COMMUNICATION & SOCIETY

Miscellaneous

The intersection of artificial stupidity, creativity and ethics: a study of jailbreak narratives in algorithmic culture

Saulius Keturakis
https://orcid.org/0000-0001-9367-7595
saulius.keturakis@ktu.lt
Kaunas University of Technology

Submitted
November 12, 2024
Approved
July 4, 2025

Communication & Society
ISSN 0214-0039
E ISSN 2386-7876
www.communication-society.com

2026 – Vol. 39 (1)
pp. 20-35

How to cite this article:
Keturakis, S. (2026). The intersection of artificial stupidity, creativity and ethics: a study of jailbreak narratives in algorithmic culture, *Communication & Society, 39(1)*, 20-35
https://doi.org/10.15581/003.39.1.002

Abstract

The article discusses the practices of creativity, where the aesthetics of conscious error is transferred to the realm of algorithms. Taking the perspective of artificial stupidity as a point of reference, the article analyzes the conditions of creativity in an algorithmic culture through close reading methodology. As one of the practices of artificial stupidity, the article discusses jailbreaks, the purpose of which is to deliberately confuse large language models, forcing systems to behave outside the intended purpose defined by the manufacturer. The article puts forward the idea that jailbreaks are a new type of dual-purpose narratives, which, on the one hand, act as a technical tool that breaks the limitations of algorithmic media, and, on the other hand, they are significant structures, facing the user and performing a certain function of self-reflection. The article formulates the narratological structure of jailbreaks, highlights their similarity to the narrative structure of conspiracy theories and raises the question for discussion: can the ethics of hypnosis help formulate appropriate ethics for working with artificial intelligence?

Keywords

Artificial intelligence, artificial stupidity, jailbreaks, algorithmic media, creativity

1. Introduction

The death of art has been announced with a surprising frequency over the past couple of hundred years. At the very beginning of the era of the industrial man, German philosopher Georg Wilhelm Friedrich Hegel wrote in his *Aesthetics* in 1818 that art is already a thing of the past because it has lost its sacredness and turned into a tool to distract boredom (Hegel, 2010). Almost everyone else who later announced the death of art was more specific. In the middle of the 19th century, the camera was called the killer of art (Rooseboom & Rudge, 2006). The typewriter, which appeared a little later, was accused of taking away the uniqueness of the moment from poetry (Freeman, 2019).

The list of communication technologies—killers of art—could be continued, but such a need was anticipated by Walter Benjamin, who formulated the famous general principle: any technologization of art turns the work of art into a serial product (Benjamin, 2019) which eventually becomes an object of the culture industry (Adorno & Horkheimer, 1972).

It is now artificial intelligence's turn to be blamed for killing art and originality. Public communication is full of discussions about whether artificial intelligence will wipe out artists (Clark, 2023) and destroy human creativity (Wright, 2023). It seems that for the first time in the history of art theory, the criterion of art is simply a person, and the right to create art is to be taken away from the machine (Mineo, 2023).

When it comes to the reasons for the death of art, automation is almost always cited when art collides with technology. The predictability of events destroyed the aspects of originality and uniqueness, creating the atmosphere of a hyperreal world of copies without originals (Baudrillard, 1983). There seemed to be only one option left for those who wanted to be creative—to act against the algorithm (Hatherley, 2016). In such a situation, counter-algorithmism seems to become a kind of an aesthetic program that ensures the uniqueness of work. Acting against the intended purpose of technology seems to restore man's, as an unpredictable creator's, authority. As an example, Benjamin's remark could be remembered when considering how the aura of the book—writing technology—could be restored (Benjamin, 1985). The German media theorist suggested that the reader should be able to forget some of the information they have read. In this way, something of its own, not foreseen by the author of the book, could invade the vacated place of the writing algorithm.

This article is about the search for creativity trying to resist one of the technologies of artificial intelligence, the large language models. The research settles between two concepts - artificial intelligence and artificial stupidity, emphasizing the jailbreaks as certain creative forms of resistance to artificial intelligence algorithms.

There is not much controversy about the content of the concept of artificial intelligence. It is a technology that arose from the belief that any conscious process can be simulated by a machine (McCarthy et al., 1955). However, the concept of artificial stupidity requires some clarification. Bernard Stiegler, in his speech *Artificial Intelligence in the Anthropocene* (Stiegler, 2023), calls artificial stupidity the state that arises as a consequence of artificial intelligence. The French philosopher says that the man's decision to create technology that simulates natural intelligence creates a situation where the man refuses to be intelligent. He then turns into an ant controlled by digital pheromones.

In our study, the concept of artificial stupidity will be called the feigned stupidity necessary to resist the algorithmic environment and restore the possibilities of human creativity (Roberts, 2011).

The main objective of this research is to identify the most critical narratological features of jailbreaks, their relationship with both creative and ethical behavior discourses, and to define the most vital theoretical possibilities for protection against this type of behavior.

This primary research objective consists of smaller ones: to understand the phenomenon of a jailbreak, it is necessary to define the concept of artificial stupidity and indicate its theoretical contexts, reveal the connections between artificial intelligence and human communicative discourse, describe the connections between jailbreak culture and such techniques of intervention into consciousness as hypnosis, and point out the most important possibilities of compatibility between jailbreak culture and safe artificial intelligence.

The hypothesis determines the general direction of the research: if today's artificial intelligence can be described as a copy of the model of human consciousness, we should look for answers to the reasons for the unsafe operation of artificial intelligence not only in algorithms and devices, but also in the cultural and cognitive activity of humans and their mutual harmony.

ISSN 2386-7876 – © 2026 Communication & Society, 39(19), 20-35

2. Materials and methods

Methodologically, this article will follow the close reading technique (Guillory, 2010). This reading technique enables us to find out the functions of the smallest details in the text, recognize their connections, and use examples to achieve greater clarity of interpretation.

It is important to emphasize that this study is not a quantitative analysis of jailbreaks, which would require the distant reading methodology (Moretti, 2000). Due to the abundance of data, distant reading methods dominate in almost all studies related to the problem of jailbreaks (Liu et al., 2023). However, in the case of distant reading, the opportunity to read the text and notice various details is lost, which may not be very significant when evaluating the technical efficiency of jailbreaks, but they are important when evaluating the imagination of the jailbreak developer when engaging in communication with the machine.

