AI & Propaganda: A Philosophical Workshop

Institute of Philosophy, Czech Academy of Sciences, Prague
Jilska 1, Praha 1
6–7 November 2025

Organizer: Institute of Philosophy of the Czech Academy of Sciences, Prague
Main Coordinator: Tomas Koblizek, Department of Analytic Philosophy
Supported by Strategie AV 21

If you are interested in participating, please write to koblizek [at] flu.cas.cz. The event is open to the public.

Keynote speakers:
Teresa Marques (University of Barcelona)
Neri Marsili (University of Turin)

Programme

6 November 2025
(meeting room of the Institute of Sociology, 2nd floor)

14-15.15 Keynote Speech
AI generated Images and the Power of Propaganda
Teresa Marques (University of Barcelona)
(chair: Tomas Koblizek)

15.15-16.00
Insidious LLM Propaganda
Nick Young (University of Genoa)
(chair: Tomas Koblizek)

Coffee break

16.15-17.00
The Evidence Machine: On Generative Artificial Intelligence and Disinformation
Keith Raymond Harris (University of Vienna)
(chair: Amelia Godber)

17.00-17.45

Undermining Democratic Deliberation?

Algorithmic Mindshaping and the Crisis of Epistemic Agency

John Dorsch (Czech Academy of Sciences)

(chair: Amelia Godber)

7 November 2025

(meeting room of the Institute of Philosophy, 1st floor)

10-11.15 Keynote speech

Rethinking Assertion in the Age of AI

Neri Marsili (University of Turin)

(chair: Teresa Marques)

Coffee break

11.30-12.15

Rethinking Manipulation: AI and Computational Propaganda

Tuğba Yoldaş (Czech Academy of Sciences)

(chair: Teresa Marques)

Lunch

14-14.45

AI, Propaganda, and X-Risks

Petr Jedlička (Czech Academy of Sciences)

(chair: John Dorsch)

14.45-15.30

A Conceptual Foundation for Training Large Language Models to Identify Propaganda

Amelia Godber

(chair: John Dorsch)

Coffee break

15.45-16.30

What should philosophers say about AI?

Glenn Anderau (University of Zürich)

(chair: Tomas Koblizek)

Abstracts

What should philosophers say about AI?

Glenn Anderau (University of Zürich)

The topic of AI has received a lot of attention in the philosophical literature recently. This paper aims to take stock of the current philosophical literature on AI and tries to answer the question what we should (not) expect from a philosophical discussion of the topic. Because of the vast variety of sub-disciplines for which AI is a topic of interest, I will focus on work on AI within epistemology, although some of the takeaways might be of interest to other philosophers as well. The paper will try to determine what a fruitful discussion of AI in epistemology should look like with regard to three main topics: 1.) Technical Knowledge 2.) Specificity 3.) Normativity. Regarding the first point, I will try to gauge how much (if any) understanding of AI on a technical level is necessary to gain philosophical insight into the topic. For the second point, the goal is to determine the merits of picking very specific examples compared to broader discussion of AI as a whole. I will also consider whether it is necessary to speak of AI as a whole when the topic discussed is simply a very specific form of AI (say LLM's). And the final aim of this paper is to determine how discussions of AI contribute to normative debates within epistemology. While descriptive discussions of AI are still valuable, a lack of genuine normative insights might lead us to reevaluate our approach to AI within epistemology.

Undermining Democratic Deliberation? Algorithmic Mindshaping and the Crisis of Epistemic Agency

John Dorsch (Czech Academy of Sciences)

Abstract

Does artificial intelligence (AI) threaten deliberative democracy? While early accounts blamed filter bubbles, and recent research shifts blame to user self-selection, this paper offers a new diagnosis: AI technologies, particularly recommender systems, propagate intellectual arrogance by optimizing content for cognitive fluency, which produces a metacognitive feeling of rightness. This metacognitive signal leads individuals to mistake ease of processing for epistemic justification, reinforcing confidence in their own beliefs while dismissing opposing views as irrational. The result is not ideological isolation, but contingent credibility asymmetry chambers, where disagreement persists but compromise collapses—fueling affective polarization. Integrating findings from cognitive science, epistemology, and the study of political discourse in algorithmic environments, we argue

that AI is not merely determining the content people see, but engaging in mindshaping—that is, shaping how individuals evaluate reasons, assign credibility, and engage in public reasoning—often yielding beliefs that are true only by epistemic luck, including reflective luck, where agents lack access to the justificatory basis of belief formation. It concludes by proposing a path forward: AI should be redesigned to explicitly introduce epistemic friction, while supporting reliable epistemic practices and cognitive affordances as a corrective.

