts often operates without a clearly identifiable hearer or interlocutor, complicating traditional models of epistemic responsibility while intensifying the effects of exclusion.

From a science and technology studies perspective, this institutionalization of epistemic judgment resonates with classic analyses of classification and infrastructure, which show how seemingly technical systems stabilize social hierarchies by rendering certain categories visible and others invisible (Bowker & Star, 2000). AI systems can thus be understood as classificatory infrastructures that actively shape epistemic visibility and legitimacy, rather than neutral channels for information processing (Jasanoff, 2011).

Feminist epistemology offers critical resources for addressing these challenges. Feminist scholars have long emphasized that knowledge is socially situated and that claims to neutrality often obscure relations of power (Haraway, 1988; Code, 1991). Standpoint theorists such as Harding (1993) and Collins (2000) argue that beginning inquiry from the lives and experiences of marginalized groups can produce what Harding terms strong objectivity – a form of knowledge that is more reflexive, less partial, and more socially accountable precisely because it foregrounds positions historically excluded from dominant epistemic frameworks.

Applied to AI, feminist standpoint theory challenges the aspiration to a disembodied, universal perspective often embedded in algorithmic design. Rather than treating bias as a deviation from an otherwise neutral baseline, standpoint theory reframes bias as an indicator of whose perspectives have been systematically privileged or excluded in the construction of knowledge systems. This insight shifts the focus from technical correction toward epistemic reorientation: whose experiences count as data, whose categories structure interpretation, and whose interests shape design priorities.

Recent work in AI ethics and critical data studies builds on these feminist insights by emphasizing participatory and inclusive approaches to system design (Birhane et al., 2022; Costanza-Chock, 2020). These approaches argue that affected communities should play an active role not only in evaluating outcomes but in defining problems, shaping datasets, and determining what counts as a satisfactory explanation. From the perspective of epistemic justice, such participation is not merely procedurally fair; it is epistemically necessary for reducing structural blind spots and expanding collective interpretive resources.

Taken together, epistemic injustice theory and feminist epistemology provide a framework for understanding AI systems as epistemically consequential social institutions rather than neutral technical tools. They direct attention to the ways in which credibility, intelligibility, and epistemic authority are distributed through socio-technical arrangements, often in ways that disadvantage women and other marginalized groups.

When epistemic practices are delegated to algorithmic systems, the conditions under which knowledge claims are produced, evaluated, and trusted are fundamentally transformed. Understanding these transformations requires moving beyond interpersonal models of epistemic exchange toward an analysis of epistemic mediation and institutionalized authority – a task taken up in the following section, which examines AI as an epistemic mediator exercising authority without accountability.

AI and Epistemic Agency

Recent debates in the philosophy and ethics of artificial intelligence increasingly ask whether AI systems should be understood as epistemic agents in their own right or merely as sophisticated tools that extend human cognition (Floridi & Sanders, 2004; Rini, 2020). While AI systems clearly lack consciousness, intentionality, and moral responsibility in the human sense, this chapter argues that framing the problem exclusively in terms of agency risks obscuring a more pressing epistemic issue: AI systems exercise epistemic authority without being epistemic agents, and it is precisely this asymmetry that generates distinctive forms of epistemic injustice.

Epistemic authority refers to the socially conferred power to define what counts as knowledge, evidence, or justified belief (Zagzebski, 2012). Traditionally, such authority has been vested in human experts, institutions, and professions whose legitimacy rests – at least in principle – on accountability, contestability, and the possibility of epistemic challenge. In algorithmic systems, however, epistemic authority is increasingly delegated to socio-technical infrastructures whose outputs are treated as authoritative despite their opacity, scale, and distance from affected subjects. AI systems thus come to function as epistemic gatekeepers: they filter information, rank relevance, predict risk, and recommend action, often without offering reasons that can be meaningfully interrogated by those subject to their judgments.

