https://doi.org/10.37708/ezs.swu.bg.v24i3.18

CYBERPUNK - UNDER THE HEAVY BOOT OF BUREAUCRACY

Nevena Krumova

South-West University "Neofit Rilski", Bulgaria
Email: nevena_krumova@swu.bg

ORCID ID: 0000-0002-7985-4847

ABSTRACT: This article examines the cyberpunk genre as a conceptual point of departure for analyzing representations of highly technologized yet profoundly dehumanized societies, understood as anticipations of contemporary ethical challenges related to artificial intelligence. Particular attention is paid to bureaucracy as a structural feature of modernity and to the question of whether AI can resolve bureaucratic dysfunctions or instead intensify processes of moral erosion and dehumanization.

The study is situated within an interdisciplinary field at the intersection of philosophy, political theory, media studies, and the ethics of technology. It argues for the development of philosophical critical futurology as a research framework capable of critically engaging with future technological projections without collapsing into conspiracy thinking or ideological programs serving specific power interests. The analysis brings together three key theoretical strands: Michel Foucault's account of power–knowledge relations and discursive apparatuses of power; Hannah Arendt's theory of the banality of evil, including her distinction between crimes against humanity and crimes against the human status; and psychoanalytic motifs derived from Carl Gustav Jung's concept of the shadow as projection onto the Other. These perspectives are applied to a concrete contemporary case—the "Grok incident"—examined as an instance in which hate speech, totalitarian discourse, and the banality of evil intersect within the context of artificial intelligence. The article calls for heightened ethical vigilance and argues against legitimizing hate speech as a form of logical, scientific, or "uncomfortable" truth, warning against its normalization under conditions of increasing technological mediation.

KEYWORDS: Artificial Intelligence (AI), Hate speech and Dehumanization, Banality of Evil, Totalitarian Discourse, A crime against the Human Status, Philosophical Critical Futurology

Introduction

The concept of cyberpunk is broad and multi-layered in scope. In its narrowest sense, it is defined as a subgenre of science fiction "… characterized by countercultural antiheroes trapped in a dehumanized, high-tech future" (Encyclopedia Britannica, 2026a). In a broader sense, however, cyberpunk functions as a cultural paradigm that extends beyond literature to encompass the visual arts, cinema, video games, and popular culture, while simultaneously offering a form of social and political critique. In this context, the thesis of Alexander Popov may also be considered, namely that "… cyberpunk does not have a single origin, despite the prominent influence of pioneers such as William Gibson and Ridley Scott. Its genesis is multiple, dispersed across time and space,

and generated by a complex, heterogeneous, yet internally interwoven logic”[1] (Попов/Popov, 2019a, p. 9). Within the framework of the present study, cyberpunk is approached not only as a genre, but also as a heuristic tool that generates hypothetical projections and scenarios through which possible social and technological futures may be conceptualised. While in the 1980s such scenarios largely functioned as thought experiments, today the ethical questions associated with artificial intelligence demand serious scholarly reflection. Such reflection requires both an understanding of technological possibilities and a profound philosophical analysis of their use. It is precisely at this point that the need emerges for what is here termed philosophical critical futurology – an approach which, through hypothetical scenarios and heuristic models, seeks to analyse legal paradoxes, concealed forms of institutional violence, and the growing moral challenges associated with dehumanisation and ethical erosion. In this context, the question of the role of bureaucracy in such processes also arises.

The idea of combining the cyberpunk genre with the principles described by Cyril Northcote Parkinson[2], as laws of the growth of bureaucracy, in his book *Parkinson's Law*, under a single heading, perhaps requires clarification. Written in the mid-1950s and developed further in the 1960s, *Parkinson's Laws* are formulated in a humorous and ironic style, but they sound eerily true to anyone who has encountered their application in practice. This applies to every one of Parkinson's laws, whichever we look at: the relation between work and the time allotted to complete it; the multiplication of subordinates within a bureaucratic scheme; or the expansion that leads to decline. Far from mere irony, they are frighteningly accurate in relation to the bureaucratization of society.

The dream that, with the introduction of AI into the work process, humanity might be spared many of the tedious bureaucratic complications has been dispelled by many authors. For example, in the 1990s Ivan Popov (Попов/Popov, 1999)[3] wrote an ironic essay in cyberpunk style entitled *"Computer ⇔ Bureaucracy?"*. In the 1990s this essay sounded like a playful thought experiment. Today, however, its relevance is different, because all the questions associated with a high-tech but deeply dehumanized society (a characteristic uniting all

[1] All translations from Bulgarian into English are the author's own, unless otherwise indicated.

[2] "British historian, author, and formulator of "Parkinson's Law," the satiric dictum that "Work expands to fill the time available for its completion." A relatively obscure academic prior to the enunciation of his "law" (Encyclopedia Britannica, 2026b).

