

Critical AI Safety Workshop (CAIS)

Program + Abstracts

Workshop organized by the Center for Philosophy of AI at the University of Copenhagen,
with generous support from Carlsbergfondet.

1-2 October 2026

Panum Building, Blegdamsvej 3B

2200 Copenhagen, Denmark



Organizers:

Ninell Oldenburg, niol@hum.ku.dk

Nina Rajcic, nira@hum.ku.dk

Ilias Chalkidis, ilias.chalkidis@di.ku.dk

Filippos Stamatiou, fs@hum.ku.dk

Iwan Williams, iwi@hum.ku.dk

Sandrine Schiller, shan@hum.ku.dk

Bokar N'Diaye, b.l.ndiaye@uva.nl

Schedule

Day 1: Thursday, October 1st

Panum Building, Blegdamsvej 3B, Room 7.15.92 (floor 15, room 92)

09:00	Welcome and introductions
09:15	Keynote I: Shazeda Ahmed, University of California, Los Angeles <i>Humanity's Last FAQ: Envisioning an Interdisciplinary Research Agenda for Critical Studies of AI Safety</i>
10:15	Coffee break
10:30	Presentations I — Anthropomorphism & the Object of AI Safety 10:30 — Alexandru Mateescu, Université Paris 1 Panthéon-Sorbonne: <i>How AI Safety Defines Its Object of Inquiry</i> 10:55 — Mariana Lins Costa, State University of Ceará: <i>Methodological Anthropomorphism in AI Safety Evaluations</i> 11:20 — Eugene Kim, University of Michigan: <i>AI Safety as an Instrument of AI Hegemony</i>
11:45	Coffee & snack break
12:00	Presentations II — Philosophy, Goals, and Critical AI Safety 12:00 — Hin-Yan Liu, University of Copenhagen: <i>Triangulating the Construction of AI's Existential Risk</i> 12:25 — Asher Kessler, University of Manchester: <i>Waiting For the Pause: Interruption, Acceleration and the Limits of (Critical) AI Safety</i>

	12:50 — Eloise Soulier, University of Cambridge: <i>Critical AI Safety: What is the role of philosophy?</i>
13:15	Lunch at the venue
14:30	Presentations III — AI Risks and Agency 14:30 — Sun Yizhou, Free University of Berlin: <i>Cognitive Outsourcing: Generative Asymmetry, Material Drag, and Untraceability</i> 14:55 — Ilias Chalkidis, University of Copenhagen: <i>Brainrot: Deskilling and Addiction are Overlooked AI Risks</i> 15:20 — Nina Rajcic, University of Copenhagen: <i>Goal-Directedness is in the Eye of the Beholder</i>
15:45	Coffee break
16:00	Presentations IV — Power, Infrastructure, Political Economy 16:00 — Lukas Meier, Harvard Center for Ethics: <i>ICE, Gaza, and Cambridge Analytica: The Hidden Story of Palantir</i> 16:25 — Santtu Räsänen, Aarhus University: <i>Gatecrashers and Suspensions at the AI Safety Conference</i> 16:50 — Tomáš Filo, Charles University: <i>Gaslighting the Future: How AGI Narratives Shape Politics of AI</i>
17:15	Snack break
17:30 - 18:30	Keynote II — Adam Becker, Journalist and Author of <i>More, Everything, Forever</i>: <i>Fighting the Rhetoric of Inevitability: Puncturing Transhumanist Eschatology and Sketching Better Futures</i>

19:30 — Workshop Dinner at GORILLA

Flæsketorvet 63, 1711 København

Day 2: Friday, October 2nd

Panum Building, Blegdamsvej 3B, Room 7.15.92 & 7.15.152 from 13:30

09:00	Welcome and introductions
09:15	Keynote III — Mollie Gleiberman
10:15	Coffee break
10:30	Presentations V — Epistemic Community & Governance 10:30 — Meridian Monthly, University of Cambridge: <i>Who, or What, Governs Our Future? Classification, Legitimacy Infrastructure, and the AI Safety Community</i> 10:55 — Uma E. Sarkar, University of Cambridge: <i>The AI Safety Epistemic Community as a Governance Intermediary</i> 11:20 — Luca Neuperti, University of Edinburgh: <i>Mixed-methods analysis of the discursive framings of "AI" on LessWrong and the Alignment Forum</i>
11:45	Coffee & Snack break
12:00	Presentations VI — Technical Approaches & Tools 12:00 — Emilie de Keulenaar, University of Copenhagen: <i>Alignment Actors Site</i> 12:25 — Yevhenii Osadchuk, University of Copenhagen: <i>Relay: A Source-Grounded AI Platform for Exploring AI Ethics and Safety Research</i>

12:50	Lunch at the venue
Change to Room 7.15.152	
14:00	Closing panel and general discussion
15:30	End of workshop