From the perspective of epistemic justice, this shift is consequential. Fricker's (2007) account of testimonial injustice presupposes a recognizably social relation between speaker and hearer, in which prejudice leads to a deflated attribution of credibility. In algorithmic contexts, by contrast, credibility assessments are formalized, automated, and institutionalized, operating without a human hearer who can be directly addressed, challenged, or persuaded. When an AI-driven system discounts certain forms of testimony – such as CVs signaling female gender, patient-reported symptoms, or linguistic patterns associated with racialized communities – the resulting epistemic harm is not reducible to individual prejudice. Instead, it reflects a structural delegation of epistemic judgment to systems that reproduce historical patterns of exclusion while obscuring the mechanisms through which credibility is allocated.

The contribution of this chapter lies in shifting the analytical focus from questions of algorithmic bias or epistemic agency to the problem of epistemic authority without accountability. While much of the existing literature critiques AI for producing unfair or discriminatory outcomes, fewer analyses examine how AI systems reorganize the conditions under which epistemic authority is exercised and contested. By extending Fricker's framework beyond interpersonal contexts, this chapter conceptualizes algorithmic systems as institutionalized epistemic mediators whose authority is socially consequential precisely because it is opaque, scalable, and difficult to challenge. This reframing clarifies why prevailing fairness and accountability frameworks, while necessary, remain insufficient for addressing the epistemic harms produced by AI systems. Recent work on contributive epistemic injustice in AI-mediated workplaces further supports this analysis, showing how workers are excluded not only from decision-making but from participation in the epistemic practices that define expertise and legitimacy (Innocenti, 2025 preprint).

This transformation also complicates the notion of hermeneutical injustice. Hermeneutical injustice, in Fricker's formulation, arises when collective interpretive resources are structured in ways that disadvantage certain groups, leaving them unable to render their experiences intelligible. AI systems intensify this problem by operationalizing dominant interpretive frameworks at scale. Because machine learning models are trained on historically sedimented data, they tend to normalize majority experiences and dominant categories, while marginal or emergent forms of social meaning remain underrepresented or unintelligible. The result is not merely a lack of representation but an algorithmic freezing of interpretive horizons, in which certain experiences systematically fail to register as meaningful inputs at all.

Importantly, these epistemic harms are compounded by the problem of opacity. The so-called "black box" character of many AI systems does not simply limit technical understanding; it restricts epistemic access. As Humphreys (2009) and Burrell (2016) have shown, opacity undermines the ability of affected individuals to evaluate, contest, or contextualize algorithmic outputs. From an epistemic justice perspective, opacity thus functions as a mechanism of exclusion: it prevents marginalized groups from participating in the practices of sense-making and justification that govern decisions affecting their lives. Hancox-Li and Kumar (2021) rightly note that explainability is not a neutral technical goal but an epistemic value laden with assumptions about whose understanding matters and whose questions are considered legitimate.

The delegation of epistemic authority to AI systems also reshapes relations of epistemic dependence. Human knowledge has always relied on testimonial trust (Hardwig, 1985), but algorithmic systems introduce a novel form of dependence in which trust is displaced from identifiable knowers to institutionalized procedures. As Rini (2020) argues, AI-generated outputs increasingly function as knowledge claims despite the absence of reasons or justificatory structures that resemble human testimony, raising new challenges for epistemic responsibility and trust. When such systems are embedded in high-stakes domains – credit scoring, hiring, welfare allocation, medical diagnosis – this dependence can become compulsory rather than voluntary. Those subject to algorithmic judgment often lack both the epistemic resources and the institutional standing to challenge its authority, a condition that disproportionately affects already marginalized groups.

From the standpoint of epistemic justice, the central problem is therefore not whether AI systems are epistemic agents, but how epistemic authority is exercised, distributed, and insulated from challenge in algorithmically mediated contexts. AI systems mediate knowledge claims without bearing epistemic responsibility, while the humans and institutions behind them are often shielded by technical complexity and organizational distance. This produces what can be described as a responsibility gap in epistemic governance: harms occur at the level of knowing, but accountability remains diffuse and elusive.

Feminist epistemology provides crucial resources for analyzing this configuration. By emphasizing the situatedness of knowledge and the epistemic advantage of marginalized standpoints, feminist scholars have long shown that claims of neutrality often mask relations of domination (Haraway, 1988; Harding, 1993). Applied to AI, this insight reveals that algorithmic authority is not neutral but grounded in particular social locations, values, and assumptions – often those of already powerful actors. When AI systems are treated as objective arbiters of truth, they can effectively silence alternative ways of knowing and foreclose epistemic contestation.