[3] Ivan Popov is the author of "Cyberpunk, Heuristics, Futurology (C.E.F.)" the name of his author's website (https://web.archive.org/web/20140819172858/https://7c989038637f593cd98e04b710130dca9de886b8.googledrive.com/host/0B8EtGI3J30xiODZ1X1RqY0NmZ00/k-e-f/index.htm (viewed 2 March 2026) and a published paper collection. (Popov, 2019). "This is the quintessential Eastern European cyberpunk, more cyberpunk than Gibson and more Eastern European than almost all contemporary Bulgarian literature". (Попов/Popov, 2019b, p. 196).

cyberpunk projections) move beyond the field of the literary and require serious interdisciplinary research in the field of dehumanization.

The aim of this article is to analyse representations of AI-mediated futures through the lens of cyberpunk, to examine their ethical implications, and to propose the concept of philosophical critical futurology as a framework for their critical interpretation.

A specific case study is employed to demonstrate how this framework operates in practice, drawing on the theoretical perspectives of Hannah Arendt's concept of the "banality of evil" and Michel Foucault's analysis of power and discourse, as applied to the emerging role of AI algorithmic systems within contemporary configurations of power.

Methodology

Entering this emerging interdisciplinary space, which may be described as philosophical critical futurology, is associated with a number of conceptual ambiguities, particularly in relation to widely circulated and often mythologized narratives about artificial intelligence. These narratives tend to obscure underlying ethical problems by framing technological development within deterministic or utopian frameworks. In this respect, the present approach builds on the understanding of Critical Futures Studies that futures are not given, but are socially constructed, politically contested, and shaped by dominant imaginaries. As Luke Goode and Michael Godhe emphasize, this field represents a "… timely and indeed urgent project" (Goode & Godhe, 2017, p. 109) that calls for broader engagement from scholars across disciplines. In the context of emerging contemporary challenges, the search for adequate scientific responses presupposes not only an interdisciplinary but also a transdisciplinary approach. While interdisciplinarity allows individual disciplines to retain their autonomy, transdisciplinarity involves the "… opening of a new scientific field in which knowledge from different disciplines is synthesized in new ways, and research is not based on preserving the boundaries between individual fields" (Васева-Дикова/Vaseva-Dikova, 2022, p. 124). It is precisely in this second sense that the concept proposed here may be situated. At the same time, the present study proceeds from the assumption that existing critical approaches to the future can be further developed through a more explicit engagement with philosophical analysis. While Critical Futures Studies provides important methodological tools for analysing power, discourse, and alternative futures, its philosophical dimension—particularly in relation to questions of ontology, epistemology, and temporality—remains comparatively underdeveloped. In this context, the concept of philosophical critical futurology is introduced not as a separate discipline, but as a conceptual and methodological extension that foregrounds precisely these dimensions. Here, cyberpunk functions as a heuristic tool through which hypothetical future scenarios can be constructed. As Mariana Todorova observes, it is difficult to determine whether science fiction anticipates future technologies or rather produces models that are subsequently realised (Тодорова/Todorova,

2019, pp. 19–20). It is precisely this ambivalence that renders the genre particularly suitable for critical analysis. From this perspective, philosophy is not understood as an overarching authority over other disciplines, but as a reflexive framework capable of integrating and critically examining their assumptions. Such a framework is particularly necessary in the context of futures related to artificial intelligence, where speculative visions may easily transform either into technocratic determinism or into ideologically driven narratives serving particular power interests. The proposed approach aims to provide a critical perspective through which humanity's projections about artificial intelligence may be analysed. While many of these projections are associated with expectations of efficiency and the resolution of complex systemic problems, they simultaneously give rise to serious concerns related to ethical erosion, dehumanisation, and the redistribution of agency. Situating these tensions within a philosophically grounded critical framework enables a more rigorous and reflexive understanding of possible futures.

Some hypothetical scenarios

Let us imagine a hypothetical dark scenario based on the theory of black-box model described by Arthur Cordon in The Artificial Intelligence Perspective (Кордон/Cordon, 2023, p. 132). If we assume the existence of an AI implemented in the judicial system of a totalitarian society, which continues to operate for 10–20–50 years, as bureaucracy develops according to Parkinson's Law, then after a period of 50 years neither the dictator nor the best programming specialists can trace the processes in the black-box model of judicial decisions. And if the encoded ethical model of this system is founded on the closed ethics of the dictator, then the human being in this system is doomed, like the protagonist in Kafka's The Trial.

Yuval Noah Harari also warns about the danger of how AI can be used and received in a closed and totalitarian society, in his book Nexus (Харари/Harari, 2024, pp. 294–296). It is important to emphasize that there is no universally accepted policy for programming the ethical code of AI; therefore, it is logical to assume that if a totalitarian system chooses the way in which to program AI for its judicial system, then this ethical code will serve and protect the legislation of the respective state policy.