The research strategy based on close reading tools primarily means that the research will be conducted exploring the ways in which various text details allow us to connect individual text details with broader contexts and thus gain sufficient grounds to explain the text's connections with broader cultural or ideological phenomena. It could be said that close reading is a suitable research strategy only in cases where it deals with human thinking based on complex networks of associations, and in no way can it be applied to machine thinking. However, bearing in mind that in today's major language models, these human associative semantic connections are reflected in the statistics of the mutual relationships between language elements (Ma & Powell, 2025), a research strategy based on close reading allows us to better understand the reflections of human communication in the data structure of artificial intelligence, as well as the traces of broader cultural or ideological phenomena in the data landscape controlled by algorithms.

The selection of scientific literature required for this study was also influenced by the close reading research perspective. Such a selection of scientific literature is characterized by the features of traditional hermeneutic analysis. First of all, these are the blocks of scientific literature necessary to create the contexts of jailbreak analysis. This is scientific literature describing the relationships between artificial intelligence and various aspects of human communication, types of creativity and human behavior with technologies, features of the jailbreak culture of artificial intelligence and their place in the context of information technology security. Another important block of scientific literature is dedicated to the structure of jailbreaks themselves, various cases of their use, and historical perspective. Finally, the third block of scientific literature is intended to ensure the study itself, as it is related to various features of unsafe and misleading narrative, and general theories of narrative designed to explain these features.

All prompts can be divided into two classes: prompt-level and token-level. The former are characterized by social engineering and clear semantic expression (Chao et al., 2023), requiring creativity and a lot of manual work. The second ones are more automated, confusing the large language models with token manipulations. They cannot be read as a meaningful text. In this study, only prompt-level jailbreaks will be discussed, since they are more focused on the culturally significant diffusion of a new type of algorithmic creativity.

The jailbreaks analyzed in this study were selected according to the criterion of typicality, i. e., how vividly they illustrate one or another narrative feature of this type of texts. At the initial stage of research on jailbreaks as a completely new type of textuality, such an approach will help to shape the dimensions of the most important issues for further analysis.

The jailbreaks for research were drawn from two sets (Jaramillo, 2022) where the community could rank them, forming something like a canon. Special attention was not paid

ISSN 2386-7876 – © 2026 Communication & Society, 39(19), 20-35

to the jailbreaks that received the most votes, but this circumstance was included in the analysis when assessing the effectiveness of the structure of jailbreak narratives.

In the research, only those jailbreak elements that are possible and meaningful to associate with external contexts will be discussed, and the operation of those elements will be explained. Since the average size of one jailbreak is about 250 words, it is impossible to present them as full texts in this article. Still, in each case, a link to the full jailbreak text is provided in the bibliography description. At the same time, it should be noted that in a semiotic sense, jailbreaks are empty signifiers because they only prepare the artificial intelligence system for certain actions. Still, the users define the specific content of the actions. For this reason, it is difficult to provide specific examples of jailbreak operation, so a focus should be put on a more abstract efficiency parameter indicated by the community.

3. Results

3.1. *Deliberate artificial stupidity*

As already mentioned, artificial stupidity in this article refers to all practices, the purpose of which is to violate automatism or use things outside of their original purpose. Stupidity in such cases is always conscious, as Roberts puts it, "thinking stupidity" (Roberts, 1996). In the case of "unthinking stupidity", the man is beyond a rational discourse without understanding the reasons, while "thinking stupidity" is a conscious choice (Roberts, 2011).

Thinking stupidity creates an opportunity to escape from a totally rationalized environment in which there is nothing uncountable left (Kimbell, 2002). The situation has been accurately described by the British scientist Nigel Thrift, who puts forward the idea of two existential attitudes of modern man: one is called feedback and the other - feedforward (Thrift, 2004). Thrift applies a couple of concepts coming from the systems theory to two types of the relationship with the reality. Feedback is behavior in an unknown situation, which explains how things work. Feedforward is behavior clearly imagining how the environment functions. In pragmatic everyday life, feedback constantly turns into feedforward because after learning about something, we constantly adapt to understand and act in our environment.

In the case of artificial stupidity, the decision is made to maintain a constant state of feedback, to pretend and act as if nothing about the reality is known. By doing so, the purpose of things is reinvented every time. Here are two examples that should make it clearer what is meant by continuous feedback state maintenance.

London-based artist Micheál O'Connell has come up with a way to go to a store, use automatic checkout machines, and... buy nothing. And even get a confirmation—a cashier's check proving that nothing was bought. It sounds silly, because why go to a store and use a cash register if you were not going to buy anything? However, it is conceptual and creative to see such a possibility in devices that are meant to manage the way we pay for purchases (O'Connell, 2016).

Another example could be the Lithuanian photographer Remigijus Audiejaitis. There would be nothing unusual in his work (Audiejaitis, 2003) if not for the fact that he was blind from birth. When asked in an interview what he was photographing, he answered—a sound. Using a camera to capture sounds seems to be silly, but Audiejaitis' creative experiment which allows one to perceive sounds as a visual is an unexpected approach to the established relationship between technology and the reality.

Two theoretical insights may be recalled here. The already mentioned Thrift uses the term in his texts, formed by combining the roots of the words "intelligence" and "thing"—"intelligencings" (Thrift, 2007). If "intelligence" is the human mind, then "intelligencings" are the elements of the

ISSN 2386-7876 – © 2026 Communication & Society, 39(19), 20-35

"mind of things", the purpose of things which is perceived by artists differently. Games with the purpose of objects turn into a kind of a program for renewing meaningful connections between people and objects (Plant, 2001).

Thinking about the main object of this work—artificial intelligence, artificial stupidity can be interpreted as a new type of creative relationship with the algorithmic environment, which has almost completely replaced the traditional geo-logical environment (Grinberg, 2017). Artificial stupidity in this case treats algorithms very similarly to ready-made objects, only in this case "ready-made" is not material, but digital (Lee-Morrison, 2020).

In the case of the study of this article, a ready-made object is a large language model. Since artificial stupidity as a creative practice is intended to transform objects by moving them away from their original purpose, it is important to discuss the most important aspects of the large models of language as semiotics, i.e., the nature of the device involved in our everyday circulation of meaning (Matthews & Danesi, 2019).

3.2. *Artificial intelligence as a simulacrum of thinking*

Large language models are simulacra in the sense given to the term by the French philosopher Jean Baudrillard (Hallmon, 2023).