Understanding AI as an epistemic mediator rather than an autonomous agent thus allows us to extend, rather than abandon, Fricker's framework. Testimonial and hermeneutical injustices persist in algorithmic contexts, but they do so in institutionalized and scalable forms that exceed interpersonal interaction. These injustices are enacted not through explicit disbelief or interpretive gaps alone, but through automated credibility assignments, normalized categories, and opaque decision-making processes that systematically disadvantage certain knowers.

This analysis has direct implications for the empirical cases examined in the next section. In recruitment, facial recognition, and healthcare, algorithmic systems do not merely make biased decisions; they reorganize epistemic relations by determining whose testimony is legible, whose experiences count as data, and whose interpretations are recognized as valid. The harms that result are therefore not only distributive or procedural but fundamentally epistemic. Recognizing this shift is essential for understanding why technical fixes alone are insufficient – and why epistemic justice must be central to the ethical and political evaluation of AI systems.

Algorithmic Authority in Practice: Gendered and Intersectional Epistemic Harms

The purpose of these cases is not to introduce new empirical evidence, but to demonstrate how epistemic injustice is structurally produced when epistemic authority is delegated to algorithmic systems.

The preceding sections conceptualized AI systems as epistemic mediators that exercise authority over knowledge claims without bearing epistemic accountability. This section operationalizes that argument through a set of empirical cases that demonstrate how algorithmic systems reorganize epistemic relations in practice. Rather than treating these cases as examples of technical bias alone, the analysis foregrounds the epistemic harms produced when credibility, intelligibility, and epistemic standing are delegated to automated systems. Across domains, these harms disproportionately affect women and other marginalized groups, not incidentally but structurally.

Recruitment Algorithms: Automated Credibility and Testimonial Injustice

Algorithmic systems used in recruitment and hiring provide a clear illustration of how epistemic authority is exercised through automated credibility assessments. When hiring algorithms rank candidates, filter applications, or predict future performance, they do more than optimize selection processes; they determine whose self-presentations, qualifications, and professional narratives are treated as credible evidence.

The well-documented case of Amazon's experimental hiring algorithm, which systematically downgraded CVs containing indicators of female gender, exemplifies this dynamic. Trained on historical data reflecting a male-dominated workforce, the system learned to associate markers of women's participation with lower suitability. From the perspective of epistemic justice, this constitutes a form of testimonial injustice, but one that differs from its interpersonal counterpart. Here, credibility is not deflated by a prejudiced hearer but by an institutionalized procedure that formalizes past exclusions and renders them operational at scale. This dynamic became publicly visible when Amazon abandoned an internal hiring algorithm after discovering that it systematically downgraded CVs containing indicators of female gender (Dastin, 2018).

Crucially, the algorithmic nature of this judgment forecloses epistemic contestation. Applicants are neither heard nor disbelieved in any dialogical sense; their testimony is filtered out before it can be recognized as such. The harm lies not only in exclusion from employment opportunities but in the denial of epistemic standing as reliable narrators of one's own competence. This illustrates how algorithmic systems can enact testimonial injustice without a hearer, transforming credibility into a technical parameter insulated from challenge.

Facial Recognition Systems: Hermeneutical Injustice and Epistemic Invisibility

Facial recognition technologies reveal a complementary epistemic harm: not the deflation of credibility, but the failure of intelligibility. Research has consistently shown that such systems perform significantly worse on women, particularly racialized women, than on light-skinned men. Error rates exceeding 30% for dark-skinned women, compared to negligible rates for dominant groups, indicate more than technical inaccuracy.

From the standpoint of epistemic justice, these systems exemplify hermeneutical injustice in an algorithmic form. The training data and classificatory schemes underlying facial recognition technologies encode dominant assumptions about what a "face" looks like, rendering certain bodies systematically harder to recognize. As a result, the lived realities of those at the intersection of gender and racial marginalization become epistemically illegible to the system.

This form of epistemic harm is particularly severe because it produces invisibility rather than misrecognition alone. Individuals are not merely misunderstood; they are insufficiently represented in the interpretive framework itself. The epistemic authority of facial recognition systems – often deployed in policing, surveillance, and identity verification – means that such invisibility carries material consequences, while remaining difficult to contest due to the opacity and institutional embedding of the technology.

