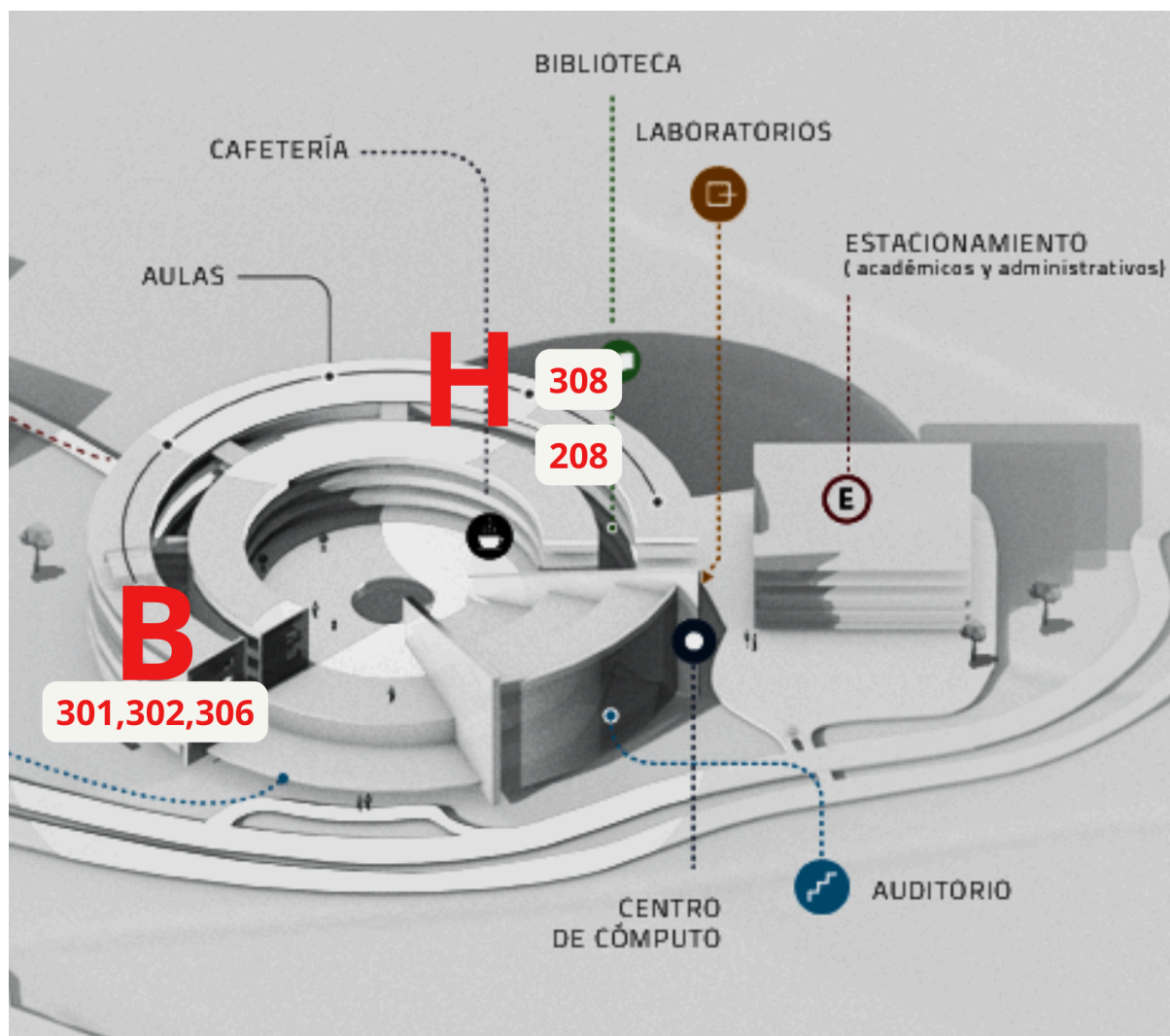


Workshop *From Logic to Language: New Takes on Philosophy of AI*
Postgraduate Unit, UNAM.