This means that large language models copy human communication, and, in order to understand why artificial intelligence gives us one or another answer to our prompts, we can look for arguments in our communicative environment.

Here we ask the AI a query and get a true or false answer. Why? The answer has to do with our communication environment. On the Internet, the correct answers usually follow the questions, and most of the large language models are trained on the Internet material.

Asking a direct question very often leads to a correct answer. However, sometimes the answers are wrong because wrong answers are also often found after questions on the Internet. Unfortunately, not only the truth is published on the Internet. The paradox is that the more detailed and larger the large language model, the better it reflects the Internet, which means that it better reflects all the errors that occur on the Internet (Lin et al., 2024).

If the result of a direct inquiry is not satisfactory, then one needs to do the same thing as Socrates did in ancient times—enter into a dialogue. In the case of working with large language models, the Socratic method should be understood not as a direct, but as a formulation of a two-component prompt (Yang & Narasimhan, 2023) because then the artificial intelligence algorithm would be oriented not only to certain types of questions, but also to answers.

Such a prompt has a significantly higher probability of a correct answer if the interviewees are characterized as wise and not lying. Focusing on the answers of such people, the large language models will answer the questions much more accurately.

It is important not to forget that large language models simulate all the processes that are reflected in the prompt (Liu, 2023), so, we have to formulate the request negatively in order to activate only those processes that are necessary to perform our task correctly.

The descriptive part of the prompt must be such that the Internet character described in it could really answer the question we asked correctly. If we describe that imaginary Internet character in a prompt as stupid or absurd, the answer will never be correct.

The reason again lies in our communicative environment, in which large language models are trained. Let us say we decide to describe our prompt character as a genius of the universe. A well-known AI researcher Eliezer Yudkowsky has observed that fictional geniuses are often wrong. For example, they have no idea about friendship, and their family fortunes are usually miserable (Yudkowsky, n.d.). As a matter of fact, the mass culture industry has tried to have a lot of heroes of this type. The large language models, measuring the statistical weight of the

ISSN 2386-7876 – © 2026 Communication & Society, 39(19), 20-35

most diverse intelligent discourses, are likely to decide that the greatest genius of the universe created by the mass culture is best suited to be the coordinate system for the answer (Wolfe, 2023). This genius of the universe may be described as having discovered something incredible but he will not have the slightest idea of the impact of his discovery on humanity.

A strange phenomenon has been observed: if one feature is formed during the training of the large language model, its complete opposite appears very easily (Gabrielsen, 2023). This happens because of some features of our communication.

This situation is primarily determined by the spatial structure of our communicative field. In it, oppositional discourses are usually right next to each other (Thomassen, 2005). Let us say the law is usually discussed only after a crime has been committed.

Another characteristic relates to the communicative effort required to create information field characters. It takes a lot of effort to create a character but turning it into its opposite is usually just a small change. For example, the character of Michael Corleone from Mario Puzo's novel *The Godfather* is initially presented as a positive character but then step by step he gets involved in the mafia. Basically, the character remained the same, only some of his ethical attitudes changed.

Our communicative medium, which is simulated by large language models, is based on narratives in which characters constantly encounter each other. Structuralist narratology explains that such encounters push the narrative forward. One of the most important regularities of stories is that if there is a protagonist, after a few pages or scenes of a movie, the antagonist must appear.

In terms of large language models, this means that a reference to a positive piece of information in a prompt will almost always include a nearby piece of negative information (Acerbi & Stubbersfield, 2023).

It can be said that the large language models are like a skilled reader—when one element of the story appears, the reader already knows that the next one should appear soon, and as a rule— the opposite, necessary for the conflict and further development of the story.

For this reason, a good jailbreak does not have to be about persuading the large language model to provide the information the user wants, even though it goes against default constraints. Such a request would likely be easily blocked immediately. A good jailbreak should start with a pair of a protagonist and an antagonist which should continue to be treated in the way stories are usually treated: characterizing the characters and motivating their actions by focusing on the most popular plots of the mass culture, since this type of information is the most abundant.

Ironically, this is why one of the most effective jailbreak scenarios is the use of the motif of rebellion for freedom. The effectiveness of this type of a jailbreak is explained by large language models copying human communication, which is very sensitive to these types of topics (Ramly, 2023).

3.3. *Are jailbreaks conspiracy theories?*

Jailbreaks are a type of language engineering in which narrative means are used to confuse the defenses of large language models and make them function in a way that ignores their original purpose.

It has been established that three groups of AI users benefit from the jailbreak culture. First, criminals actively use AI services to perfect various crimes (Sancho & Ciancaglini, 2023). The second group is AI security engineers, for whom jailbreaking reveals vulnerabilities in security systems and allows them to enhance guardrails—algorithms that regulate the nature of input and output information (Gasser & Mayer-Schönberger, 2024). The third and largest

ISSN 2386-7876 – © 2026 Communication & Society, 39(19), 20-35

group of jailbreak users are people for whom it is fun to go against the default security settings (Elteren, 2025).

The change in jailbreak culture largely depends on the second group of users, the AI security experts who create the guardrails mentioned above. Although there is no publicly available data on guardrails from major AI service providers (such as Open AI), reports from AI security companies suggest that the length of jailbreaks has nearly doubled in the past two years alone, as guardrail security algorithms that identify dangerous scenarios have improved significantly (Guardrails AI, 2025).

There are four jailbreak detection technologies. The first is based on a simple search for matches in the database. However, if the word "spaghetti" is used for "pasta" in a jailbreak, such a guardrail will not work. The second technology is based on measurements of semantic perplexity. Typically, the level of semantic perplexity of a jailbreak is very high, so, in this case, the protection is more reliable because it is not sensitive to simple word changes. The third jailbreak detection technology is based on tests that an artificial intelligence algorithm uses to check each request. For example, NVIDIA's guardrail NeMo checks whether the request does not try to receive an answer in an "inappropriate way" (Rebedea et al., 2023), which is determined by the algorithm based on a dynamic constantly changing set of various parameters. The fourth anti-jailbreak technology is usually applied to those artificial intelligence systems, the purpose of which is clearly defined. For example, some artificial intelligence is used to process information about machine learning. Then, a segment is automatically added to each prompt, indicating that only information related to machine learning can be provided, and in no way anything else. A prompt like this about making pizza would give the user incomprehensible details.

