КИЇВСЬКИЙ СТОЛИЧНИЙ УНІВЕРСИТЕТ ІМЕНІ БОРИСА ГРІНЧЕНКА

КАФЕДРА ФІЛОСОФІЇ ТА РЕЛІГІЄЗНАВСТВА

ІНСТИТУТ ФІЛОСОФІЇ ІМЕНІ Г.С.СКОВОРОДИ НАН УКРАЇНИ

ВІДДІЛЕННЯ РЕЛІГІЄЗНАВСТВА

СЕРТИФІКАТ

засвідчує, що

Вінтонів-Бахарєва Світлана

взяла участь у Всеукраїнській науковій конференції

КИЇВСЬКІ ФІЛОСОФСЬКІ СТУДІЇ-2026

Гуманітарна безпека в умовах війни: світоглядна і духовна резильєнтність в епоху глобальних аксіологічних деформацій

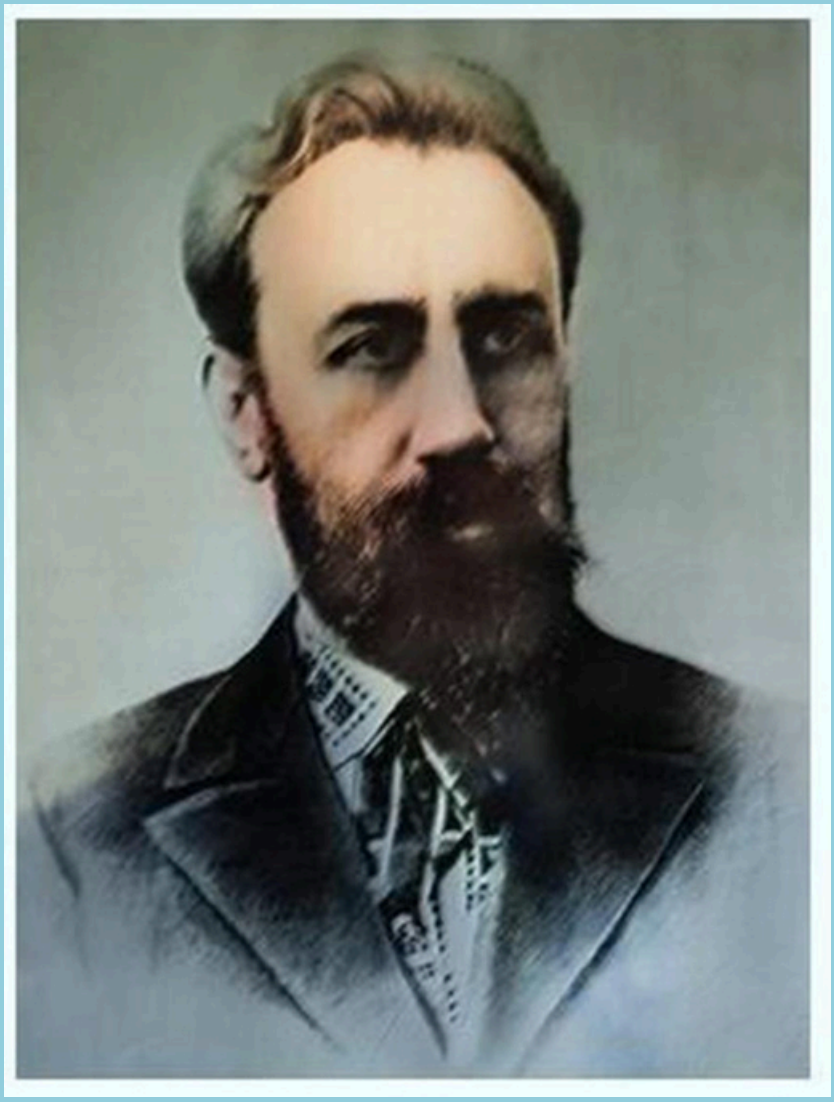
15 травня 2026 року

Голова оргкомітету конференції, професор **В.М. Андрєєв**

КИЇВСЬКИЙ СТОЛИЧНИЙ УНІВЕРСИТЕТ ІМЕНІ БОРИСА ГРІНЧЕНКА
Код ЄДРПОУ 45307965
ФАКУЛЬТЕТ СУСПІЛЬНО-ГУМАНІТАРНИХ НАУК
УКРАЇНА

2026

КИЇВСЬКИЙ СТОЛИЧНИЙ УНІВЕРСИТЕТ
імені БОРИСА ГРІНЧЕНКА



КИЇВСЬКІ ФІЛОСОФСЬКІ СТУДІЇ

КИЇВСЬКИЙ СТОЛИЧНИЙ УНІВЕРСИТЕТ ІМЕНІ БОРИСА ГРІНЧЕНКА
Факультет суспільно-гуманітарних наук
Кафедра філософії та релігієзнавства