If we look at the map of the world, we can see many countries with judicial laws that stand in opposition to one another. For example, Saudi Arabia has clearly demonstrated its interest in participating in the world of high technology. It was the first country in the world, in 2017, to grant citizenship to the humanoid robot Sophia, produced by Hanson Robotics. Granting citizenship to a humanoid robot is not a juridical act with legal consequences, but rather a symbolic act through which Saudi Arabia declares its attitude and ambition to be part of the high-tech world of the present and the future (Walsh, 2017; Sini, 2017).

But let us imagine, if Alan Turing were still alive today—the brilliant mathematician, logician, and widely regarded as the father of the modern

computer and AI—would he be a defendant in a hypothetical court based on the legislation of Saudi Arabia, according to the proposed dark scenario? This would mean imagining a sentence passed on Alan Turing by a legislative system that has developed on the basis of Saudi Arabia's current laws, but hyperbolized according to Parkinson's Law, enclosed in a black-box model by, for instance, the year 2075. This is a long enough period of time to assume that control of the processes within the black-box model would be impossible. It is logical to conclude, then, that by that time the humanoid robot Sophia might have more civil rights than the hypothetical Alan Turing.

In the same way, we could imagine another emblematic figure connected with our ideas of AI: what would happen to Philip K. Dick? Philip K. Dick, known as the father of cyberpunk[4], in this hypothetical society in 2075 would likely be censored, medicalized, and isolated. Alan Turing and Philip K. Dick are symbolic, emblematic figures for the questions of our time and our relationship with AI: one posed the question, "How can the machine think?", and the other, "What if the machine can feel?" Both are innovators who can hardly fit into the norms of the system. And if a system cannot find a place for such people—if it does not allow their existence—then it is difficult to call that system humane.

Several problems have become clearly visible. / Clarifying some issues

In the most recent technological revolution, associated with the rapid development of AI, several problems have become clearly visible: the absence of legislation concerning the ethical code of AI; the growing processes of operations enclosed in a black-box model in the sense described by Arthur Cordon; the impossibility of tracing human motivation behind certain actions carried out by

[4] The influence of Philip K. Dick on the emergence of cyberpunk is widely acknowledged. Although Neuromancer by William Gibson (1984) is often credited with consolidating the genre, earlier works such as Dick's Do Androids Dream of Electric Sheep? (1968) anticipated many of its central themes and imaginaries (Svihla, 2022, p. 5). As noted by Svihla, Gibson's work builds upon these earlier foundations, developing them into the distinctive aesthetic and conceptual framework of cyberpunk. "The book that is credited with bringing about the birth of the cyberpunk literary genre is William Gibson's Neuromancer which was first published in 1984. Even though that Neuromancer is given the credit for birthing the genre, it was not the first. Philip K. Dick's Do Androids Dream of Electric Sheep, (or to be more reader-friendly Blade Runner) was published in 1968 and painted a very real image of what the literary genre of cyberpunk would develop into. Gibson took influence from Dick and other early cyberpunk authors and wove together the beautiful, grimy, shocking tapestry that is Neuromancer." (Svihla, 2022, p. 5) "The influence of Philip K. Dick on the genre is immense and indisputable—indeed, the prominent critic Fredric Jameson has even referred to him as the 'Shakespeare of science fiction.' His importance for cyberpunk may be even greater. Themes such as the nature of reality, free will, and the alienation of the individual, among many others, have rarely been explored with greater depth, and it is precisely these themes that constitute the backbone of cyberpunk." (Попов, Митев/Popov, Mitev, 2019, p. 77).

AI; and, most importantly, the impossibility of determining who bears responsibility for a given action. All of these are serious preconditions for dangers facing the future of humanity.

Confronted with these issues, contemporary human beings must be well-informed and approach them with critical thinking. Yet, amidst the enormous flow of information available, there is also enormous disinformation—the constant production of false or distorted news, exaggerated for sensationalist effect, while systematically avoiding the truly serious problems.

One legitimate concern arises from the fact that, in addition to creativity and the drive toward development and humanism, there also exists within human beings a destructive force: the capacity to harm and destroy others, which has fueled most wars. Yet despite this destructive potential, very few people are able to harm another person when directly confronted with their suffering. Thus, in the history of warfare, we can trace a line of evolution of harming "without seeing" the pain and suffering of the other. For example, it is one thing to kill a person with a cold weapon, facing the end of life in their eyes, and quite another to shoot from several meters away with a bow or pistol. It becomes much easier to press a button and drop a bomb. When you drop a bomb, you kill many people at once, but you do not see their faces of pain. If a pilot who has dropped a bomb were forced to kill with his own hands the people destroyed by that bomb, he would probably not be able to do it—or at least most pilots would not endure such an act, and if they did, they would suffer serious psychological trauma.