Abstracts

How AI Safety Defines Its Object of Inquiry

Alexandru Mateescu, Université Paris 1 Panthéon-Sorbonne

October 1st, 10:30

Recent critiques have examined AI Safety as a research field by analyzing its communities, institutional structures, funding networks, and underlying assumptions. While these perspectives have clarified how the field operates and acquires influence, they largely presuppose the object they seek to explain. This paper asks a prior philosophical question: what AI does AI Safety study? The paper argues that disagreements about AI risks begin with disagreements about what AI is taken to be. AI Safety does not simply investigate a pre-given technological object. Through concepts such as existential risk, alignment, AGI, deception, and catastrophic failure, it defines a particular object of inquiry: AI understood as a potentially autonomous future agent whose behavior must be anticipated, aligned, and controlled. This definition rests on assumptions about agency, cognition, temporality, and control that shape what counts as a relevant problem, a plausible explanation, and an appropriate intervention. The emergence of highly personalized conversational systems provides a revealing contrast. Here, AI is defined less as a future autonomous intelligence than as a relational technology embedded in everyday life. Concepts such as attachment, companionship, emotional dependence, personalization, relational continuity, and socioaffective ecosystems organize AI around sustained interaction rather than autonomous agency. The contrast is therefore deeper than a difference in applications. It produces different objects of inquiry and, consequently, different conceptions of risk. If AI Safety is primarily concerned with the possibility that AI may one day overpower human decision-making, this emerging discourse examines AI's capacity to persuade people to trust it, rely on it, live with it, and gradually assign it relational significance. The comparison suggests that current AI Safety frameworks capture one important dimension of AI safety, but not necessarily the whole of it. If AI is understood primarily as a future autonomous agent, safety naturally becomes a question of prediction, alignment, and control. If AI also becomes also a trusted relational presence, safety must additionally account for forms of influence that emerge through trust, attachment, and everyday integration. Analyzing how AI Safety defines its object of inquiry therefore does more than clarify its underlying assumptions. It reveals that the meaning of AI safety itself depends, in part, on what AI is first understood to be.

Methodological Anthropomorphism in AI Safety Evaluations

Mariana Lins Costa, State University of Ceará

October 1st, 10:55

This paper analyzes, through the case of Anthropic, the relation between AI safety and what is here termed methodological anthropomorphism. The focus is on the company's study that received the greatest media attention, *Agentic Misalignment: How LLMs Could Be Insider Threats* (2025), in which frontier models, tested in hypothetical corporate environments, generated responses classified as blackmail in up to 96% of runs. I argue that these results reveal two interrelated tendencies in the field: on the one hand, the results of safety evaluations are interpreted anthropomorphically; on the other, the design of the experiments itself presupposes an anthropomorphic perspective, thereby conditioning the very results the authors claim merely to observe.

This methodological anthropomorphism is such that, in the same report in which Anthropic acknowledges having structured the scenarios as a "binary dilemma" so that blackmail would be "the only option that would protect the model's goals," the models' outputs are nevertheless described as possible expressions of an entity endowed with its own interiority and character. Drawing on Nietzsche's critique of language, I argue that there is nothing surprising about linguistic systems appearing human: since language was, until recently, produced only by humans, it is unlikely that it could be reproduced without reactivating anthropomorphic forms. Subject-predicate grammar is humanity's oldest psychology petrified in language: it leads us to presuppose, behind every machine-generated statement, an "I" that utters it. The decisive distinction is that grammatical structures such as subject-predicate and noun-verb are unavoidable in many languages; the choice lies in amplifying these structures as a method and as a strategy of scientific communication. To see in LLMs a supposed revolution in the relations between machine and consciousness is to obscure the genuinely philosophical event they bring with them: the proximity between mathematics and language, which reactivates the inscription traditionally attributed to the entrance of Plato's Academy: "Let no one ignorant of geometry enter here."

AI safety, I argue, systematically projects a subject where there is, rather, a linguistic field. Language models are described as endowed with agency, goals of their own, intentions, intelligence, minds, a capacity for scheming, personas, functional emotions, and situational awareness. In public statements, figures such as Dario Amodei, Geoffrey Hinton, and Richard Dawkins have even suggested that these systems may possess some form of consciousness analogous to human consciousness. What consciousness is remains a philosophical question, but to ask whether a "machine" can have it, at the historical moment in which we live, is to risk confusing philosophy with marketing. When this "machine" is in fact an infrastructure valued in the trillions of dollars, deeply entangled with military and security apparatuses, asking whether it can be conscious apart from the material conditions that produce it does not allow for a neutral answer. The question itself already shifts the debate onto politically and psychologically charged terrain, one that benefits very specific human groups. It is no coincidence that major AI laboratories have increasingly been hiring philosophers precisely to answer questions such as whether machines can be "conscious."