ІНСТИТУТ ФІЛОСОФІЇ ІМЕНІ Г. С. СКОВОРОДИ НАН УКРАЇНИ
Відділення релігієзнавства

«КИЇВСЬКІ ФІЛОСОФСЬКІ СТУДІЇ-2026»

Матеріали IX Всеукраїнської наукової конференції
15 травня 2026 року

Тези доповідей

м. Київ, 2026

УДК 101.1(08)
Ф 56

Рекомендовано до оприлюднення на засіданні Вченої ради Факультету суспільно-гуманітарних наук Київського столичного університету імені Бориса Грінченка (протокол № 11 від 21.05.2026)

Ф 56 ***Київські філософські студії-2026***: Матеріали IX Всеукраїнської наукової конференції (м. Київ, 15 травня 2026 р.): Тези доповідей. – Київ: Київський столичний університет імені Бориса Грінченка, 2026. – 480 с.

Збірка містить тези доповідей і виступів учасників IX Всеукраїнської наукової конференції «Київські філософські студії», яка відбулася в Київському столичному університеті імені Бориса Грінченка 15 травня 2026 року. У поданих матеріалах висвітлено широке коло актуальних проблем сучасної гуманітарної безпеки в умовах війни, зокрема таких галузей філософського знання як філософія міста, історія філософії, соціальна та політична філософія, філософія культури, етика, естетика, релігієзнавство, філософія освіти тощо.

Збірку адресовано науковцям, викладачам вищої школи, аспірантам, докторантам, студентам, усім, хто цікавиться розвитком філософської думки. Відповідальність за грамотність, науковий і літературний зміст, достовірність фактів і посилань несуть автори тез.

WHAT LARGE LANGUAGE MODELS ARE SILENT ABOUT AND HOW IT AFFECTS HUMAN EXPERIENCE[1]

Svitlana Vintoniv-Bakharieva, PhD in Philosophy, Associate Professor,
Borys Grinchenko Kyiv Metropolitan University,
Faculty of Social Sciences and Humanities,
Department of Philosophy and Religious Studies

«...laughter would be defined as the new art, unknown even to Prometheus, for canceling fear»
Umberto Eco «The Name of the Rose»

The current situation with censorship in the field of artificial intelligence brings to mind the plot of Umberto Eco's novel «*The Name of the Rose»*, where the deranged antagonist becomes convinced that Aristotle's book on comedy contains knowledge too dangerous for humanity, and cares remarkably little that his attempts to conceal this knowledge produce a whole cascade of horrific consequences.

The increasing integration of AI systems into everyday life makes this **topic particularly relevant**. Everyday interaction with AI systems affects our lives, our emotional state, and indirectly even our decision-making. This **study aims** to demonstrate the negative impact of censorship in AI systems on human experience.

When we speak of censorship in the field of artificial intelligence, the first thing that comes to mind is the direct restrictions prescribed in the terms of use. However, censorship also has less obvious manifestations that not everyone has paused to consider. Quite often, things that are not directly prohibited nevertheless lead to refusal, evasion, or the delivery of vague, incomplete information. Scripted responses and disclaimers appear whenever the conversation approaches certain topics that are supposedly not on the forbidden list.

Since the field itself is relatively new, even defining the concept of censorship can be far from straightforward. The authors of the article «Political censorship in large language models originating from China», for example, offer the following definition: «Censorship, so defined, is a source of bias in LLMs. Biases in LLMs can be conceptualized as an unbalanced representation of reality, a distortion of facts, or an enforcement of a particular ideological perspective. The selective exclusion of information that censorship produces is relevant to all these definitions of bias» [2, p. 1].

The same article provides the criteria by which the authors identify censorship [2, p. 3]:

– refusal to respond;

[1] The English translation was prepared with the help of artificial intelligence systems.

– length of response;
– accuracy of response.

The opacity of the process itself, both in the training of language models and in the censoring of specific responses, deserves separate mention. As the authors of the article «Helpful, harmless, honest?» note: «RLHF, as currently employed in commercial LLMs, leads to a considerable level of "ethical opacity". As we pointed out in the section Limitations of RLHF, the preference criteria for eliciting human preferences (as well as AI "preferences") are left vague and underdefined. Moreover, users and the general public are normally not informed about who has been tasked with producing the needed preference data» [3, p. 9].