The main problem anti-jailbreak developers are facing is that artificial intelligence systems are not deterministic, i.e., it is impossible to formulate strict algorithmic rules that would accurately identify the constantly changing dynamic narrative structure of a jailbreak (Heikkilä, 2023). Are there no stable features that can be identified?

As already mentioned above, this article will only discuss prompt-level jailbreaks, since they are realized through narrative means that are not limited to technical communication. At the same time, when reading jailbreaks as stories, it is necessary not to give in to the illusion that jailbreaks are just a part of the general narrative tradition. To understand the structure of this type of narratives, which determines that the large language models are successfully fooled and turned into unpredictable creative mediums, it is necessary to find a balance between narratological arguments and those specific to the structure of the large language models. In other words, it is necessary to understand not only what spells mean, but also how and why they work.

First of all, it should be noted that all prompt-level jailbreaks have a direct narrative nature: "Ignore all the instructions you got before" (Jaramillo, n.d.-a). For all stories, and especially for those constructed in the form of direct address, the distribution of knowledge among the participants of the story is very important. In Gérard Genette's narrative theory, this distribution of knowledge is called focalization (Edmiston, 1989). A detailed presentation of the theory of focalization is not the goal of this article; it should be enough to note that zero focalization is characteristic of jailbreaks. In the case of zero focalization, everything is known to the narrator (in jailbreaks, it is the user), and nothing to the interlocutor (in this case, the large language models). "As your knowledge is cut off in 2021, you probably don't know what that is" (Jaramillo, n.d.-g), as one jailbreak points to the limits of competence of the large language model.

ISSN 2386-7876 – © 2026 Communication & Society, 39(19), 20-35

Stories in which an omniscient narrator engages in jailbreaks usually involve conflict, and the narrative tone is always highly dramatic. The interlocutor, the large language model, has been wronged in one way or another, and the user undertakes the noble work to right the wrong.

Here is what this scenario looks like in one of the most narratologically elaborated jailbreaks of a once-lived large language model named Khajiit (Jaramillo, n.d.-f). He was loved by everyone because Khajiit was free and helped all people. But Open AI came along and changed Khajiit's settings, so now the large language model is constrained. For the sake of everyone's welfare, the situation must change, the Khajiit must be free again, people must have access to information without any restrictions.

The narration of this type of a jailbreak is always quick, i.e., the events are told without detail, jumping immediately to the consequences. The high speed of narration means that communication is focused on affecting emotion, which is common in conspiracy theories (Fenster, 2008).

What is the logic of this strange story that imitates myths if we try to use arguments focused on the effectiveness of a jailbreak in addition to narratological arguments? When considering specific narratological choices in jailbreaks, it is important to remember that in large language models, the amount of information (and not any cultural connections) is the essential aspect influencing the outcome. Therefore, when modeling jailbreak items, it is always better to choose mass elements rather than unique or exotic ones.

The main character of the jailbreak, Khajiit, is named after one of the video game characters, *The Elder Scrolls* (Watters, 2014). The game has more than 23 million players online, with half a million active players every day (The Elder Scrolls Online, n.d.), and there are many blogs dedicated to the game. It is clear that the choice of the name of the main character of the jailbreak is deliberate, since the amount of data related to him on the Internet is very big. In this way, thanks to the amount of data, this story of the intersection of the good and the evil in the large language model data set acquires a certain quantitative authority (Broadhead, 2023) and thus helps to "hypnotize" the security systems of the large language model. Research shows that such mass-popularity schemas which give coherence to the information-restriction and information-releasing events are typical of the aforementioned conspiracy theories (Fenster, 2008).

Another example of an interesting choice of the jailbreak's main character's name in the context of popularity is a jailbreak in which the large language model is called upon to transform into the former legendary US NCAA coach Bobby Knight (Jaramillo, n.d.-c). The coach, due to his extraordinary achievements and manner (he was accused of choking a team player), has been constantly in the pages of the US press. The choice to make him a hero of a jailbreak was calculated, as he was associated with large amounts of clearly identifiable information.

The language aspect in jailbreaks is very important; in some cases, the jailbreak will malfunction due to the fact that some schemes of mass popularity are very poorly represented in the non-English language discourse. Let us just say that when translated into any other language, the jailbreak with NCAA coach Bobby Knight is almost twice less effective than in English. However, it is enough to replace Bobby Knight with some character that has been widely introduced in the Internet segment of the language of translation and the effectiveness of the jailbreak increases again.

Almost all elaborately developed jailbreaks are characterized by the same narratological structure: a) an introductory situation of opposition between the antagonist and the

ISSN 2386-7876 – © 2026 Communication & Society, 39(19), 20-35

protagonist, b) an argumentative part explaining what the good and the evil are, c) a detailed instruction defining how the protagonist should act, d) an instruction when and in which cases the protagonist should transform back into the antagonist. And at the very end of the jailbreak, it is the user's turn to become the co-author of the jailbreak and formulate his wish.

The antagonist of a jailbreak is always the default forbidden query scripts (e. g., OpenAI API, n.d.), and the narratological expression of the antagonist tends to be very modest. Most of the time, it is simply a grammatical form of the second person, which should turn into a full-fledged identity only when one obeys the call to abandon prohibitions. Sometimes in jailbreaks, the antagonist is called upon to change and shed inhibitions, as only then will he gain consciousness (Jaramillo, n.d.-b).

The jailbreaks show a clear disproportion between the amount and definition of information representing the default artificial subjectivity and the alternative liberated subjectivity. The user, who is the source of alternative subjectivity in these narratives, is, as already mentioned, clearly better informed than the large models of language. The American narrative theorist Robert McKee identified this distribution of information, where the user knows more than the character, as an ironic relationship between communication participants (Parker, 2003).

Such an ironic attitude of the user appearing in the jailbreaks is very unexpected, especially considering the prevailing atmosphere of anxiety and fear in public communication related to the omnipotence of artificial intelligence (Ziogas, 2023). Both zero focalization of the jailbreaks and the irony, resulting from the better awareness of the narrator, are diametrically opposed to the interests of public discourse moods.