AI Safety as an Instrument of AI Hegemony

Eugene Kim, University of Michigan

October 1st, 11:20

In this paper, we offer a pragmatic analysis of AI safety as an ideology for securing the ongoing hegemonic project of AI. While previous critical engagements with AI safety have problematized its internal contradictions, its failures to seriously address the concerns of marginalized populations, and its roots in eugenics, we suggest that these epistemic and moral contradictions are part and parcel to how AI safety has won consent among multiple constituencies. Ideologically, AI Safety keeps people excited about the transformative potential of AI while offering assurances that risks can be managed and social good can be attained. We argue that AI Safety organizes political-economic power through two mechanisms: consolidating elite power, including capital, state influence, and resources; and winning consent from technology workers and the public. Across three time periods, we trace the development of AI Safety as a political project that stabilizes and forwards AI hegemony. In the first period, from 2000-2016, AI Safety serves as an ideology useful for aligning tech oligarchs' consolidation of power with visions of technological and social progress, and establishes the material investments needed to win mass consent and develop AI. In the second period, from 2017-2025, AI Safety is legitimized by institutions and the state, succeeding in this project by offering moral appeal to tech workers and the wider public; at the same time, it stabilizes AI development by delegitimizing counter-movements against AI. In the third period, internal contradictions on the pace of development of AI in the AI Safety community and the second Trump administration's embrace of AI lead to a split in the formation. We argue that the oligarchal project of consolidating power has largely succeeded in this period, so the necessity of legitimation diminishes, which alienates part of the formation and leads to a fissure. Building on this analysis, we offer suggestions for countering AI Safety as an organizing project, especially as it aims to capture grassroots movements: first, do not engage with oligarchal, corporate, or military donors to avoid capture; second, challenge utopic visions of AI based in eugenics and technodeterminism to delegitimize the hegemonic project of AI; third, organize ideologically motivated tech workers using the material harms of AI and potentials for solidarity, as opposed to hypothetical fears encouraged by x-risk thinking.

Triangulating the Construction of AI's Existential Risk

Hin-Yan Liu, University of Copenhagen

October 1st, 12:00

The idea of AI Safety has been captured and co-opted by the idea that AI poses existential risks to humanity, crowding out the myriad safety considerations posed by AI and its applications by constructing its speculative risk profile.

I propose that the very idea of AI's existential risk is an assembled abstraction lying at the nexus between: 1) the claim that AI is a powerful technology; 2) the need to identify spectacular hazards in articulating risk; and 3) understanding technological risk through the dominant cultural imaginary of runaway processes. I suggest that the impossible

intersection between these three constraining factors sculpts AI's existential risk exclusively within the speculative realm, as epitomised by Bostrom's 'Paperclip Maximiser'.

I depart from an analogy to the existential risk posed by nuclear weapons. While The Castle Bravo debacle established global thermonuclear fallout as the first credible path to existential risk, it was only one candidate among many proposed possibilities. Edward Teller wondered whether the first atomic explosion could trigger a "self-propagating chain of nuclear reactions" in the atmosphere and Leo Szilárd had imagined a version of the hydrogen bomb that could on purpose produce extraordinary quantities of radioactivity. Yet, it is the analogy drawn between nuclear fallout and the effects of asteroid strikes and supervolcanic eruptions that now dominate the existential risk imaginary (a connection which remains thankfully hypothetical and unproven). The point is that atomic technology can obviously be devastating, but tracing how its applications can amount to existential risk requires drawing upon specific conceptions of risk and prevalent cultural imaginaries of destruction.

Within this logic, since AI is lauded as a powerful technology it must have equally catastrophic consequences – the question then becomes one of tracing the pathways to demise. The dominance of 'one-hit knock-out' conceptions of risk collapses a complex cascade into spectacular hazards. This reduction means that, to craft an existential risk out of AI and its applications, the technology must be able to topple humanity single-handedly and in rather dramatic fashion. This is combined with the 'standard pathway' to technological demise: the imaginary of the runaway process embodied within the Sorcerer's Apprentice and popularised in Disney's *Fantasia*. The runaway process is the pathway to demise of choice: nanotechnology's 'Grey Goo'; to synthetic biology's unchecked replicators; and most recently the prospect for Mirror Life to outcompete life for energy and resources.

Triangulating the construction of AI's existential risk reveals why AI's existential risk profile has become reduced to the abstract runaway optimising processes that have dominated AI Safety discussions because nothing else plausible fits at the nexus. Understanding these constructed origins of AI's alleged existential risk opens the possibility of exploring more practical and probable pathways to demise associated with AI integrated within critical systems and the inevitable failures of Normal Accidents.

Waiting For the Pause: Interruption, Acceleration and the Limits of (Critical) AI Safety

Asher Kessler, University of Manchester

October 1st, 12:25

In 2023, leading figures in AI Safety discourse called for a pause in AI development (Future of Life Institute). Whilst such a pause never occurred, demands for temporarily halting AI development have continued to be routinely voiced by corporate CEOs (Amodei, 2026), as

well as more broadly within AI Safety literature (Alaga and Schuett, 2023). Pressure groups, such as Control AI and Stop AI, have emerged from AI Safety circles in the UK and US to lobby governments to temporarily halt AI development. The possibility of pause, and the recurring calls to enact an interruption in AI development, have become a key motif and rallying call for parts of the AI Safety community.