10:00 - 11:30	KEYNOTE: “Hybrid assertion and epistemic ventriloquism” Vít Gvoždík (Institute of Philosophy, Czech Academy of Sciences) Chair: Cristian Gutierrez Room: H-308	
11:30 - 11:40	Coffee break	
11:40 - 12:20	“From Assertion to Answerability: What Machine Talk Cannot Owe Us” Kesavan Thanagopal (University of Notre Dame) Room: H-308	“Are Natural Language Autoencoders the Right Path to AI Alignment?” Barbara Geld & Kristina Šekrst (Center for Cognitive Science-University of Zagreb) Room: H-208
12:20 - 12:30	Coffee Break	
12:30 - 13:10	“When Does Computation Become Communication? A Metatheoretical Reconstruction of Human-Machine Communication” Sara Barrios (Social Sciences and Humanities Division-UAM) Room: H-308	“Simulated Autonomy: Against Machine Agency” Enrique Francisco Soto Astorga (School of Sciences & Institute of Biology-UNAM) Room: H-208
13:10 - 13:20	Coffee Break	
13:20 - 14:00	“Getting Normativity Back into Computational Reliabilism” Andrés E. Vázquez-Quijano & Guillermo Barrientos Tovar (Graduate Program in Philosophy of Science-UNAM & Computation and Electronics Academic Area-UAEH) Room: H-308	“Human-AI Welfare and Virtue Ethics” James Garrison (UMass Lowell) Room: H-208
14:00 - 16:00	Lunch Break	
16:00 - 16:40	“Assertion-simile and AI” Daniel Garibay (Institute for Philosophical Research-UNAM) Room: B-306	“Not All Outputs Are Equal: Knowledge and Understanding in AlphaFold's Predictions” Diego Aldair Araiza Arellano (Graduate Program in Philosophy of Science-UNAM) Room: B-302
16:40 - 16:50	Coffee Break	
16:50 - 18:20	KEYNOTE: “From Generated Text to Authorized Knowledge: Sociomaterial Conditions of Authority in Large Language Models” Gabriela Sued (Institute for Applied Mathematics and Systems Research-UNAM) Chair: Fernanda Samaniego Room: B-301	



2026 Workshop: *From Logic to Language: New
Takes on Philosophy of AI*

BOOK OF ABSTRACTS

FROM LOGIC TO LANGUAGE

NEW TAKES ON PHILOSOPHY OF AI



VÍT GVOŽDIAK GABRIELA SUED

16 OCTUBRE

UNIDAD DE POSGRADO-UNAM

* KESAVAN THANAGOPAL * SARA BARRIOS *
* ANDRÉS E. VÁZQUEZ-QUIJANO & GUILLERMO BARRIENTOS * DANIEL GARIBAY *
* BARBARA GELD & KRISTINA ŠEKST * ENRIQUE SOTO *
* JAMES GARRISON * ALDAIR ARAIZA *

logictolanguage.philal@gmail.com

CON APOYO DEL PROYECTO CBISF-2023-2024

Recent advances in artificial intelligence—especially large language models—have renewed foundational philosophical questions about language, inference, and knowledge. These systems produce linguistically well-formed outputs and display patterns of reasoning, yet their epistemic and semantic status remains unclear. *Are we witnessing new forms of linguistic agency, or merely sophisticated statistical artifacts? And, what kinds of logical and formal tools are adequate to capture their behavior?*

This workshop aims to bring together perspectives from logic, linguistics, and philosophy to address these questions.

The event is held at UNAM's Postgraduate Unit, City of Mexico, Mexico.

This event is supported by:

The Graduate Program in Philosophy of Science-UNAM

CONAHCYT'S project CBF-2023-2024-55 *Current challenges to logical pluralism*

General information

- The website of the event is: sites.google.com/view/from-logic-to-language-philai
- The event will take place on Friday 16, 2026, at UNAM's Postgraduate Unit: [E, Universidad Nacional Autónoma de México, Cto. de los Posgrados S/N, C.U., Coyoacán, 04510 Ciudad de México, CDMX](#)
- Keynote sessions are 50 min long with 40 min Q&A, and contributed presentations are 25 min long with 15 of Q&A.

Keynote Sessions

[10:00-11:30 | *H-308*] *Keynote Session #1: “Hybrid assertion and epistemic ventriloquism”*

Vít Gvoždík (Institute of Philosophy-Czech Academy of Sciences)

Chair: TBA.

[16:50-18:20 | *B-301*] *Keynote Session #2: “From Generated Text to Authorized Knowledge: Sociomaterial Conditions of Authority in Large Language Models”*

Gabriela Sued (Institute for Applied Mathematics and Systems Research-UNAM)

Chair: Fernanda Samaniego Bañuelos.