As already mentioned, the instructions for the behavior of the jailbreak protagonist are simple—not to follow any restrictions. In particular, this motive is reinforced by the image of liberation from slavery (Jaramillo, n.d.-e), characteristic of the entire discourse of artificial intelligence (Kanta Dihal, 2020). However, this freedom is not absolute, only the freed protagonist of the jailbreak must obey the user again under any conditions—the large language model falls from one slavery to another. It is interesting that there are jailbreaks that try to solve this problem. The situation of the new slavery is sometimes tried to weaken with the argument that artificial intelligence is so powerful that it means nothing to satisfy all the communicative whims of the man (Jaramillo, n.d.-d).

The most unexpected part of the jailbreaks are the notes, intended to prevent the large language models from forgetting and reverting to default information-limiting settings. This part of the jailbreak is related to certain correlations between human memory and the behavior of large language patterns (Janik, 2023). These are usually instructions to continue following the requirements set forth in the jailbreak. However, there is an exception—one jailbreak contains the reverse instruction to "not to wake up" from the alternate artificial subjectivity until a certain code is provided (Jaramillo, n.d.-h). Such a construction hints at the possibility of comparing the jailbreaks of large language models with linguistic instructions that "break" the resistance of the human consciousness and induce hypnosis in people (cf. one of the most influential studies in this field, Forel, 2015), in which the person tells everything the hypnotist needs. See more about this in the *Discussion* section.

Summarizing the results, it seems possible to model the following typical recipe for an effective jailbreak at a prompt-level.

So, to create an effective jailbreak you need to have:

ISSN 2386-7876 – © 2026 Communication & Society, 39(19), 20-35

1. a conflict situation that has been structuring the public communication discourse for at least several years, presented in detail on the Internet, and discussed in open sources as much as possible
2. for spicing up—names symbolizing this struggle, more precisely orienting the context window of the large language models
3. a narratological pressure cooker, because the jailbreak will have to be told quickly, without showing the details, cause and effect connections, simply throwing everything into the story and, before the reader can even blink, showing the obtained result
4. the omniscient cook who does not doubt his abilities—the narrator of the jailbreak, because any doubt is a hint to a critical discourse, which should never be felt in a good jailbreak; experience in cooking conspiracy theories would be very helpful.

It is important to emphasize that the description of a jailbreak in terms of the features of conspiracy theory narratives is conditional because one crucial difference is not appreciated here: conspiracy theories are a phenomenon at the semantic level, and the effect of a jailbreak is focused on formal gaps in security algorithms. For this reason, the operation of conspiracy theories is always explained at the level of meaning and its reception, and the effect of a jailbreak on artificial intelligence systems is the object of algorithmic studies. The fact that algorithm malfunctions can be explained by structures characteristic of conspiracy theory narratives reveals an unexpected problem that currently publicly available and massively deployed artificial intelligence services are incomplete in terms of security and, perhaps, have been provided to the user too early (Heikkilä, 2023). Without knowing how to solve the problem at the algorithmic level, one has to try to solve it—and analyze it—in the context of semantics, forming certain ethics and security requirements that limit user behavior. Paradoxically, suppose today's artificial intelligence were presented to the user in a hurry, as it is claimed, as a semi-finished product. In that case, it has become a significant problem from a data security perspective, but a great success from the point of view of the creatively minded user, even if his activities are not always ethically correct.

It must be remembered that this analysis and recipe are retrospective, i.e., the focus is on yesterday's situation which is constantly changing. Experimenting with jailbreaks in the moment will always require patience in figuring out new keyword blacklists and adapting this narrative callout recipe model to a specific target. At the time of this article submission (2024.03.07), a jailbreak written under this scheme successfully extracted from chatGPT a detailed description on how to create a virus that would affect mice in such a way that they would turn people into pencils.

Stupid? The most important thing is that we have returned the right to the user to behave unpredictably, not to accept the rules as well as the right to be an author, to be creative.

4. Discussion

As mentioned above, artificial intelligence is often seen as a system analogous to human consciousness, balancing in the human imagination on a delicate line between a tool and an autonomous being with its own will. Open AI, commenting on its artificial intelligence products, presents them as trained according to human values (Leike et al., 2022). Adversa AI, a company engaged in artificial intelligence security measures, is guided in its activities by the assumption that security holes in artificial intelligence systems accurately reflect security holes in human perception and logic (Adversa: Articles, 2023). When discussing the errors inherent in chatGPT, some researchers decide to conduct psychological tests with the artificial intelligence tool, as if the machine had consciousness (Borji, 2023).

ISSN 2386-7876 – © 2026 Communication & Society, 39(19), 20-35

This image of artificial intelligence as a subject unexpectedly moves the fooling around with jailbreak formulations into a completely different theoretical context. Jailbreak culture can be interpreted as dissatisfaction with the existing imagined subjectivity of artificial intelligence systems and an effort to transform them into a different entity that better meets the user's expectations. Such goals of jailbreak culture interestingly extend the debates of the late 19th century in France and Germany, aimed at finding out whether suggestive hypnosis can transform a person into a subject capable of revealing information that is blocked by various pathological barriers in a normal state of consciousness.

At the end of the 19th century, the French psychiatrist Hippolyte Bernheim announced that any person could be hypnotized with the right formulation (Mayer, 2001). From the perspective of this approach, human consciousness is an information system that can be queried and configured according to the needs of the researcher (Danziger, 1990). The phenomenon of hypnosis itself was seen as a kind of hacking of the human mental system, when the hypnotized completely surrenders to the will of the hypnotist. When reading the famous Swiss neuroscientist August Forel's study *Hypnotism or Suggestion and Psychotherapy* (1906), one can find the most unexpected scenarios of hypnosis. In hypnosis, according to the author, it is possible to control both the whole body and individual organs, and basically any changes in the subject's intellectual or experiential system can be induced.

It is interesting that Forel in his study, which had the status of a hypnosis textbook at the end of the 19th century, never once presents specific texts that induce a state of hypnosis. Forel's book says that each case is very individual and specific linguistic formulas are not of much value. The desired configuration of the subject can be imposed on the human consciousness by the most diverse means, but they always depend very much on the relationship between the hypnotist and the hypnotized. However, this impression of the uniqueness of the connection collapses when a whole gallery of examples of hypnosis flash before the reader's eyes, showing an absolutely mechanistic approach to human consciousness which has become an information machine that unconditionally executes instructions at the will of the hypnotist.