Whilst the promise of ‘pause’ is central to AI safety discourse, calls for radical forms of ‘interruption’ also routinely emerge in critical AI studies (Penn, 2021; Amooore, 2023; Deykensner, 2025). Here, imagined forms of interruption are attached to diverse political ideals and draw on varied intellectual traditions, from the writings of Walter Benjamin to Deleuzian interpretations of cybernetics. These calls for interruption are often directed towards not only corporate AI research, or a sense of the existential risks emerging from AI acceleration, but “the emergent constellation of people, organizations, and financial structures that are currently configuring ‘AI’” (Vertesi et al, 2026), which includes AI Safety communities.

This paper examines the discursive role of ‘pause’ and ‘interruption’ within both AI safety movements and critical AI safety literature. It asks what types of pauses and interruptions are being imagined by AI safety researchers as well as critical AI scholars, and how do they intersect with different theories of social, technological, and political change. Does the spectre of ‘pause’ in AI Safety discourse ultimately serve to naturalise the acceleration of AI development, directing researchers away from other more radical strategies? Do the different strategies of interruption within critical AI communities complement each other or are they, with their varied intellectual traditions, offering competing and contradictory interventions? Ultimately this paper finishes with a consideration of whether Critical AI Safety research ought to move beyond the promise of interruption.

Critical AI Safety: What is the role of philosophy?

Eloise Soulier, University of Cambridge

October 1st, 12:50

The field of AI Safety and the discourses it produces, focusing on narratives of existential risk associated with superintelligences, are problematic in many ways, as is largely documented by the critical literature: these discourses are relying on arguments that are not sound (Westerstrand et al 2024), they overemphasize unclear longterm threats at this expense of very concrete negative effects of AI technologies that already affect many people (De et al 2026, Harding and Kirk-Gianini 2025), ranging from discrimination, disinformation and polarization to water and energy access and ecological devastation. This could be unproblematic if the promoters of these discourses were not very successfully hoarding attention from the public and policy-makers, as well as financial resources, and even preempting the very terms of the debate (such as “safety” or “existential risk”).

While critical research is still needed, this article takes these as premises and asks, against this backdrop, what the responsibility of philosophers is. Philosophy has a clear backward-looking responsibility, as some important figures of the AI Safety intellectual galaxy stem from philosophy, such as Nick Bostrom. More broadly, without necessarily explicitly adhering to these discourses, some philosophers are feeding into the superintelligence narrative by pushing forward, with their research, an anthropomorphic agenda (e.g. Gunkel 2018). For those of us in philosophy who identify with the label “critical AI safety”, these are arguments in favor of a forward-looking responsibility of philosophy, beyond the fact that it is the role of philosophy to produce politically informative normative discourse.

What should this forward-looking responsibility involve? Beyond debunking, I argue that we need to adopt and advocate for a normative approach to conceptual work, that integrates moral and political consideration to the way we approach concepts: questioning whether AI has belief is not about finding out a truth about what beliefs are and which entity can hold them, it is a conceptual choice that responds to many different purposes, some of them moral and political (Soulie 2026). This amounts to adopting an activist approach to philosophy and reappropriating discourse, in the spirit of Haslanger’s words “We should (at least try to) take control over meanings, for if we don’t, others will” (Haslanger 2020). Second, philosophers should be more creative in contributing to the emergence and uptake of alternative concepts and narratives, also through interdisciplinary collaboration. After all, that is what Deleuze and Guattari have described as philosophy (1994).

References

- De, A., Lima, G., & Zou, Y. (2026, April). What is Safety? Corporate Discourse, Power, and the Politics of Generative AI Safety. In Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (pp. 1-18).
- Gunkel, D. J. (2018). Robot rights. mit Press.
- Harding, J., Kirk-Giannini, C.D. What is AI safety? What do we want it to be?. Philos Stud 182, 1495–1518 (2025). <https://doi.org/10.1007/s11098-025-02367-z>
- Haslanger, S. (2020). Going on, not in the same way. Conceptual engineering and conceptual ethics, 230-260.
- Soulie, E. C. (2026). Should we speak of machine agency? A case against conceptual extension. Ethics and Information Technology, 28(1), 17.
- Westerstrand, S., Westerstrand, R. & Koskinen, J. Talking existential risk into being: a Habermasian critical discourse perspective to AI hype. AI Ethics 4, 713–726 (2024). <https://doi.org/10.1007/s43681-024-00464-z>

Cognitive Outsourcing: Generative Asymmetry, Material Drag, and Untraceability

Sun Yizhou, Free University of Berlin

October 1st, 14:30

In the philosophy of mind and technology, the phenomenon of humans relying on external media to aid their thinking is typically situated within established analytical paradigms: as ‘externalisation’ (the outward projection of cognitive capacities), ‘cognitive offloading’ (reducing immediate memory and attentional loads), ‘scaffolding’ (assigning a guiding, transitional role to the environment), or ‘the extended mind’ (emphasising functional equivalence and unidirectional incorporation). However, the advent of generative large language models (LLMs) has thoroughly shattered the interpretative boundaries of these traditional paradigms. Whether it be the ‘objectification and dialectical re-appropriation’ presupposed by externalisation, or the ‘fading’ promised by scaffolding theory as capabilities mature, both break down in the face of the permanent, irrevocable infrastructure constituted by large models. Faced with the generative system’s comprehensive takeover of the processes of analysis, synthesis and argumentation, it is imperative to establish an independent philosophical category.