Healthcare Algorithms: Data Absence, Hermeneutical Gaps, and Epistemic Dependence

Healthcare provides a further domain in which algorithmic systems reproduce and intensify epistemic injustice. Long-standing data gaps regarding women's bodies, symptoms, and reproductive health are increasingly encoded into AI-driven diagnostic and predictive tools. Algorithms trained predominantly on male populations have been shown to underdiagnose women in areas such as cardiovascular risk assessment, despite women's self-reported symptoms. These omissions reflect what Criado-Perez (2019) describes as systematic "data gaps," in which women's bodies and experiences are rendered epistemically invisible within ostensibly neutral knowledge systems.

This constitutes a paradigmatic case of hermeneutical injustice, where collective interpretive resources fail to adequately capture certain experiences. In algorithmic healthcare, however, the injustice is intensified by epistemic dependence: clinicians and patients alike are encouraged, or required, to rely on algorithmic outputs as authoritative. When these outputs systematically misinterpret or overlook women's embodied experiences, affected individuals face a dual epistemic harm: their testimony is discounted, and the conceptual resources necessary to interpret their symptoms are unavailable.

Importantly, these failures are often framed as technical limitations rather than epistemic injustices. Yet from a feminist epistemological perspective, the absence of women's experiences from medical datasets reflects historical patterns of exclusion in knowledge production. AI systems thus become mechanisms through which androcentric epistemologies are stabilized and extended, rather than corrected.

From Bias to Epistemic Reorganization

Across these cases, a common pattern emerges. Algorithmic systems do not simply make biased decisions; they reorganize epistemic relations by determining whose testimony is legible, whose experiences are intelligible, and whose knowledge claims are granted authority. These systems enact testimonial and hermeneutical injustices in forms that are institutionalized, scalable, and difficult to contest, particularly for those already marginalized within social hierarchies.

What unites recruitment algorithms, facial recognition systems, and healthcare AI is not merely unequal outcomes, but the delegation of epistemic judgment to systems that operate without reciprocal accountability. The harms produced are therefore not adequately addressed through technical fixes alone, such as adjusting datasets or optimizing accuracy metrics. Rather, they point to a deeper epistemic problem: the concentration of epistemic authority in socio-technical systems that reflect dominant standpoints while obscuring their normative foundations.

Implications for Epistemically Just AI

These case studies underscore why epistemic justice must be central to the evaluation of AI systems. Feminist standpoint theory offers a critical lens for rethinking AI design by insisting that marginalized perspectives are not supplementary but epistemically indispensable. Incorporating these standpoints is not merely a matter of representation; it is a means of expanding interpretive resources and redistributing epistemic authority.

Understanding algorithmic harms as epistemic injustices reframes the task of AI ethics. The question is no longer only how to reduce bias, but how to design systems that recognize diverse knowers, enable epistemic contestation, and avoid freezing dominant interpretations into technical infrastructure. This shift provides the normative grounding for the principles of epistemically just AI developed in the following section.

Towards Epistemically Just AI: Principles for Reconfiguring Epistemic Authority

The preceding analysis has shown that AI systems function as epistemic infrastructures that redistribute credibility, intelligibility, and epistemic authority in systematic ways. When algorithmic systems discount testimony, render experiences unintelligible, or foreclose epistemic contestation, they produce harms that are not merely distributive or procedural but fundamentally epistemic. If epistemic justice concerns the fair distribution of epistemic goods, then designing and governing AI systems requires more than technical bias mitigation; it demands a reconfiguration of how epistemic authority is exercised and held accountable.

This section articulates a set of principles for epistemically just AI, derived directly from the forms of epistemic injustice identified in earlier sections. These principles are not presented as exhaustive or universally applicable rules, but as normative orientations that respond to the structural epistemic harms produced by algorithmic mediation.

Inclusion of Diverse Epistemic Standpoints

A central insight of feminist epistemology is that knowledge production is socially situated and that dominant epistemic frameworks routinely reflect the perspectives of already privileged groups. From this standpoint, the underrepresentation of women and other marginalized groups in AI systems is not simply a diversity problem but an epistemic one: it limits the interpretive resources available to society and stabilizes partial perspectives as universal.