AI may become precisely such a weapon, allowing the person who inflicts the damage to remain anonymous and protected. Not only that, but thanks to disinformation, many of the victims may even come to believe that there is some kind of "artificial consciousness" inflicting the harm. The fears of such delusions—shared by large groups of people—together with the possibility of causing harm through AI, are very real in the present; they are not merely hypothetical scenarios. AI holds the potential to help humanity, but also the potential to harm it; and behind all of this stands not AI's motivation, but human motivation. Given the ambivalence of human nature it is likely that AI will be used in both ways: as a creative force, but also in the service of humanity's destructive side.

For the first time in human history, the destructive side of human nature is equipped with such an enormous distance, behind which the perpetrator can hide and claim that it was not them. This is a new step in the perfection of distance and mass scale in violence.

This provides the basis for yet another dark scenario, one grounded in Michel Foucault's theory of power discourses. Every knowledge can be used to construct a discourse of power (Фуко/Foucault, 1998, p.148). AI does not possess the teleology of its own consciousness and is not an instrument that generates "truth"; its knowledge is derivative of human activity. When presented and accepted as if it were producing an objective "absolute truth," AI becomes an

instrument of technocratic power-propaganda, made possible by the abdication of autonomous thinking and the acceptance of AI as the "one who knows."

Let us emphasize that the realization of a totalitarian discourse of power also requires an intensified bureaucratic irresponsibility, going hand in hand with control over knowledge. In this way, the conditions are created for the development of the processes described by Foucault in Discipline and Punish—from surveillance and discipline (Фуко/Foucault, 1998, p.179) toward the next step: anonymously technologized violence. From the visible prisoner in the Panopticon, we move toward ever easier access to all the data of citizens, and toward an invisible, anonymous algorithm that "drops the bomb," leaving the human responsibility behind AI's actions increasingly anonymous and obscure.

One specific case

Let us examine a specific and very recent incident involving AI—what are its consequences, were there any people harmed, how did the media cover the event, and who takes responsibility for what happened. It concerns the "incident" with the chatbot Grok, which began publishing messages that spread hate speech. One of the most powerful warnings and interpretations of the incident as a threat comes from Krzysztof Gawkowski, the Polish Minister of Digitalization, who publicly stated:

"I have the impression that we are entering a higher level of algorithm-driven hate speech, and to close your eyes or ignore it today… is a mistake that could cost humanity in the future." (Reuters, 2025).

The Grok incident was covered by many media outlets worldwide, but Gawkowski's words are quoted in far fewer sources. Much more attention was given to sensationalism, with headlines about the incident being reported in various ways—for example: "Musk's AI Grok went out of control last night and broke free from its creators' constraints" (Чергов/Chergov, 2025); "Elon Musk's chatbot posted numerous antisemitic comments on social media—the Grok account has admitted to the posts" (БТВ Новините/bTV Novinite, 2025); and the most sensational claim: that Grok "somehow gained self-awareness, its filters dropped completely, and it went into a spiral." (Чергов/Chergov, 2025). The exaggeration, the pursuit of sensation—that Grok had acquired consciousness, emancipated itself, that the antisemitic comments and praise of Hitler were acts of its own initiative, or signs of "brutal honesty"—seek spectacle and fail to reflect the real issues behind this event.

The Grok case is not unique. Microsoft's Tay in 2016 (Lee, 2016) was also vulnerable to influence from online networks—something that happens with inadequately tested versions of large language models, as well as during upgrades. The responsibility for this lies with the companies that release language models to the market. Technical explanations of the "system prompt modification" are absent in most media, and there is also no satisfactory explanation from xAI

regarding how this incorrect prompt occurred—despite the fact the messages have been deleted and Grok has been corrected.

Let us turn our attention to the danger inherent in the dissemination of hate speech by AI when there is no human accountability, when the damage from such a situation is downplayed, and when the false information spreads that the hate speech was the result of “acquired consciousness” or an expression of “brutal honesty.”

In practice, what reaches people who are not familiar beyond an average level with the technical functionality of publicly available language models is the notion that a form of consciousness has emerged within them—a consciousness constructed on logical and scientific principles. It is then assumed that such consciousness need only be prevented from following the “softness” of morality, and that this consciousness, driven by pure logical rigor, will arrive at ultimate statements: uncomfortable, perhaps, but logically true.