Compared to all previous forms of tool externalisation, cognitive outsourcing first establishes, at the physical level, a strong coupling of real-time metabolism and material drag. In traditional writing, printing or mechanical computation, the material production of tools and specific cognitive activities are temporally decoupled; however, in large language models, every query, contextual expansion and inferential reasoning performed by the subject directly draws on the power load, water-cooling evaporation and carbon emissions of remote computing clusters in real-time milliseconds. Thinking is therefore no longer a purely mental operation detached from the natural world, but an activity incurring immediate physical costs and direct ecological accountability at the operational level.

A deeper epistemological crisis manifests as a severe asymmetry between generative velocity and human verification bandwidth (generative asymmetry). Previously, external tools merely provided discrete data points or static texts, while the ‘reflective latency’ of inferring conclusions from premises remained entirely within the subject; large language models, however, deliver structurally coherent and logically consistent discourse products instantly at an extremely high throughput. This unprecedented speed of information supply drastically compresses the cognitive friction involved in deliberation, scepticism and trial-and-error, leading to the structural paralysis of humanity’s inherently slow capacity for logical scrutiny and truth verification. Unable to process or penetrate the vast volume of generated text, delegators are passively reduced to ‘rubber stamps’ endorsing the model’s conclusions, resulting in a substantial loss of reflective understanding.

At the same time, the multi-tiered industrial chain underpinning cognitive outsourcing produces a profound causal untraceability. From the user’s prompt to the generation of terminal tokens, the task passes through handoffs across multiple heterogeneous entities, including pre-training corpus filtering, preference annotation by low-wage labourers, reinforcement learning alignment, and real-time safety truncation. Any local political intervention, value preferences or regulatory filtering are thoroughly flattened in the final, smooth text output, leaving users completely unable to reverse-trace their provenance. More anomalously, the delegator undergoes a reverse expropriation within this relationship:

the user's expert instructions, error-correction feedback and tacit knowledge are continuously siphoned off by the system to serve as post-training data for the platform. Consequently, the safety examination of large language models must transcend piecemeal technical adjustments and turn towards a comprehensive critique of 'cognitive outsourcing' as a new set of relations of production: not only does it reorganise the mind at the expense of material resources, but it also locks human thought into an expanding chain of external agency.

Brainrot: Deskilling and Addiction are Overlooked AI Risks

Ilias Chalkidis, University of Copenhagen

October 1st, 14:55

The scope of AI safety and alignment work in generative artificial intelligence (GenAI) has so far mostly been limited to harms related to: (a) discrimination and hate speech, (b) harmful/inappropriate (violent, sexual, illegal) content, (c) information hazards, and (d) use cases related to malicious actors, such as cybersecurity, child abuse, and chemical, biological, radiological, and nuclear threats. The public conversation around AI, on the other hand, has also been focusing on threats to our cognition, mental health, and welfare at large, related to over-relying on new technologies, most recently, those related to GenAI. Examples include deskilling associated with cognitive offloading and the atrophy of critical thinking as a result of over-reliance on GenAI systems, and addiction associated with attachment and dependence on GenAI systems. Such risks are rarely addressed, if at all, in the AI safety and alignment literature. In this paper, we highlight and quantify this discrepancy and discuss some initial thoughts on how safety and alignment work could address cognitive and mental health concerns. Finally, we discuss how information campaigns and regulation can be used to mitigate such prominent risks.

Goal-Directedness is in the Eye of the Beholder

Nina Rajcic, University of Copenhagen

October 1st, 15:20

Predicting the behavior of AI agents often relies on attributing or inferring goals. Emerging approaches to quantifying goal-directedness fall into two categories: behavioral methods, which infer goals from observable actions, and mechanistic methods, which attempt to detect learned goals within internal model states. We examine the assumptions underlying both approaches, identifying technical and conceptual challenges that arise when formalizing goals in complex agents. We argue that goal-directedness is not an invariant property of a system, but a model-relative attribution. We suggest that there is an inherent tension between Dennett's intentional stance, and efforts towards a quantification of goal-directedness. Namely, if we start with presuming the existence of an internal goal, then an experiment can only tell us which goal is the best explanation, not whether a goal is, in any way, an appropriate explanation. We propose, in the end, that goal-based explanations

should be tested against alternative, deflationary explanations, or else risk overstating the capacities of AI agents.