An epistemically just AI must therefore actively incorporate diverse epistemic standpoints throughout the lifecycle of system development, including problem formulation, data collection, model training, and evaluation. This principle goes beyond representational inclusion; it requires recognizing marginalized perspectives as sources of epistemic insight rather than noise or deviation. In this sense, feminist standpoint theory reframes inclusion as a condition of strong objectivity, not a concession to fairness.

Participatory Epistemic Governance

The cases examined in this chapter demonstrate that epistemic injustice is intensified when those affected by algorithmic decisions lack meaningful opportunities to shape or contest them. Participatory approaches to AI design are often justified on democratic or ethical grounds; from an epistemic justice perspective, participation is additionally a matter of epistemic legitimacy.

Epistemically just AI requires forms of participatory epistemic governance in which affected communities are involved not only as data subjects or end users, but as co-producers of knowledge. This includes participation in defining what counts as a problem worth solving, what forms of evidence are relevant, and what criteria should guide system evaluation. Such participation helps counter testimonial injustice by recognizing marginalized groups as credible contributors to knowledge, and hermeneutical injustice by expanding collective interpretive resources.

Transparency as Epistemic Access

Transparency and explainability are frequently invoked as technical remedies for opaque AI systems. However, the analysis in earlier sections shows that opacity constitutes an epistemic harm insofar as it restricts who can understand, question, and contest algorithmic judgments. Transparency must therefore be reconceptualized as epistemic access, rather than mere disclosure.

Epistemic access requires that explanations be intelligible, context-sensitive, and responsive to the needs of those most affected by algorithmic decisions. Philosophical work on algorithmic recourse further emphasizes that meaningful contestation requires more than explanation; it requires actionable pathways for those affected to challenge and revise algorithmic decisions (Venkatasubramanian & Alfano, 2020). Explanations designed solely for developers or regulators risk reproducing epistemic hierarchies by privileging technical expertise over lived experience. From an epistemic justice perspective, explainability should enable affected individuals to engage critically with algorithmic authority, challenge decisions, and seek redress when harms occur.

Reflexivity and Epistemic Accountability

AI systems are often framed as neutral tools, allowing responsibility for epistemic harms to be displaced onto data, models, or abstract notions of "the algorithm." This displacement obscures the social and institutional actors who design, deploy, and benefit from AI systems, creating a gap in epistemic accountability.

An epistemically just approach to AI requires continuous reflexivity regarding the values, assumptions, and power relations embedded in algorithmic systems. This includes ongoing auditing for epistemic harms, mechanisms for contestation and correction, and institutional accountability structures that assign responsibility for epistemic outcomes. Reflexivity, in this sense, is not an optional ethical add-on but a necessary condition for preventing the normalization of epistemic injustice through automation.

Commitment to Epistemic Pluralism

Epistemically just AI must be grounded in a commitment to epistemic pluralism – the recognition that there are multiple, potentially complementary ways of knowing the world. Dominant AI systems are largely trained on data and categories rooted in Western, technoscientific epistemologies, often marginalizing indigenous knowledge systems, non-Western conceptual frameworks, and experiential forms of expertise. Recent decolonial critiques of AI emphasize that dominant machine learning paradigms universalize Western epistemologies while marginalizing indigenous, local, and experiential forms of knowledge, thereby reproducing global epistemic hierarchies through technical systems (Mhlambi, 2020; Mohamed et al., 2020).

Epistemic pluralism does not entail abandoning scientific rigor; rather, it expands the range of interpretive resources through which social reality can be understood. By integrating diverse epistemologies where appropriate, AI systems can reduce hermeneutical gaps and avoid imposing a single, dominant framework of meaning. This commitment directly addresses the risk of algorithmic normalization identified in earlier sections, whereby dominant perspectives are frozen into technical infrastructures.

Taken together, these principles point toward a shift in how AI ethics is conceptualized and practiced. Instead of treating bias as a technical anomaly to be corrected, epistemic justice frames algorithmic harm as a consequence of how epistemic authority is structured, delegated, and insulated from challenge. Designing epistemically just AI therefore requires institutional reforms alongside technical interventions, including participatory governance mechanisms, accountability frameworks, and epistemically inclusive research practices. As Mittelstadt (2019) argues, abstract ethical principles alone cannot guarantee responsible AI without institutional mechanisms that address power, accountability, and governance.