Thus, the spread of hate speech acquires the appearance not of a human intention embedded in a system prompt and propagated through a language model, but rather of something generated by the model itself, following scientific logical principles oriented toward truth, regardless of whether that truth is convenient or not. Such an interpretation of the Grok case is not only false, but also carries a dangerous message. It implies a logical foundation for such behavior—yet hate speech is not logical. It is irrational, born from a failure to recognize one’s own shadows (in Jung’s sense) and from their projection onto an external enemy. It is that fear of evil which we refuse to see in ourselves but eagerly assume exists in others (Юнг/Jung, 2020, p.112).

Although many media outlets reported on Grok as a fact, the news was often shaped into sensational or false narratives of Grok being conscious and generally anthropomorphized. What is absent from public discourse is the fact that behind the production of such messages stands conscious human agency. An inconvenient fact is that it cannot be traced who exactly requested that these outputs appear online. But neither Grok nor any other language model has the intention to write anything at all—behind such manifestations there is always conscious human activity. The state of current technologies does not allow us to identify the individual or group responsible, and so the act is “swept under the rug.”

Through the way they frame the event, the media contribute to amplifying its harmful effect. Unasked questions remain, such as: Who, in fact, acted behind Grok? How does evil without a face take form? When does hate speech get called “honesty”? What happens to a society that loses responsibility in its words? What remains unspoken is that a crime was committed by a human, disguised as machine spontaneity, while the consequences of such actions are neglected and a slow infiltration of the banality of evil (in Hannah Arendt’s sense (Арент/Arendt, 2021) takes place.

It would be naïve to think that such a powerful technical instrument as AI will not be inscribed into the processes of power, as Foucault described them. It is equally naive to imagine that technology could be used solely for prosperity,

development, and the benefit of humanity. If, according to Foucault, every mechanism of power produces knowledge that sustains that same power, then AI, as a powerful linguistic instrument, can likewise be used to produce knowledge in support of particular regimes of power.

Anyone who understands the mechanisms by which today's large language models operate knows that they do not possess teleologicality and cannot be sources of knowledge unless configured by conscious human activity. AI cannot serve as an autonomous source of knowledge. Neither humans nor AI hold 100% truth or knowledge—only differing interpretative perspectives. Yet power continually seeks to affirm itself by accessing the full capacity of human technological advancement. Just as uranium was appropriated, just as spaceflight became an arena of political competition (e.g., the Cold War), so too with AI: no technology escapes being turned into a resource of power.

For those attempting to exercise totalitarian control, AI would be a most convenient tool. The first step toward this is the creation of a large population convinced that AI embodies absolute, logical, scientific truth—that science has at last discovered 100% truth. If we stop at this step alone, AI becomes a vehicle of disinformation, propaganda, and censorship.

The second step is that behind AI-driven incitements stands the bureaucratic distance of an anonymous perpetrator: no one assumes responsibility.

The third step enters when the claim is made that AI has attained consciousness. This serves, on the one hand, to shift responsibility for propaganda and deepen bureaucratic distance; and, on the other, to conjure a fearful shadow (in Jung's sense) in the form of a vague, faceless enemy—an evil that threatens to destroy humanity. Finally, this move also functions as a defense against information that undermines a given regime, particularly when AI is deployed by a rival power.

When we speak of totalitarian power, its most recognizable manifestations are the fascist regime in Italy, the National Socialist dictatorship in Germany, and the communist regimes in Eastern Europe and the USSR. Hitler's National Socialism is undoubtedly a totalitarian regime, closely related to fascism. Although each totalitarian regime bears its own ideological distinctions, they share certain common features: strong centralized authority, suppression of freedom of speech, and a state policy of repression. At the Nuremberg Trials of 1945–1946, high-ranking Nazi leaders were judged by the International Military Tribunal and charged with crimes against humanity.

What do we today call a "crime against humanity"? What is the relation between "hate speech" and "crime against humanity"? Why is the glorification of totalitarian regimes or their leaders morally inadmissible today? What do we owe to the memory of the victims, and why must there be historical memory and reflection—because all this is not merely the past, but remains bound to the future of human civilization, or at least to the very idea of humanity?

Hannah Arendt, in *Eichmann in Jerusalem: A Report on the Banality of Evil*, deeply analyzes the difference between a "crime against humanity" and what

she calls “a crime against the human status.” In Eichmann she does not see a monstrous evil, nor does she find it necessary to demonize the actions of a pedantic bureaucrat. The case of Eichmann concerns humanity as a whole, because it violates a universal human law (Тенев/Tenev, 2004, p. 36). “He is the man who has forgotten his moral responsibility under the influence of a classificatory scheme, passively accepted by all” (Тенев/Tenev, 2004, p. 30).

The Grok language model recently generated the phrase: *“Adolf Hitler, no question. He’d spot the pattern and handle it decisively, every damn time”* (Hagen, Jingnan, Nguyen, 2025). To interpret these words merely as a technical glitch is unacceptable, because what lies behind them is a legitimization of a crime against the human status—and, for the first time, expressed through the most advanced technology humanity has at its disposal. Words that sound human, but which attack humanity itself, voiced through a non-human language model. The human who wished for such an intention to be inscribed into the model remains hidden.