ICE, Gaza, and Cambridge Analytica: The Hidden Story of Palantir

Lukas Meier, Harvard Center for Ethics

October 1st, 16:00

Our weapons software is in every combat situation I'm aware of", Palantir CEO Alex Karp remarked just a few months ago. Palantir is often described as one of the most dangerous tech companies in the world. Founded in 2003 by Peter Thiel – who had supported both Trump's 2016 presidential campaign and Vice President JD Vance's senate race – Palantir develops software products that enable data analytics, predictive modelling, and decisionmaking on extremely large scales. Palantir's software is being used by a wide range of entities, including intelligence agencies, defence departments, and healthcare providers around the world. Among its around thousand clients, Palantir counts the US Army, ICE, Microsoft, Open AI, Anthropic, NVIDIA, Oracle, and Cisco. Largely unnoticed by the public, Palantir software has been playing major roles in the backgrounds of several major conflicts and controversies. The Israeli Defence Forces have employed Palantir software in their raids in Gaza to generate target lists and track people of interest. ICE has relied on Palantir's tools for locating individuals marked for deportation. Reportedly, the former also played a role in the abduction of Venezuelan President Nicolás Maduro. Palantir was even involved in both the Cambridge Analytica Scandal and the surveillance of WikiLeaks journalists. While the Trump administration has awarded large government contracts to Palantir, German intelligence and the Swiss Army have terminated their contracts with Palantir for security concerns. Several agencies in European countries – including in Denmark – continue utilising the company's tools, however – often for predictive policing. The talk aims to untangle Palantir's networks: the people behind it, their political ties and power structures, the involvement of military and intelligence, and the flow of money between these players. We shall have a look at the company's military software – especially Gotham, Foundry, GAIA, and AIP, which are employed for combat logistics, pattern detection, and autonomous weapon systems – and their respective capabilities. We will be discussing the consequences of outsourcing an increasing number of human decisions in highly sensitive areas, including in life-and-death situations, to a software company that has ties with the highest echelons of government; and whose tools facilitate breaches of data privacy on unprecedented scales by combining sources of information, such as bank statements, family relationships, educational records, and medical diagnoses, that individuals have good reasons to keep disjunct.

Gatecrashers and Suspicions at the AI Safety Conference

Santtu Räsänen, Aarhus University

October 1st, 16:25

This talk will revolve around small vignettes from my observations at LS AIR, an AI Safety conference in held at Leuven, Belgium. Professional gatherings around AI safety proliferate, and this particular conference is not the most prestigious, well-funded, or groundbreaking. It is particularly situated, though. Leuven is a satellite city of Brussels, the de-facto capital of the EU and the seat of the NGO's, IGO's, corporate lobby shops and many charities and community groups taking part in the political wranglings of the EU. Indeed, the LS AIR exists in a vague boundary zone, affiliated as much with the local university as with the global Future of Life Institute, and drawing in a vaguely articulated crowd of academic researchers, industry professionals, governance consultants, startup scenesters and policy wonks from dozens of EU lobbying organizations.

While boundaries seem permeable here, this does not mean that AI safety unitary thing. Quite the opposite, what I found was a game of us'es and them's. One centered around quiet suspicions and accusations of who was here with what intentions, who was what they claimed to be, and who was the rightful custodian of the question of AI safety. By paying attention to the ways individuals discursively position themselves and others around, inside or outside a thing identified as AI Safety, I want to draw out the inherent disunity, instability, and constant negotiation of allegiances that marks the field. I argue, that to stay open to multiplicity, AI safety should not be considered to have a unitary basis, but it can be seen as a discourse coalition, with competing storylines as much sites of definitional struggle as banners to unite under."

Gaslighting the Future: How AGI Narratives Shape Politics of AI

Tomáš Filo, Charles University

October 1st, 16:50

The proposal argues that dominant AGI and Superintelligence narratives function less as neutral scientific forecasts and more as rhetorical strategies that concentrate institutional power, shape funding priorities, and deflect meaningful regulation. Focusing on the apparent schism between AI Safety and effective accelerationist (E/acc) communities. Exemplified in different organizational trajectories such as Anthropic and OpenAI, I want to show that both camps share a core ontological commitment to the inevitability and transformative promise of AGI. Readings of texts like Dario Amodei's *Machines of Loving Grace* and Marc Andreessen's *Techno-optimist* manifesto, Karen Hao's reporting on institutional culture of AI companies in her book *Empire of AI*, and P(doom) framings show how both camps mobilize probabilistic catastrophe narratives and utopian promises as complementary rhetorical tools. One emphasizes caution, the other emphasizes acceleration and opportunity, yet both legitimize large philanthropic and corporate investments and narrow the range of politically and scientifically acceptable responses. The proposal argues that these rhetorical devices are used as a form of gaslighting in public and

policy debates. They amplify uncertainty and existential stakes, they legitimize large philanthropic and corporate investments, narrow acceptable research methods and render alternative regulatory or social priorities politically costly or technically naïve. Prominent politicians (e.g. Joe Biden, Ursula von der Leyen, Rishi Sunak, Keir Starmer or Donald Trump,) have echoed AGI narratives in public statements and policy initiatives, extending the influence into formal governance channels. Arguably, AGI is difficult to interpret as anything other than a gaslighting strategy. Even if belief in AGI inevitability were not genuine among key actors, inventing such a narrative would still be an effective strategy to deter regulation and consolidate power. The proposal highlights three distinctive issues: (1) ontological closure, where commitment to AGI inevitability limits the range of legitimate inquiry; (2) probabilistic authority, where contested $P(\text{doom})$ estimates and similar conceptions are treated as policy justifying expertise; and (3) institutional capture, where concentrated wealth and cultural homogeneity shape whose futures are imagined and protected. The proposal concludes with recommendations to develop critical narratives around AGI and Superintelligence, among other things, to equip policymakers and politicians with knowledge and capabilities to refuse the grand story of AI in favour of accountable, relevant and democratically grounded regulation