A Conceptual Foundation for Training Large Language Models to Identify Propaganda

Amelia Godber

Generative AI enables diverse actors to rapidly produce manipulative and false information tailored to influence public opinion at scale. A growing body of analysis registers concern about the politico-epistemic ramifications of AI-generated propaganda, and points to a need to safeguard the epistemic conditions necessary for democratic deliberation. AI may be used as a propaganda tool, but it also holds promise as a means for detecting and debunking propaganda in public discourse. This paper aims to demonstrate how philosophical analysis can inform empirical evaluation by helping to lay a conceptual foundation for conditioning large language models to identify political propaganda. It proposes a theory of propaganda based on a typology of rhetorical strategies that subvert an audience's rational capacities, arguing that the concept avoids overgeneralising and undergeneralising while expressing two key throughlines in the philosophical literature (§1). §2 motivates the use of generative AI as a content moderation tool capable of both detecting propaganda and providing evidence-based counterspeech to support societal resilience to discourse that bears negatively on epistemic attainment. §3 sets out to show that the proposed concept yields a promptable rubric that a general large language model can reliably follow. The model is conditioned on the conceptual criteria and a set of labeled exemplars, then asked to classify unseen cases and justify its evaluations. These results are compared with those of a baseline test in which the model is given neither the rubric nor exemplars and is asked to classify the same cases without guidance. The findings suggest that the theory functions as an effective guide: it improves accuracy and yields superior reasons for the model's assessments, where these are marked by better evidence use, fewer unsupported claims and errors, and more consistent rationales across cases. Finally, §4 addresses limitations and potential concerns, and considers the scope of the paper's contribution to the nascent debate on AI and propaganda.

The Evidence Machine: On Generative Artificial Intelligence and Disinformation

Keith Harris (University of Vienna)

Generative artificial intelligence gives rise to a host of epistemic challenges related to its ability to rapidly produce fake but realistic answers, references, data, images, videos, and so on. How severe are these challenges, how do they contribute to disinformation campaigns, and how can they be mitigated? I argue, first, that some of the most dramatic epistemic consequences associated with generative artificial intelligence are likely to be muted. Some commentators have argued, for example, that generative artificial intelligence might be used for the purposes of mass deception or to drive skepticism toward genuine evidence. Some fears about such consequences, I argue, are based on an overly simplistic view of the epistemic and psychological force of evidence. Second, I argue that, despite limitations on its direct effects on individuals' beliefs, AI-generated content is likely to effectively support disinformation campaigns. I conclude with some remarks on how the epistemic ill-effects of generative AI likely can (and can't) be mitigated.

AI, Propaganda, and X-Risks

Petr Jedlička (Czech Academy of Sciences)

My presentation examines several pathways through which AI-enabled propaganda and disinformation can threaten society. First, I consider their growing role in existential risk assessments (e.g., Kasirzadeh 2024), particularly within "gradual disempowerment" scenarios that emphasize the slow accumulation of harms leading to social disruption, as opposed to abrupt X-risk trajectories associated with intelligence explosion (Yudkowsky 2013, Bostrom 2012). In these scenarios, AI's impact will hinge on its capabilities (AGI vs. ASI), number of actors (single dominant actor vs. multiple actors), and the extent of human involvement (full, partial, or none), the quality of system's alignment etc. I also analyze risk scenarios that explicitly model AI deployment within interstate arms races—most prominently between the United States and China (Kokotajlo et al. 2027). Building on this, I examine how the People's Republic of China has already integrated novel AI tools into its propaganda apparatus across traditional and digital channels, and how this aligns with the PRC's broader policy strategy.

AI generated images and the power of propaganda

Teresa Marques (University of Barcelona)

Hyska (2025, forthcoming) has recently argued that uses of AI in propaganda, such as in so-called 'deepfakes', undermine our capacity to acquire knowledge from shared media. Hyska further argues that, by interfering in our capacity to show each other information, deepfakes undermine "our relationships in the context of collective political action". Although I agree that an environment rich in fake images or videos can undermine knowledge, I dispute some of Hyska's assumptions, and their aptness to explain the effects and power of propaganda. Her argument relies on two crucial ideas: First, she assumes a broadly Gricean understanding of communication that relies on speaker's intentions. Second, her argument assumes a framework that is heavily focused on the transmission of information. I will present some challenges to both assumptions. Against the first, I argue that communication requires the existence of conventionalized practices that are operative independently of the actual intentions of communicators. Second, I argue that the effectiveness of propaganda does not depend on the veridically, or otherwise, of the communicative acts it exploits. It can contribute to strong relationships and collective political action. This is not a good in itself. Understanding propaganda requires attending to other dimensions of collectives, namely collective emotions and plans. This is consistent with the intrinsic defectiveness of propaganda. Whether relying on AI or not, it interferes in something more basic: our capacity to see reality and each other.

References:

Bonard, Constant; Contesi, Filippo & Marques, Teresa (2024). The Defectiveness of Propaganda. Philosophical Quarterly (4).