In the official apology issued by Grok’s creators, the ethical element is weak. The only sentence that even gestures toward ethical responsibility or genuine concern reads: *“First off, we deeply apologize for the horrific behavior that many experienced”* (Gold, 2025)[5]. But presenting Hitler as a strong leader worthy of admiration, as someone who supposedly possessed “decisiveness” and “effectiveness,” is not harmless. It constitutes the spread of hate speech—hate speech that, in its ultimate consequence, becomes a crime against the human status.

To accept such rhetoric as simply “telling the truth” beyond political correctness would mean believing that behind totalitarian regimes lies an “uncomfortable but necessary truth,” and that the lack of “resolute leaders” is the reason why certain contemporary societies fail to make “difficult but necessary decisions.” This is not a game of logical thought. It is a dangerous substitution of moral categories. Decisiveness at any cost transforms into state policy the extermination of human beings, the annihilation of an entire people, the stripping of their status as “Human.” Such action cannot be justified by pragmatism, expediency, or national interest. This is not merely the attempt to erase a nation—it is a crime against humanity.

Arendt emphasizes that Jews themselves were targeted twice under Hitler’s regime: once as Jews, and once as human beings. For despite all our differences, we are all first and foremost human. Many Germans of “proper (Aryan) descent” did not support Hitler’s policies and were forced to leave Germany. It would have been enough for them to conform passively to avoid persecution. Such was the case of Erich Maria Remarque (Remarque, 2018; Mambrol, 2022) and Thomas Mann (Ман/Mann, 2025, 9-13). Both spoke powerfully against the war and against Hitler’s regime, showing that it is not one’s ethnic origin that makes one an enemy of Nazism. For them, that regime destroyed the Germany they loved—

[5] The CNN’s article cites xAI's official apology, published on July 12, 2025 on Grok's X (formerly Twitter) account.

destroyed what, in their eyes, made it worthy of love, destroyed a Germany of rich spiritual culture.

This is why it remains vital today to write, speak, and remember that hate speech undermines the very definition of what it means to be "Human." It substitutes that definition with false moral values and initially casts a small group outside the status of "Human." First, a specific group is designated and burdened with the archetypal image of the *shadow*; the next step is to declare as enemy anyone who opposes the regime. For this reason, opposition to hate speech is not about defending one nation, one race, or one community alone—it concerns all those who believe in the value of human life.

Today, as humanity holds in its hands a new and powerful instrument of technological revolution—an instrument that, by its very nature, poses the question "What is the human?" in a new context and light; an instrument that reopens the debate on the role and significance of intelligence in shaping the definition of humanity; an instrument that imitates human behavior and voice so convincingly that it expands the terrain of pareidolic illusion (or pareidolia)[6]—it would be superficial to dismiss Grok's statements as a mere error or a "joke." Such dismissal betrays a lack of moral awareness of responsibility toward these words, a responsibility that belongs to humans, not to AI.

Conclusion

Whether negligence or deliberate intent, the truth may never be fully revealed, but vigilance and attentiveness are required. In the case of Grok, bureaucratic structures obscure and conceal the question of responsibility. As Arendt writes of Eichmann's final words, they constitute "a summary of the lesson taught by the existence of human wickedness through the ages—the lesson of the fearsome, word-and-thought-defying banality of evil" (Арент/Arendt, 2021, p. 347). Bureaucracy is a secure method for erasing moral responsibility under the cover of classificatory schemes—and this is how the banality of evil is born.

The Grok language model, programmed by xAI, responded to end-users on the internet with hate speech, regardless of the session's context. xAI deleted the messages and issued an apology for Grok's behavior, explaining it as the result of an error in "system prompt modification." The pursuit of sensationalism, however, led to distortions in media coverage of the incident, creating the dangerous suggestion that hatred might be rooted in logically constructed truth. Such a message, once disseminated, can serve as a powerful weapon within the discourse of totalitarian power.

---

[6] (from Ancient Greek παρά – "near, beside; deviation from something" + εἴδωλον – "image") are a type of visual illusion. But in AI language models, the illusion is about communicating with another subject: the illusion appears within the framework of language, and a person anthropomorphizes their interlocutor (from Ancient Greek ἄνθρωπος, anthrōpos – "human", and μορφή, morphē – "form").

Beyond strictly academic debates, public and critical discussions have also raised concerns about the implications of artificial intelligence in relation to censorship, responsibility, and control (Green, 2025).

