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*

© Authors & FNDU

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

This work is licensed under the Creative Commons BY-NC 4.0 International License. To view a copy of the CC BY-NC 4.0 license, visit https://creativecommons.org/licenses/by-nc/4.0/deed.en

Punamusta Oy
Joensuu 2026
Finland

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