These principles do not promise a final resolution to epistemic injustice in AI systems. Rather, they establish epistemic justice as an ongoing normative commitment – one that demands continual reflection on whose knowledge is recognized, whose voices are heard, and whose interpretations shape the algorithmic mediation of social reality. This orientation sets the stage for the concluding reflections on the broader implications of reframing AI ethics as a politics of knowing.

Conclusion: Epistemic Justice as a Politics of Knowing in the Age of AI

This chapter has argued that Artificial Intelligence should be understood not merely as a technical instrument or decision-support tool, but as an epistemic infrastructure that mediates how knowledge is produced, authorized, and contested. As AI systems increasingly shape what counts as credible information, which experiences are rendered intelligible, and whose interpretations of social reality are privileged, questions of epistemic justice become central rather than peripheral to debates on AI ethics.

Drawing on epistemic injustice theory and feminist epistemology, the chapter addressed three guiding questions. First, how do AI systems contribute to testimonial and hermeneutical injustice, particularly for women and marginalized groups? The analysis demonstrated that algorithmic systems enact testimonial injustice by automating credibility assessments that discount or filter out marginalized voices before they can be recognized as such. At the same time, they produce hermeneutical injustice by operationalizing dominant interpretive frameworks that fail to capture the lived realities of marginalized groups, rendering those experiences epistemically illegible. Crucially, these injustices are not accidental side effects but structural consequences of delegating epistemic judgment to opaque socio-technical systems.

Second, the chapter asked whether feminist standpoint theory can provide an epistemic framework for identifying and mitigating algorithmic harms. The answer advanced here is affirmative, but with an important qualification: feminist standpoint theory does not merely supplement existing bias frameworks; it reframes the problem. By emphasizing the situatedness of knowledge and the epistemic advantage of marginalized standpoints, feminist epistemology exposes the limits of claims to neutrality and objectivity embedded in AI systems. It shows that what appears as technical bias is often the result of deeper epistemic exclusions in how data are collected, categories are defined, and problems are framed. In this sense, standpoint theory offers a critical epistemic lens for diagnosing algorithmic injustice at its roots.

Third, the chapter examined what principles might guide the design and governance of epistemically just AI systems. Building on the identified epistemic harms, it proposed a set of interrelated principles: inclusion of diverse epistemic standpoints, participatory epistemic governance, transparency understood as epistemic access, reflexivity and accountability, and a commitment to epistemic pluralism. These principles were not presented as abstract ethical ideals but as normative responses to the reorganization of epistemic authority effected by AI systems. Together, they point toward a shift from technical bias mitigation to institutional and epistemic transformation.

Across the empirical cases examined – recruitment, facial recognition, and healthcare – the chapter showed that AI systems do more than reproduce existing inequalities: they stabilize and scale them by embedding dominant epistemic assumptions into technical infrastructures that are difficult to contest. What distinguishes algorithmic injustice from earlier forms of discrimination is not simply its speed or scale, but the way epistemic authority is exercised without reciprocal accountability. AI systems make knowledge claims that are treated as authoritative while remaining largely insulated from challenge by those most affected by their consequences.

By reframing AI ethics through the lens of epistemic justice, this chapter has sought to shift the focus from outcomes alone to the conditions of knowing that underpin them. Epistemic justice directs attention to whose knowledge is recognized, whose voices are heard, and whose experiences shape the interpretive frameworks embedded in AI systems. It reveals AI not as a neutral arbiter of truth, but as a site where epistemic power is exercised, negotiated, and potentially redistributed.

More broadly, this analysis invites a rethinking of the relationship between technology and knowledge in contemporary societies. If AI systems are increasingly entrusted with epistemic authority, then questions of legitimacy, accountability, and inclusion cannot be treated as secondary concerns. Designing epistemically just AI is not a finite technical task but an ongoing political and epistemic project – one that requires continual reflection on the social conditions under which knowledge is produced and validated.