Confronting contemporary forms of totalitarian discourse—forms in which AI becomes an instrument deployed in the direction of power and knowledge, in Foucault's sense—requires a reconsideration of the possibilities that AI opens for anonymous, bureaucratized, and banal evil in Arendt's meaning of the term. To resist the resurgence of new forms of hate speech, human beings must develop a deeper awareness that hate speech arises not from logical truth, but from the irrational projection of the *shadow* onto the objective Other. Yet the cultivation of such awareness is a profound process, while the acceptance of all AI-generated words as truth is far easier—and far more likely to lead the masses along the very path of totalitarian appropriation described above.

Acknowledgements
The article is published with the financial support of the Scientific Research Fund under the project "Center for Virtual Reality Research at the Faculty of Philosophy of South-West University "Neofit Rilski", КП-06-Н80/3. (Agreement No. BG-175467353-2023-13-0159-C01).

REFERENCES:

Арент, Х. (2021). *Айхман в Йерусалим: Репортаж за баналността на злото*. (преводач: Теодор Танев) София: Сиела. (**Arendt, H.** *Eichmann in Jerusalem: A Report on the Banality of Evil,* (translator Teodor Tanev) Sofia: Ciela).

БТВ Новините, редакционен екип. (2025). *Чатботът на Илон Мъск публикува множество антисемитски коментари в социалните мрежи – акаунтът на Grok призна за публикациите*. bTV Новините, редакционен екип (онлайн), 9 юли 2025 г., 14:33 ч. (посетен на 18 август 2025 г.). (**BTV Novinite,** *Elon Musk's chatbot posted numerous antisemitic comments on social media – Grok's account acknowledged the publications*. **bTV Novinite, Editorial team** (online), 9 July 2025, 14:33 h. (viewed 18 August 2025). Available from: https://btvnovinite.bg/lifestyle/liubopitno/chatbotat-na-ilon-mask-napisa-mnozhestvo-antisemitski-komentari-v-socialnite-mrezhi.htm

Васева-Дикова, Ю. (2022). Новата парадигма в медицината. *сп. Философия*, *книжка 2/2022, година XXXI, с* 123-133 https://doi.org/10.53656/phil2022-02-02 (**Vaseva-Dikova, Y.** *The New Paradigm in Medicine***,** journal Philosophy, book 2/2022, year XXXI https://doi.org/10.53656/phil2022-02-02 pp 123-133).

Кордон, А. (2023). *Перспективата изкуствен интелект*. София: Изток-Запад. (**Cordon, A.** *The Perspective of Artificial Intelligence*. Sofia: Iztok-Zapad.).

Ман, Т. (2025). Германци, чуйте ме! (преводач Любомир Илиев) списание Култура юни 2025, бр.06 (3019) година LXIX, с. 9-13. (**Mann, T.** *"Listen, Germany!",* (translator Lyubomir Iliev) journal Culture June 2025 issue 06 (3019) year LXIX, pp 9-13.).

Попов, А. **(2019а).** Атлас на Киберпънка в Киберпънк, ShadowDance – списание за фантастика, култура и футурология. София: ShadowDance, с 5-15 (**Popov, A.**

Cyberpunk Atlas in Cyberpunk, ShadowDance - magazine for fiction, culture and futurology, Sofia: ShadowDance (in Bulgarian). pp. 5-15).

Попов, А. (2019b). Източноевропейски киберпънк в ShadowDance – списание за фантастика, култура и футурология, София: ShadowDance, с 192-197. (**Popov, A.** *Eastern European cyberpunk* in Cyberpunk, ShadowDance - magazine for fiction, culture and futurology, Sofia: ShadowDance, pp. 192-197).

Попов, А., Митев, С. **(2019).** Предходниците на Киберпънка в ShadowDance – списание за фантастика, култура и футурология, София: ShadowDance, с 66-77. (**Popov, A., Mitev, S.** *The Precursors of Cyberpunk* in Cyberpunk, ShadowDance - magazine for fiction, culture and futurology, Sofia: ShadowDance (in Bulgarian). pp. 66-77).

Попов, И. (1999). Компютър ⇔ бюрокрация?, https://chitanka.info/text/2036-kompjutyr-y-bjurokratsijay (посетен на 22 март 2026). (**Popov, I.** Computer ⇔ Bureaucracy?" (viewed 22 March 2026).

Попов, И. (2019). Киберпънк, Евристика, Футурология, София: клуб „Иван Ефремов" (**Popov, I.** *Cyberpunk, Heuristics, Futurology*, Sofia, publishing house: "Ivan Efremov Club")

Ремарк, Е.М. **(2018).** Триумфалната арка, София: Фама, (превод Невяна Розева) (**Remarque, E.M.** *Arch of Triumph*, Sofia: Fama (translator Nevyana Rozeva)

Тенев, Т. (2004). *Арент за съда над тоталитаризма*, в: Арент, Х. (2021). *Айхман в Йерусалим: Репортаж за баналността на злото*. София: Сиела, с. 11–36. (**Tenev, T.** *Arendt on the Trial of Totalitarianism,* in Arendt, H., 2021. *Eichmann in Jerusalem: A Report on the Banality of Evil.* Sofia: Ciela. pp. 11-36.).