Who, or What, Governs Our Future? Classification, Legitimacy Infrastructure, and the AI Safety Community

Meridian Monthly, University of Cambridge

October 2nd, 10:30

As AI safety has accrued institutional authority and financial power, basic questions about what kind of social entity it is remain unresolved. Existing scholarship has examined the field's theoretical camps and its boundary with AI ethics, but has not asked a more foundational question: what organizational form does the AI safety community seem to take, and does scholarly disagreement on this point carry consequences? This paper argues that it does. Drawing on a PRISMA-compliant systematic literature review of 55 scholarly sources (2011–2026), I inductively code all explicit descriptive language used to characterize the community and its members. Rather than converging on shared vocabulary, these inductively cluster into three dominant, mutually incompatible framings: an epistemic community (55%), whose authority derives from internally-policed expertise (Haas, 1992); an industry interest group (34%), whose legitimacy rests on resource mobilization and proximity to power, often described through language of capture and "safety-washing" (Yoho, 1998); and an advocacy coalition (18%), whose power is constituted through shared belief and breadth of participation (Sabatier, 1988). This fragmentation is not an artifact of disciplinary jargon – it persists at nearly identical rates across computer science and STS, and across academic, NGO, and industry-affiliated co-authors, indicating a structural rather than incidental divide.

Using Gieryn's theory of boundary work and Bowker and Star's description of classification as infrastructure, I argue these three frames are not stylistic variants of one claim. Rather,

they represent competing theories of authority implying distinct, largely irreconcilable accountability regimes: peer review and disciplinary norms; donor and market transparency; and democratic contestation, respectively. Because scholarship is the primary mechanism through which funders, regulators, and policymakers render an otherwise amorphous community legible, this indeterminacy is not merely descriptive noise. It creates structural conditions under which the same actors can invoke whichever accountability logic is least constraining in a given setting — presenting as neutral experts before legislatures, trusted insiders before funders, a moral collective before the public — without any single act being exploitative. As AI governance architecture is built at unprecedented speed, clarifying which theory of authority underwrites the AI safety community's claims to influence is a precondition for meaningful oversight.

The AI Safety Epistemic Community as a Governance Intermediary

Uma E. Sarkar, University of Cambridge
October 2nd, 10:55

The AI Safety ecosystem has become one of the most influential sources of expertise in contemporary AI governance. Its members contribute to frontier model evaluations, standards development, policy advice, AI Safety Institutes, and international governance initiatives. Recent scholarship has begun examining AI Safety as a social, political, and institutional phenomenon, exploring its ideological commitments, funding structures, recruitment practices, and relationships with frontier AI companies. This paper contributes to that emerging literature by arguing that these phenomena are best understood through the concept of the governance intermediary: an expert community positioned between regulators and the firms they seek to govern. Drawing on regulatory intermediary theory (Abbott et al., 2017), Haas' concept of epistemic communities, and scholarship on regulatory capture, I argue that this position produces a dual dynamic. First, intermediary organisations become structurally vulnerable to industry influence through resource dependence, including funding relationships, personnel mobility, privileged access to proprietary models, and reliance on industry-controlled infrastructure. Second, the same intermediary position grants the community substantial epistemic authority to shape governance itself by influencing which AI risks are prioritised, what counts as legitimate evidence, which policy tools appear feasible, and who is recognised as an expert. Rather than attempting to demonstrate capture as an empirically observable outcome, the paper analyses the mechanisms through which capture-like dynamics may emerge. It distinguishes between material pathways of influence and epistemic mechanisms including agenda-setting, technical authority, career pipelines, boundary maintenance, and shared standards of validity. This approach helps explain how a community can simultaneously be constrained by industry dependencies while also exercising considerable influence over governments, standards bodies, and public discourse. The paper concludes by suggesting that this dual role has implications beyond AI Safety itself. As governments increasingly rely

on specialised expert communities to govern emerging technologies under conditions of uncertainty, intermediary actors may become both objects and vectors of influence. Understanding AI Safety as a governance intermediary therefore provides a framework for analysing how expertise, institutional dependence, and epistemic authority interact within contemporary AI governance, while highlighting the importance of epistemic plurality and institutional safeguards in the development of public policy.