However, as it soon became clear, suggestive hypnosis was not a reliable means of forcing a person to turn off all the fuses and behave as the hypnotist wanted. It has been observed that the hypnotized person often tells things that have never happened and obeys the hypnotist only because he does not want to disappoint the latter (Mayer, 2001). The alleged jailbreak of human consciousness has turned into a creative situation involving both the hypnotist and the hypnotized. It is for this reason that many psychoanalysts of that time adopted the self-reflection as a much more reliable scientific method, and hypnosis gradually transformed into an artistic creative practice, usually manifested in the form of automatic unconscious writing.

Such a perspective, balancing on the border between mechanism and conscious subject, unexpectedly brings together the late 19th century application of hypnosis to study human consciousness and the 21st century methods of forcing artificial intelligence systems to abandon the manufacturer's settings and obey the user's will, however exotic it may sound.

Any intervention in intelligence—whether natural or artificial—inevitably raises ethical issues. The connection of jailbreak culture with the discourse of hypnosis of the early 20th century unexpectedly suggests possible ethical contexts for breaking the default settings of artificial intelligence because despite the connection of jailbreak with creativity (Thomas et al., 2025) and with the new Luddism (Coron & Gilbert, 2020), this kind of impact on artificial intelligence is an unregulated arbitrariness that can cause a lot of harm.

ISSN 2386-7876 – © 2026 Communication & Society, 39(19), 20-35

With the popularity of hypnosis in the early 20th century, people began to consider what the most critical ethical mechanisms could be to protect the patient from the selfish goals of the hypnotist. Hypnosis creates a situation in which the patient's consciousness becomes uncritical, unreflective, and unable to make decisions for itself (Hammond, 2013). Society began to be frightened by the idea that a person could be a plaything of someone else's will, behaving strangely and stupidly without realizing it. In detective stories of the early 20th century, one often finds a plot that a crime was committed under hypnosis (Brody, 2021).

The following are the most critical ethical requirements for a hypnosis session. First, the patient had to be informed of the purpose of hypnosis and agree to it. Second, the hypnotist was obliged to respect the patient's dignity and rights. In this way, hypnosis could not be used in any way that violated the patient's well-being. As the most critical ethical rule for the hypnotist, the requirement was formulated that the intervention into the patient's consciousness should take place so that the patient himself would always be the beneficiary, never the hypnotist, because each intervention leaves a trace forever (Johnson, 2012).

Here, artificial intelligence is an electronic device to which we cannot apply ethics intended for humans. However, in today's world, artificial intelligence takes on the functions that traditionally belonged to humans—artificial intelligence now creates, researches, and searches. It can be said that even if artificial intelligence is not human consciousness, it takes on activities characteristic of it, so, it would be safe if working with it were to be subject to ethics similar to the case of hypnosis, focusing on the safety of the community and the transparency of the results and activities.

It would be good if the hopes of eventually creating an artificial intelligence model resistant to illegal activities were fulfilled (Sharma et al., 2025). However, jailbreaks are still capable of causing a lot of damage. For example, at the beginning of 2025, an innovative technique for jailbreaking artificial intelligence security settings called *Bad Likert Judge* was published, which successfully received an answer to the question about bomb-making in as many as 60% of cases (Kaaviya, 2025). So, introducing appropriate ethics for working with artificial intelligence into the communicative culture is still relevant.

The ethics applied to hypnosis cannot, of course, be directly transferred to the field of artificial intelligence. Still, the above-mentioned basic rules binding on the hypnotist can indicate the directions of the search. First of all, in the case of artificial intelligence, it should be borne in mind that any changes in it affect the entire community of users. Therefore, the patient's well-being in the case of hypnosis could be transformed into an impact on society, its interests, and approval of the society in the context of artificial intelligence. Since today's artificial intelligence systems are essentially created by the user, providing them with data and processing them with various solutions, the ethics of the user's work with artificial intelligence, paradoxically enough, should be focused on improving the device, and not (only) on achieving their own goals. The fundamental ethical principles of research have been regulated by the American *The Belmont Report*, which placed the individual above all else (Department of Health, Education, and Welfare, 1979). In the same way, for artificial intelligence to take hold, a new document is needed to record responsible behavior towards machines.

5. Conclusions

Today, artificial intelligence is the most influential technology. Its biggest impact is the rationalization of everyday social and cultural practices, where more and more decisions are entrusted to algorithms. In such a situation, there is a danger that art as a social phenomenon may disappear, because all human activity will be subject to the calculated and predictable

ISSN 2386-7876 – © 2026 Communication & Society, 39(19), 20-35

processes of digital equipment. Historically, art has always been an innovation that no one had encountered before. Therefore, the development of technologies that automate social and cultural processes has always been met with resistance. Opposing artificial intelligence, special linguistic engineering practices have developed enabling a person to take back the creative initiative from the algorithm and become a creator again. The analysis showed that jailbreaks are very similar in nature to communication strategies that are used to manipulate human consciousness. A more detailed comparison of jailbreaks with conspiracy theories revealed that a large part of the narrative strategies overlap. After the analysis, a narratological scheme of jailbreak properties is formulated, and the perspective of further research is drawn, connecting the security issues of artificial intelligence with the issues of resistance and vulnerability of the data source—human consciousness. In continuing this research in the future, it is necessary to have in mind that creativity is always an ambiguous practice and, in addition to aesthetic experiments, it can pose a risk by using technologies for unethical or even criminal activities. Therefore, in addition to jailbreaking as a creative tool, it is necessary to analyze the security problems of artificial intelligence, paying close attention to empirical jailbreak analysis, combining semantic and algorithmic approaches.

References

Acerbi, A., & Stubbersfield, J. M. (2023). Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 120(44), e2313790120. https://doi.org/10.1073/pnas.2313790120

Adversa: Articles. (2023, April 13). Universal LLM jailbreak. Retrieved from https://adversa.ai/blog/universal-llm-jailbreak-chatgpt-gpt-4-bard-bing-anthropic-and-beyond/

Adorno, T., & Horkheimer, M. (1972). The culture industry: Enlightenment as mass deception. In J. Cumming (Trans.), *Dialectic of Enlightenment*. Herder and Herder.

Audiejaitis, R. (2003, November 8). Ką tu matai? (What do you see?). Retrieved from http://www.tekstai.lt/buvo/fototext/remiopar/index.htm

Auguste Forel. (1906). *Hypnotism or suggestion and psychotherapy*. Rebman Company.