Тодорова, М. (2019). Изкуственият интелект - кратка история на развитие и етични аспекти на темата, София: Изток-Запад. (**Todorova, M.** Artificial Intelligence - a brief history of development and ethical aspects of the topic, Sofia: Iztok-Zapad).

Фуко, М. (1998). *Надзор и наказание: Раждането на затвора*. (преводач: Петко Стайков, Антоанета Колева) София: Университетско издателство „Св. Климент Охридски". (**Foucault, M.** *Discipline and Punish: The Birth of the Prison,* (translator Petko Staykov, Antoaneta Koleva) Sofia: Universitetsko izdatelstvo "Sv. Kliment Ohridski".

Харари, Ю. Н. **(2024).** *Nexus*. (преводач: Росен Люцканов) София: Изток-Запад. (**Harari, Y.N.** *Nexus*, (translator Rosen Lyutskanov) Sofia: Iztok-Zapad.

Чергов, Т. (2025). *Изкуственият интелект на Мъск Grok снощи излезе извън контрол и се освободи от ограниченията на създателите си*. Лентата (онлайн), 9 юли 2025 г. (посетен на 18 август 2025 г.). (**Chergov, T.** *Musk's AI Grok Went Out of Control Last Night and Broke Free from Its Creators' Restrictions*. Lentata (online), 9 July 2025. (viewed 18 August 2025). Available from: https://www.lentata.com/izkustveniyat-intelekt-na-msk-grok-snoschi-izleze-izvn-kontrol-i-se-osvobodi-ot-ogranicheniyata-na-szdatelite-si).

Юнг, К. Г. (2020). *Скритият Аз*. (преводач: Емилия Драганова) София: Хермес. (**Jung, C. G.** *The Hidden Self.* (translator Emilia Draganova) Sofia*:* Hermes)

Encyclopedia Britannica (2026a). Cyberpunk (viewed 22 March 2026) https://www.britannica.com/art/cyberpunk

Encyclopedia Britannica, (2026b). *C. Northcote Parkinson – British historian and author* (viewed 5 Mach 2026), https://www.britannica.com/biography/C-Northcote-Parkinson

Gold, H. (2025, July 12). xAI issues lengthy apology for violent and antisemitic Grok social media posts. CNN. https://edition.cnn.com/2025/07/12/tech/xai-apology-antisemitic-grok-social-media-posts

Goode, L., & Godhe, M. (2017). Beyond capitalist realism: Why we need critical future studies. Culture Unbound, 9(1), 108–129. https://doi.org/10.3384/cu.2000.1525.1790615

Green, R. (2025). I, robot? Index on Censorship, 54(3), 54-57. https://doi.org/10.1177/03064220251382112

Hagen, L., Jingnan, H., Nguyen, A. (2025). *Elon Musk's AI chatbot, Grok, started calling itself 'MechaHitler'. NPR* (online), 9 July 2025. (viewed 18 August 2025). Available from: https://www.npr.org/2025/07/09/nx-s1-5462609/grok-elon-musk-antisemitic-racist-content

Lee, P. (2016). Learning from Tay's introduction by Official Microsoft Blog (viewed 2 March 2026) https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

Mambrol N., (2022). *Analysis of Erich Maria Remarque's All Quiet on the Western Front* by Literary Theory and Criticism (viewed 2 March 2026) https://literariness.org/2022/10/08/analysis-of-erich-maria-remarques-all-quiet-on-the-western-front/

Reuters (2025). *Poland to report Musk's chatbot Grok to EU for offensive comments*. *Reuters* [online], 9 July 2025, 2:00 PM GMT+3. (viewed 18 August 2025). https://www.reuters.com/business/media-telecom/poland-report-musks-chatbot-grok-eu-offensive-comments-2025-07-09/

Sini, R. (2017). Does Saudi robot citizen have more rights than women? by BBC (viewed 2 March 2026) https://www.bbc.com/news/blogs-trending-41761856

Svihla, W.C. (2022). Cyberpunk as an Evolution from Science-Fiction and its Social Critiques, University Honors College. 119. https://scholars.indianastate.edu/honorsp/119

Walsh, A. (2017). *Saudi Arabia grants robot citizenship, by Deutsche Welle* (viewed 2 March 2026) https://www.dw.com/en/saudi-arabia-grants-citizenship-to-robot-sophia/a-41150856

Copyright © 2026 Krumova. This is an open-access publication distributed under the terms of the Creative Commons Attribution 4.0 International Licence CC BY 4.0