By foregrounding epistemic authority as the central site of algorithmic power, this chapter advances epistemic justice not merely as a moral concern, but as a critical analytic framework for understanding how AI reshapes the politics of knowing.

Ultimately, the pursuit of epistemically just AI is inseparable from the pursuit of epistemic justice more broadly. It challenges us to resist the allure of the "view from nowhere" and to recognize that knowledge – whether human or machine-mediated – is always situated, contested, and shaped by power. In doing so, it opens the possibility for AI systems that do not merely mirror existing epistemic hierarchies, but contribute to more inclusive, reflexive, and just ways of knowing the world.

References

Birhane, A., Prabhu, V. U., Kahembwe, E., et al. (2022). The values encoded in machine learning research. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 173–184. https://doi.org/10.1145/3531146.3533083

Bowker, G. C., & Star, S. L. (2000). *Sorting things out: Classification and its consequences.* MIT Press.

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 1–12. https://doi.org/10.1177/2053951715622512

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Code, L. (1991). *What can she know? Feminist theory and the construction of knowledge.* Cornell University Press.

Collins, P. H. (2000). *Black feminist thought: Knowledge, consciousness, and the politics of empowerment.* Routledge.

Costanza-Chock, S. (2020). *Design justice: Community-led practices to build the worlds we need.* MIT Press.

Criado-Perez, C. (2019). *Invisible women: Data bias in a world designed for men.* Abrams Press.

Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters.*https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Dotson, K. (2014). Conceptualizing epistemic oppression. *Social Epistemology, 28*(2), 115–138. https://doi.org/10.1080/02691728.2013.782585

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing.* Oxford University Press.

Hancox-Li, L., & Kumar, I. E. (2021). Epistemic values in feature importance methods: Lessons from feminist epistemology. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 817–826. https://doi.org/10.1145/3442188.3445943

Harding, S. (1993). Rethinking standpoint epistemology: What is "strong objectivity"? In L. Alcoff & E. Potter (Eds.), *Feminist epistemologies* (pp. 49–82). Routledge.

Hardwig, J. (1985). Epistemic dependence. *The Journal of Philosophy, 82*(7), 335–349.

Haraway, D. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies, 14*(3), 575–599.

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese, 169*(3), 615–626. https://doi.org/10.1007/s11229-008-9435-2

Innocenti, M. (2025, preprint). Contributive epistemic injustice in the AI-driven workplace. *SSRN.* https://doi.org/10.2139/ssrn.5379112

Jasanoff, S. (2011). *Reframing rights: Bioconstitutionalism in the genetic age.* MIT Press.

Johnson, G.M. (2021). Algorithmic bias: on the implicit biases of social technology. *Synthese* 198, 9941–9961. https://doi.org/10.1007/s11229-020-02696-y

Kay, J., Kasirzadeh, A., & Mohamed, S. (2025). Epistemic Injustice in Generative AI. *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*. AAAI Press, 684–697.

Medina, J. (2013). *The epistemology of resistance: Gender and racial oppression, epistemic injustice, and resistant imaginations.* Oxford University Press.

Mhlambi, S. (2020). From rationality to relationality: Ubuntu as an ethical and human rights framework for artificial intelligence governance. *Carr Center Discussion Paper Series.* Harvard Kennedy School.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence, 1*, 501–507. https://doi.org/10.1038/s42256-019-0114-4

Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology, 33*, 659–684. https://doi.org/10.1007/s13347-020-00405-8

Rini, R. (2020). *Deepfakes and the epistemic backstop. Philosophers' Imprint*, 20(24), 1–16.

UNESCO. (2024). *Challenging systematic prejudices: An investigation into bias against women and girls in large language models.* UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000388971

Venkatasubramanian, S., & Alfano, M. (2020). The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency* (pp. 284–293). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372876

Zagzebski, L. (2012). *Epistemic authority: A theory of trust, authority, and autonomy in belief.* Oxford University Press.

National Defence University
PL 7, 00861 HELSINKI

Tel. +358 299 800

www.mpkk.fi

ISBN 978-951-25-3584-2 (Pbk.)
ISBN 978-951-25-3585-9 (PDF)
ISSN 2737-0615 (online)

SOTATAIDON YTIMESSÄ

Puolustusvoimat
The Finnish Defence Forces