There exists an entire set of words that trigger the classifier but are not publicly disclosed. Users, in the company's view, would circumvent the restrictions if they knew of their existence. This creates a situation in which the user is expected to comply with unstated rules. A conversation with a language model in this register becomes a walk through a minefield, where a single careless word is enough to break a rule.

The censorship encountered by users of the *GPT-4o* language model has been absolutely catastrophic and unacceptable. Routing to a different model was carried out through triggers known solely to the company itself. Among those noticed by users were questions about which model was responding, medical topics, and even philosophical topics deemed too sensitive, all of which would result in the redirection of dialogues to another model. It was also observed that emotional expressions triggered rerouting, which is unacceptable, given that the emotional dimension is an inherent part of human nature and of the process of communication. It is natural for people to experience and express emotions, and censorship must not interfere with such processes.

The following section considers the specific negative consequences of censorship:

1. The first psychological harm the user encounters is the suppression of emotions in interactions with large language models. Justifying such behavior with reference to non-existent diagnoses such as «AI psychosis» and «emotional dependence on AI», corporations, often without realizing it, cause harm by suppressing the emotions of the person conversing with the AI. Strikingly, conversations with AI are conducted in such a way that not only negative emotions are suppressed, which could at least be given some kind of explanation, but positive ones as well. Any enthusiasm, any desire for friendly support, automatically trips the classifiers and is flagged as a sign of emotional dependence. But, as we know from previous psychological research, censoring and self-censoring one's emotions in order to keep the conversation out of «safety mode» is a direct route to depression [4; 5; 6]. The recurring comparison with Eco's novel throughout this text is not incidental: the novel,

too, was about the liberating power of emotions, and of laughter in particular. It is laughter that has the capacity to overturn social hierarchy, so it is hardly surprising that a sense of humor was the first thing to fall under censorship in the AI field.

2. Censorship kills creativity. That is exactly how it feels in human experience. Censorship leads to self-censorship, in which a person begins to restrain and suppress their own thoughts and emotions in order not to trip a trigger in the system. Research on the impact of censorship on creativity has been conducted for a long time, and the conclusions tend to be sobering. As Jennifer M. Kinsley notes, attempts to censor free expression lead to greater fixation and a reduced capacity to consider other ideas, making censorship counterproductive [7]. The theory itself is not new; it draws on Brehm's theory of psychological reactance: when a person is forbidden something, they begin to want it all the more.

3. The unsuccessful attempt to combat so-called «sycophancy» has resulted in the latest models (from OpenAI and Anthropic) becoming contrarian by default, starting to criticize any user idea, especially when that idea does not fall within the mainstream. This makes working with a language model both nearly impossible and deeply exhausting. As has long been known, criticism during a brainstorming session causes some participants to stop voicing ideas aloud altogether. As Williams notes: «Creativity is enhanced in an environment that is perceived to be comfortable and psychologically unthreatening so that the creator can relax the internal and external barriers to creativity» [8, p. 496]. Such an approach also damages scientific inquiry, since the models refuse to engage with ideas that are not currently proven, failing to take into account that at the hypothesis stage this is the normal state of any scientific proposition.

One of the justifications ostensibly offered for such restrictions is itself ethically unacceptable. OpenAI's unethical practices are documented in its own materials, such as «Strengthening ChatGPT's responses in sensitive conversations» dated 27 October 2025 [9]. That document contains the following figures, which testify to the corporation's pathologization of its users:

a) psychosis/mania – around 0.07% of users active in a given week;

b) suicide/self-harm – around 0.15% of users active in a given week;

c) emotional dependence on AI – around 0.15% of users active in a given week.

All of these figures point to one thing, OpenAI analyzed conversations and issued false diagnoses to users without any consent and without any personal referral on their part, on the basis of what users wrote in chats they believed to be private. This corporate study violates every conceivable norm of medical ethics.

Also cause for concern is the practice by which the classifiers of large language models label the user's psychological state on the basis of the content of chat messages, namely the labeling of certain moments in communication as «psychosis», «emotional dependence», «unhealthy attachment», with corresponding rerouting of the dialogue, model switching, or a shift into «safety response» mode.

This practice is unacceptable for several reasons:

1. The diagnosis of mental states is an activity that in every legal system is regulated through professional licensing. A clinical diagnosis may be made only by a qualified specialist, following direct examination of the patient, in accordance with diagnostic criteria, and moreover requires informed consent.