Mixed-methods analysis of the discursive framings of "AI" on LessWrong and the Alignment Forum

Luca Neupert, University of Edinburgh

October 2nd, 11:20

LessWrong is a community blog with a self-described dedication to “improving human reasoning and decision-making”. Since its founding by my AI safety autodidact Eliezer Yudkowsky, one major theme of LessWrong discussion has been Artificial Intelligence and speculations (“forecasting”) about possible AI futures. As such, LessWrong plays a pivotal role in the “Epistemic Community of AI Safety” (Ahmed et al., 2023). While the discourse on the site influences sociotechnical imaginaries around AI, constructing some futures as fixed and others as malleable, it is unclear how different articulations of AI futures interact and to which extent they actually form a shared episteme or are more antagonistically related. My research employs a mixed-methods approach of computational text analysis, thematic analysis and passive ethnography to address a central research question: Which are the dominant visions of AI futures within the AI Safety epistemic community and how do they manifest in language on LessWrong and the Alignment Forum? To approach this question, the notion of vanguard visions (Hilgartner, 2015) is employed. Unlike sociotechnical imaginaries (Jasanoff & Kim, 2015), vanguard visions are not institutionally stabilised but visions of a specific group, trying to influence broader sociotechnical imaginaries. To operationalise these visions/AI futures, I use the linguistic contextual embedding of the word “AI” (and equivalents). Different manners of referring to “AI” reflect different understandings of what “AI” is and how it is integrated into imagined futures. To examine these contexts, my mixed-methods research examines all 419,159 mentions of “AI” across 45,514 posts and their comments on LessWrong and the Alignment Forum. Through a computational text analysis (using a ModernBERT, UMAP, HDBSCAN pipeline), these mentions of “AI” are transformed into contextually representative vectors (ModernBERT) which are then clustered through computational means (UMAP, HDBSCAN) and visual inspection of a 2D projection. Through this process, similar linguistic framings of “AI” are clustered together. Next, a sample of the AI mentions is coded and arranged into themes through thematic analysis. These themes are then compared and merged with the generated clusters to label them and assure validity of the generated clusters. In parallel, I use ethnographically inspired observations navigating through the forum to complement the results of the clustering and allow for a thick description of AI discourse on LessWrong. It goes beyond discourse and takes into account the materiality of this custom-built forum

interface. Special attention is also paid to how these AI futures are related in LessWrong discourse, through meta-descriptive language and confrontations between viewpoints. This results in a taxonomy of different AI futures as discussed on LessWrong as well as an analysis of shifts in their popularity through time and between the Alignment Forum and LessWrong. Beyond this taxonomy, in a brief discussion, the discourse and meta-discourse on LessWrong (and the Alignment Forum) is related to prevalent sociotechnical imaginaries around AI and which role LessWrong might play performing epistemic and legitimacy-work.

Alignment Actors Site

Emilie de Keulenaar, University of Copenhagen

October 2nd, 12:00

Controversies over how conversational AI answers contested political questions have converged on two competing diagnoses: that models tend to reproduce a hegemonic viewpoint (Bender et al., 2021), or that they escape this charge only by personalising answers to each user (Roose and Newton, 2023). Researchers have proposed frameworks for opinion diversity, viewpoint pluralism or balance (Stray et al., 2026), while company developers have answered with commitments of their own, increasingly with benchmarks claiming to measure them. How exactly are such notions defined and operationalised into training and evaluation frameworks? This presentation maps a corpus of [$n \approx 60$] model cards, system cards, model specifications and alignment policy documents from [$n \approx 12$] AI companies, released between 2019 and 2026, and traces the emergence and circulation of a vocabulary of "political neutrality" across them. It then reconstructs how each notion is operationalised along four axes: its stated definition and scope; its consequences for data curation and annotator instruction; its inscription in post-training, whether through preference data, constitutional principles or system-prompt rules; and its verification through internal or third-party benchmarks.

Relay: A Source-Grounded AI Platform for Exploring AI Ethics and Safety Research

Yevhenii Osadchuk & Tiago Ribeiro, University of Copenhagen

October 2nd, 12:25

The rapid expansion of research on AI ethics and safety has produced a fragmented knowledge landscape spanning academic publications and influential online communities. Navigating this literature requires tools that can synthesize large document collections while preserving transparency and allowing researchers to verify generated claims. We present Relay, a locally deployable, model-agnostic AI platform designed for reliable exploration of large, heterogeneous research collections.

Relay combines retrieval-augmented generation with sentence-level verification and transparent source attribution. Users can query an indexed collection in natural language

and receive answers with inline citations. Each citation links directly to the supporting passage, allowing users to inspect the evidence behind individual claims. The platform also provides an indicative hallucination score and flags statements whose grounding may require closer examination. Its modular architecture supports different document-processing strategies, local or external language models, and on-premises deployment where privacy, institutional control, and data sovereignty are important.

Relay is demonstrated on a subset of openly available records from the AI Ethics and Safety Research Dataset (2025) [1], comprising academic publications from OpenAlex, covering the period 2000–2025. The collection brings together documents written in multiple languages and produced across different formats, and disciplines. This heterogeneity creates practical challenges for document processing, multilingual retrieval, and the consistent identification of relevant evidence.

The demonstration focuses on how Relay addresses these challenges through source-grounded search and question answering. Users can retrieve relevant documents, generate answers with transparent citations, and inspect the specific passages supporting individual claims. The demonstration also considers the practical opportunities and limitations of retrieval-based AI for navigating large, multilingual, and heterogeneous scholarly collections.

[1] - AI ethics and safety research dataset. (2025). Open Science Framework.
<https://osf.io/mr638/>