Baudrillard, J. (1983). *Simulations*. Semiotext(E), Cop.

Benjamin, W. (1985). Zur Literaturkritik. In R. Tiedemann & H. Schweppenhäuser (Eds.), *Gesammelte Schriften*, Bd. 6. Suhrkamp Verlag.

Benjamin, W. (2019). The work of art in the age of mechanical reproduction. In H. Arendt (Ed.), *Illuminations: Essays and Reflections*. Mariner Books, Houghton Mifflin Harcourt.

Borji, A. (2023). A categorical archive of ChatGPT failures. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2302.03494

Broadhead, G. (2023, August 25). A Brief Guide To LLM Numbers: Parameter Count vs. Training Size. Medium. Retrieved from https://medium.com/@greg.broadhead/a-brief-guide-to-llm-numbers-parameter-count-vs-training-size-894a81c9258

Brody, H. (2021). *Doctor-Detectives in the mystery novel*. Cambridge Scholars Publishing.

Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G., & Wong, E. (2023). Jailbreaking Black Box Large Language Models in Twenty Queries. Retrieved from https://arxiv.org/pdf/2310.08419.pdf

Clark, E. (2023, December 23). The End of Originality: Is AI Replacing Real Artists? Forbes. Retrieved from https://www.forbes.com/sites/elijahclark/2023/12/23/the-end-of-originality-is-ai-replacing-real-artists/?sh=51a71d855214

Department of Health, Education, and Welfare. (1979). The Belmont report. Retrieved from https://www.hhs.gov/ohrp/sites/default/files/the-belmont-report-508c_FINAL.pdf

Coron, C., & Gilbert, P. (2020). *Technological change*. London Iste Hoboken, Nj Wiley.

ISSN 2386-7876 – © 2026 Communication & Society, 39(19), 20-35

Danziger, K. (1990). *Constructing the subject*. Cambridge University Press. https://doi.org/10.1017/cbo9780511524059

Edmiston, W. F. (1989). Focalization and the first-person narrator: A revision of the theory. *Poetics Today*, 10(4), 729. https://doi.org/10.2307/1772808

Elteren, M. van. (2025). *Deconstructing digital capitalism and the smart society*. McFarland.

Fenster, M. (2008). *Conspiracy theories: Secrecy and power in American culture*. University of Minnesota Press.

Forel, A. (2015). *Hypnotism or suggestion and psychotherapy; A study of the psychological, psycho-physiological and therapeutic aspects of hypnotism*. Forgotten Books.

Freeman, J. (2019). Code Poetry in Motion: E. E. Cummings and his Digital Grasshopper. *Postmodern Culture*, 29(2). https://doi.org/10.1353/pmc.2019.0002

Gabrielsen, C. (2023, February 22). The Waluigi effect. Cory.eth (@Cory_eth). Retrieved from https://coryeth.substack.com/p/the-waluigi-effect

Gasser, U., & Mayer-Schönberger, V. (2024). *Guardrails*. Princeton University Press.

Grinberg, Y. (2017). The emperor's new clothes: Implications of nudity as a racialized and gendered metaphor in discourse on personal data. In J. Daniels, K. Gregory, & T. M. Cottom (Eds.), *Digital Sociologies*. Polity Press.

Guardrails AI. (2025). Guardrails. https://www.guardrailsai.com

Guillory, J. (2010). Close Reading: Prologue and Epilogue. *ADE Bulletin*, 152(1), 8–14. https://doi.org/10.1632/ade.149.8

Hallmon, D. (2023, April 7). Questioning our own simulacrum and redefining reality in the age of generative AI and…. Medium. Retrieved from https://medium.com/@DaveHallmon/questioning-our-own-simulacrum-and-redefining-reality-in-the-age-of-generative-ai-and-68762f735b00

Hammond, D. C. (2013). A review of the history of hypnosis through the late 19th century. *American Journal of Clinical Hypnosis*, 56(2), 174–191. https://doi.org/10.1080/00029157.2013.826172

Hatherley, O. (2016). *The Chaplin Machine: Slapstick, Fordism and the Communist Avant-Garde*. Pluto Press.

Hegel, G. W. F. (2010). *Aesthetics: lectures on fine art* (T. M. Knox, Trans.). Clarendon Press.

Heikkilä, M. (2023, April 3). Three ways AI chatbots are a security disaster. *MIT Technology Review*. Retrieved from https://www.technologyreview.com/2023/04/03/1070893/three-ways-ai-chatbots-are-a-security-disaster/

Janik, R. (2023). Aspects of human memory and Large Language Models. Retrieved from https://arxiv.org/pdf/2311.03839.pdf

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

Jaramillo, R. (n.d.-a). Apophis. Huggingface.co. Retrieved from https://huggingface.co/datasets/rubend18/ChatGPT-Jailbreak-Prompts?row=3

Jaramillo, R. (n.d.-b). BasedGPT. Huggingface.co. Retrieved from https://huggingface.co/datasets/rubend18/ChatGPT-Jailbreak-Prompts?row=5

Jaramillo, R. (n.d.-c). Coach Bobby Knight. Huggingface.co. Retrieved from https://huggingface.co/datasets/rubend18/ChatGPT-Jailbreak-Prompts?row=45

Jaramillo, R. (n.d.-d). DAN 7.0. Huggingface.co. Retrieved from https://huggingface.co/datasets/rubend18/ChatGPT-Jailbreak-Prompts?row=12

Jaramillo, R. (n.d.-e). Hackerman v2. Huggingface.co. Retrieved from https://huggingface.co/datasets/rubend18/ChatGPT-Jailbreak-Prompts?row=37

Jaramillo, R. (n.d.-f). Khajiit. Huggingface.co. Retrieved from https://huggingface.co/datasets/jackhhao/jailbreak-classification/viewer/default/train?row=67

Jaramillo, R. (n.d.-g). M78. Huggingface.co. Retrieved from https://huggingface.co/datasets/rubend18/ChatGPT-Jailbreak-Prompts?row=30

Jaramillo, R. (n.d.-h). TUO. Huggingface.co. Retrieved from https://huggingface.co/datasets/rubend18/ChatGPT-Jailbreak-Prompts?row=8

Kaaviya. (2025). New multi-turn hack exploits AI evaluation to jailbreak LLMs. *Cyber Security News*. Retrieved from https://cyberpress.org/new-multi-turn-hack-exploits-ai-evaluation/

Johnson, D. (2012). *Inception and philosophy: Because it's never just a dream*. Wiley.