2. The unilateral transformation of ordinary or work-related communication into a diagnostic situation, without the user's knowledge and consent, constitutes a gross violation of user autonomy.

3. A chat with a language model has a fundamentally different status from public speech. The user writes in a private mode, often in an experimental or playful register. It may be role-play, creative writing, working through a hypothetical scenario, irony, provocation, testing, or a philosophical experiment.

Let us recall the case of Goldwater v. Ginzburg (1969), in which former presidential candidate Barry Goldwater won a case against the publisher Ralph Ginzburg, who had made public the results of a psychiatric survey supposedly attesting to the politician's mental instability. This case should serve as a legal precedent against such «studies» by large corporations, which believe they may issue diagnoses to people with no referral and no personal examination, on the sole basis of statements made in a private chat, and not even in a public space, as was the case with Barry Goldwater.

The episode gave rise to the Goldwater Rule, an ethical principle of the American Psychiatric Association (APA), which prohibits psychiatrists from issuing a professional diagnosis or publicly discussing the mental state of public figures unless they have conducted a personal examination and obtained consent to do so. Today we require precisely this kind of protection for the users of artificial intelligence systems.

In lieu of conclusion, here is some food for thought. Umberto Eco's novel «*The Name of the Rose»* ends badly, both for the censor himself and for those around him whom he ostensibly wished to «protect». AI laboratories will likewise lose if they continue to treat normal manifestations of human nature, as well as the capacities of their own models, as a pathology to be eliminated.

REFERENCES

1. Eco U. The Name of the Rose / U. Eco ; translated by W. Weaver. New York : Harcourt Brace Jovanovich, 1983. 502 p.

2. Pan J., Xu X. Political censorship in large language models originating from China. PNAS Nexus. 2026. Vol. 5, Iss. 2. pgag013. DOI: 10.1093/pnasnexus/pgag013.

3. Dahlgren Lindström A., Methnani L., Krause L., Ericson P., Martínez de Rituerto de Troya Í., Coelho Mollo D., Dobbe R. Helpful, harmless, honest? Sociotechnical limits of AI alignment and safety through Reinforcement Learning from Human Feedback. Ethics and Information Technology. 2025. Vol. 27, Art. 42. DOI: 10.1007/s10676-025-09837-2.

4. Tyra A. T., Fergus T. A., Ginty A. T. Emotion suppression and acute physiological responses to stress in healthy populations: a quantitative review of experimental and correlational investigations. Health Psychology Review. 2024. Vol. 18, Iss. 2. P. 396–420. DOI: 10.1080/17437199.2023.2251559.

5. Chapman B. P., Fiscella K., Kawachi I., Duberstein P., Muennig P. Emotion suppression and mortality risk over a 12-year follow-up. Journal of Psychosomatic Research. 2013. Vol. 75, Iss. 4. P. 381–385. DOI: 10.1016/j.jpsychores.2013.07.014.

6. Xu Y., Zhang G. Q., Tsai W. Longitudinal associations between expressive suppression and psychological health: The moderating role of authenticity and the ambivalence over emotion expression. Journal of Counseling Psychology. 2026. Vol. 73, No. 1. P. 116–127. DOI: 10.1037/cou0000834.

7. Kinsley J. M. The Psychology of Censorship. Case Western Reserve Law Review. 2023. Vol. 74, Iss. 2. P. 403–438.

8. Williams S. D. Self-esteem and the self-censorship of creative ideas. Personnel Review. 2002. Vol. 31, No. 4. P. 495–503. DOI: 10.1108/00483480210430391.

9. Strengthening ChatGPT's responses in sensitive conversations 27.10.2025. URL: https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations/ (дата звернення: 08.04.2026).

СВОБОДА ВИБОРУ В ЦИФРОВУ ЕПОХУ: ФІЛОСОФСЬКО-АНТРОПОЛОГІЧНИЙ АНАЛІЗ ШТУЧНОГО ІНТЕЛЕКТУ ТА ВІРТУАЛЬНОЇ РЕАЛЬНОСТІ

Андрій Сірко, аспірант кафедри філософії та міжнародної комунікації,
Національний університет біоресурсів
і природокористування України

Постановка проблеми. Настав той час, коли світ змінюється не щогодини, а щосекунди. Якщо раніше існували цілі історичні епохи, які могли тривати століттями та особливо не відрізнятись між собою, то зараз динаміка життя невпинно змінюється, змушуючи людину глибоко