Fallis, Don (2021). The Epistemic Threat of Deepfakes. Philosophy & Technology, 34 (4):623–643.

Hyska, Megan (2025). The politics of past and future: synthetic media, showing, and telling. Philosophical Studies 182 (1):137-158.

Hyska, Megan (forthcoming). Deepfakes, Public Announcements, and Political Mobilization. In Tamar Szabó Gendler, John Hawthorne, Julianne Chung & Alex Worsnip, Oxford Studies in Epistemology, Vol. 8.

Matthews, T. (2023) Deepfakes, Fake Barns, and Knowledge from Videos. Synthese, 201(2):41.

Rini, Regina (2020). Deepfakes and the Epistemic Backstop. Philosophers' Imprint 20 (24):1-16.

Salmela, M., and Nagatsu (2016). Collective emotions and joint action: Beyond received and minimalist approaches. Journal of Social Ontology 2, 1–25

Rethinking Assertion in the Age of AI

Neri Marsili (University of Turin)

Can LLMs make genuine assertions? Academics disagree. A key driver of this dispute is an underlying disagreement about what asserting requires. Rather than defending a specific characterisation of assertion, I will review different proposals, showing how each can be extended from human communication to machine communication. Artificial utterances, I will argue, satisfy different conceptions of assertoric force to different degrees. This pluralist approach better acknowledges our conflicting intuitions about machine assertion: it explains precisely why we feel that there is some sense in which LLMs can assert, and some sense in which they don't.

Rethinking Manipulation: AI and Computational Propaganda

Tuğba Yoldaş (Czech Academy of Sciences)

We influence one another in a number of ways, some innocuous, others morally problematic. One prevalent type of influence is manipulation. Philosophical accounts disagree about what exactly counts as manipulation: there is no consensus on whether intent, deception, subversion of rational capacities, or the outcome is essential (Coons & Weber, 2014; Jongepier & Klenk, 2022; Noggle, 2025). Questions concerning how to distinguish manipulation from other forms of influence, such as coercion, persuasion, or nudging, whether it is always morally problematic, and if so, under what conditions, are amplified by the rise of computational propaganda. Computational propaganda is a specific form of AI-mediated influence that involves manipulation for political, ideological, commercial, or other purposes. It has been defined as the "use of algorithms, automation, and human curation to purposefully distribute misleading information over social media networks" (Woolley & Howard, 2018). AI-enabled propaganda increasingly relies on synthetic media, algorithmic amplification, and microtargeting, which need not involve lying, identifiable agents, or deliberate communicative intent. As a result, it combines features of persuasion, manipulation, deception, and misinformation in ways that challenge traditional distinctions. I argue that AI-driven propaganda pressures two assumptions underlying the main philosophical accounts of manipulation. First, it unsettles intent-based views that treat propagandistic manipulation as tied to hidden motives or agent-directed deception. Second, it challenges process-based accounts because algorithmic targeting can shape attention and cognition without clearly bypassing or fully engaging rational deliberation. Rather than resolving these tensions, I aim to show how AI-mediated propaganda exposes them and why philosophical analyses of propaganda must be revised to be action-guiding.

Insidious LLM Propaganda

Nick Young (University of Genoa)

It is clear that generative AI can be used to create propaganda. In May 2025, we saw a crude example: system prompt tampering led XAI's chatbot Grok to insert comments about alleged "white genocide in South Africa" (Preston, 2025) in its responses to completely unrelated questions, such as those about cats or cartoons. In this talk, I consider how LLMs might become instruments of propaganda in more insidious ways. I argue that, because it is unlikely that LLMs are intentional agents in any interesting way, it is implausible to think of them as propagandists. If this is correct, then treating them as such may not be the optimal way to produce propaganda with an LLM. I then consider two training phases of modern LLMs: pre-training and post-training (e.g. alignment training). I suggest that post-training may not be able to produce propaganda with a fine enough grain as to be very effective, and argue instead that pre-training is likely more effective because it leads to a particular type of semiotic media. By this I mean that pre-trained LLMs are grounded on learned statistical patterns—relationships between concepts (paradigmatic) and sequences (syntagmatic) extracted from training data. Such a medium could be corrupted during pre-training through selective corpus design: adjusting co-occurrence frequencies, privileging certain framings, or systematically excluding alternatives. Such manipulation would embed bias directly into the model's probability distributions, not just its surface behaviour. The result would not be overt propaganda, but systematic tendencies that privilege certain ideas and frames by making them more salient. Therefore, the most effective method of producing LLM propaganda may well be to target the pre-training substrate where meaning-making patterns are first formed.

References

Preston, D. (2025, May 16). Grok's white genocide fixation caused by 'unauthorized modification'. The Verge. https://www.theverge.com/news/668220/grok-white-genocide-south-africa-xai-unauthorized-modification-employee