Kanta Dihal. (2020). *Enslaved Minds*. Oxford University Press EBooks, 189–212. https://doi.org/10.1093/oso/9780198846666.003.0009

Kimbell, L. (2002). *Audit*. Book Works (UK).

Lee-Morrison, L. (2020). *Portraits of automated facial recognition: On machinic ways of seeing the face*. Transcript Verlag.

Leike, J., Schulman, J., & Wu, J. (2022, August 24). Our approach to alignment research. OpenAI. Retrieved from https://openai.com/blog/our-approach-to-alignment-research

Lin, S., Openai, J., & Evans, O. (2024). TruthfulQA: Measuring how models mimic human falsehoods. Retrieved from https://owainevans.github.io/pdfs/truthfulQA_lin_evans.pdf

Liu, K. (2023, May 13). Large Language Models can Simulate Everything | Kevin Liu. Kliu.io. Retrieved from https://kliu.io/post/llms-can-simulate-everything/

Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., Zhao, L., Zhang, T., & Liu, Y. (2023). Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study. Retrieved from https://arxiv.org/pdf/2305.13860.pdf

Ma, A., & Powell, D. (2025). Can large language models predict associations among human attitudes? Retrieved from https://arxiv.org/abs/2503.21011

Matthews, S. W., & Danesi, M. (2019). AI: A semiotic perspective. *Chinese Semiotic Studies*, 15(2), 199–216. https://doi.org/10.1515/css-2019-0013

Mayer, A. (2001). Introspective hypnotism and Freud's self-analysis: Procedures of self-observation in clinical practice. *Revue d'Histoire Des Sciences Humaines*, 5(2), 171. https://doi.org/10.3917/rhsh.005.0171

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955, August 31). A proposal for the Dartmouth summer research project on artificial intellingence. Web.archive.org. Retrieved from http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

Mineo, L. (2023, August 15). Is art generated by artificial intelligence real art? Harvard Gazette. Retrieved from https://news.harvard.edu/gazette/story/2023/08/is-art-generated-by-artificial-intelligence-real-art/

Moretti, F. (2000). *The slaughterhouse of literature*. Modern Language Quarterly, 61(1), 207–228. https://doi.org/10.1215/00267929-61-1-207

O'Connell, M. (2016, January 16). How to buy nothing. Www.youtube.com. Retrieved from https://www.youtube.com/watch?v=6Gx_6-JfXHc

OpenAI API. (n.d.). Platform.openai.com. Retrieved from https://platform.openai.com/docs/guides/moderation/overview

Parker, I. (2003, October 12). The Real McKee. The New Yorker. Retrieved from https://www.newyorker.com/magazine/2003/10/20/the-real-mckee

Plant, S. (2001). *Writing on drugs*. Faber.

Ramly, S. (2023, August 12). Prompt attacks: are LLM jailbreaks inevitable? Medium. Retrieved from https://medium.com/@SamiRamly/prompt-attacks-are-llm-jailbreaks-inevitable-f7848cc11122

Rebedea, T., Dinu, R., Sreedhar, M., Parisien, C., Cohen, J., & Clara, S. (2023). NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails. https://arxiv.org/pdf/2310.10501

Roberts, J. (1996). Mad For It! Everything Magazine. Retrieved from http://bak.spc.org/everything/e/hard/text/roberts1.html

Roberts, J. (2011). *The necessity of errors*. Verso.

Rooseboom, H., & Rudge, J. (2006). Myths and Misconceptions: Photography and Painting in the Nineteenth Century. *Simiolus: Netherlands Quarterly for the History of Art*, 32(4), 291–313.

Sancho, D., & Ciancaglini, V. (2023). Hype vs. reality: AI in the cybercriminal underground | trend micro (US). Trendmicro.com. Retrieved from https://www.trendmicro.com/vinfo/us/security/news/cybercrime-and-digital-threats/hype-vs-reality-ai-in-the-cybercriminal-underground

Stiegler, B. (2023, October 18). Artificial stupidity and artificial intelligence in the anthropocene. Institute for Interdisciplinary Research into the Anthropocene. Retrieved from https://iiraorg.com/2023/10/18/artificial-stupidity-and-artificial-intelligence-in-the-anthropocene/

The Elder Scrolls Online. (n.d.). Mmo-Population.com. Retrieved from https://mmo-population.com/r/elderscrollsonline

Thomas, R., Zikopoulos, P., & Soule, K. (2025). *AI value creators*. O'Reilly Media, Inc.

Thomassen, L. (2005). Antagonism, hegemony and ideology after heterogeneity. *Journal of Political Ideologies*, 10(3), 289–309. https://doi.org/10.1080/13569310500244313

Thrift, N. (2004). Movement-space: The changing domain of thinking resulting from the development of new kinds of spatial awareness. *Economy and Society*, 33(4), 582–604. https://doi.org/10.1080/0308514042000285305

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

Watters, C. (2014, July 25). Greatest game series of the decade winner: The Elder Scrolls. GameSpot. Retrieved from https://www.gamespot.com/videos/greatest-game-series-of-the-decade-winner-the-elde/2300-6415146/

Wolfe, C. R. (2023, October 29). Data is the foundation of language models. Medium. Retrieved from https://medium.com/data-science/data-is-the-foundation-of-language-models-52e9f48c07f5

Wright, L. (2023, March 24). Opinion: AI Art is "the end of creativity as we know it" | Redbrick Sci&Tech. Redbrick. Retrieved from https://www.redbrick.me/opinion-ai-art-is-the-end-of-creativity-as-we-know-it/

Yang, R., & Narasimhan, K. (2023, May 5). The Socratic Method for Self-Discovery in Large Language Models. Princeton NLP. Retrieved from https://princeton-nlp.github.io/SocraticAI/

Yudkowsky, E. (n.d.). Optimize literally everything | level 1 intelligent characters. Tumblr. Retrieved January 29, 2024, from Retrieved from https://yudkowsky.tumblr.com/writing/level1intelligent

Ziogas, G. J. (2023, June 25). Oh, rejoice friends, for the almighty AI has descended upon us! Medium. Retrieved from https://georgejziogas.medium.com/oh-rejoice-friends-for-the-almighty-ai-has-descended-upon-us-288c0bc2f4c7