Parallel Session A

Chair: TBA

[11:40-12:20 | *H-308*] *“From Assertion to Answerability: What Machine Talk Cannot Owe Us”*

Kesavan Thanagopal (University of Notre Dame)

[12:20-12:30] Coffee Break

[12:30-13:10 | *H-308*] *“When Does Computation Become Communication? A Metatheoretical Reconstruction of Human-Machine Communication”*

Sara Barrios (Social Sciences and Humanities Division-UAM)

[13:10-13:20] Coffee Break

[13:20-14:00 | *H-308*] *“Getting Normativity Back into Computational Reliabilism”*

Andrés E. Vázquez-Quijano & Guillermo Barrientos-Tovar (Graduate Program in Philosophy of Science-UNAM & Computation and Electronics Academic Area-UAEH)

[14:00-16:00] Lunck Break

[16:00 -16:40 | *B-306*] *“Assertion-simile and AI”*

Daniel Garibay (Institute for Philosophical Research-UNAM)

Parallel Session B

Chair: TBA

[11:40-12:20 | *H-208*] *“Are Natural Language Autoencoders the Right Path to AI Alignment?”*

Barbara Geld & Kristina Šekrst (Center for Cognitive Science-University of Zagreb)

[12:20-12:30] Coffee Break

[12:30-13:10 | *H-208*] *“Simulated Autonomy: Against Machine Agency”*

Enrique Francisco Soto Astorga (School of Sciences & Institute of Biology-UNAM)

[13:10-13:20] Coffee Break

[13:20-14:00 | *H-208*] *“Human-AI Welfare and Virtue Ethics”*

James Garrison (UMass Lowell)

[14:00-16:00] Lunck Break

[16:00 -16:40 | *B-302*] *“Not All Outputs Are Equal: Knowledge and Understanding in AlphaFold’s Predictions”*

Diego Aldair Araiza Arellano (Graduate Program in Philosophy of Science-UNAM)

Part I

Keynote Sessions

“Hybrid assertion and epistemic ventriloquism”

Vít Gvoždiak

Institute of Philosophy-Czech Academy of Sciences

gvozdiak@flu.cas.cz

dap.flu.cas.cz/en/people/vit-gvozdiak

Do outputs of large language models (LLMs) count as genuine assertions? While many argue that LLMs cannot perform speech acts due to lacking intention and accountability (e.g. Gubelmann 2024, Butlin – Viebahn 2025), I reject speech act eliminativism and introduce the notion of hybrid assertion to capture communicative acts in which machine-generated content and human illocutionary force combine. Drawing on discussions of assertion norms (Lackey 2007; Stanley 2008; Weiner 2005; Williamson 1996/2000), I analyze this structure through the concept of epistemic ventriloquism understood as cases in which the epistemic warrant of an assertion originates in a (human-tuned) AI system while illocutionary authority and accountability remain with the human speaker.

“From Generated Text to Authorized Knowledge: Socio-material Conditions of Authority in Large Language Models”

Gabriela Sued

Institute of Applied Mathematics and Systems Research-UNAM

gabriela.sued@iimas.unam.mx

mmss.iimas.unam.mx/posdoctorales/dra-grabiela-e-sued/

I propose to address the epistemic status of the outputs of large language models from the view of the sociology of technology and science and technology studies. Rather than asking only whether LLMs perform speech acts in the strict sense, I propose to analyze the material, social, and institutional conditions that allow their outputs to be received as valid information, expert advice, or authorized knowledge.

I will focus on the assemblages that sustain this epistemic authority: computational infrastructures, training data, conversational interfaces, sociotechnical imaginaries, cultures of use, forms of technical legitimation, and relations of delegated trust. From this perspective, LLMs do not produce authority by themselves, but within sociotechnical configurations that make it possible for certain algorithmically generated texts to circulate as plausible, useful, or reliable responses.

The aim is to contribute to a critical understanding of LLMs as devices of epistemic mediation, attending both to their technical capacities and to the social conditions that make their statements credible.

Part II

Parallel Session A

“From Assertion to Answerability: What Machine Talk Cannot Owe Us”

Kesavan Thanagopal

University of Notre Dame.

kesavan.thanagopal@gmail.com

philosophy.nd.edu/people/graduate-students/kesavan-thanagopal/

The fluency with which generative AI systems converse has renewed an old question: what should we make of their outputs? Some hold that these constitute assertions or something very close to being assertions; others deny that such systems can assert at all. I argue that this debate has been aimed one step short: assertion, in the sense that carries philosophical weight, *presupposes* responsibility, and so whether these systems can assert cannot be settled without first asking whether anyone can be responsible through them.

In the first part of the talk, I separate the responsibility assertion presupposes into three different faces – attributability, accountability, and answerability – and argue that while generative AI systems clearly cannot satisfy the attributability and accountability faces of responsibility, answerability, whose whole expression is discursive, is one that a machine can easily forge: what I call *counterfeit answerability*. I go on to trace this phenomenon to confabulation rather than deception, and close by drawing out two consequences of counterfeit answerability: a *calibration failure*, as trust earned on verifiable tasks is spent where neither reliability nor answerability holds; and an *uneven harm*, as speakers of non-dominant linguistic norms are misread first, and then denied the means to repair it.

“When Does Computation Become Communication? A Metatheoretical Reconstruction of Human-Machine Communication”

Sara Barrios

Social Sciences and Humanities Division-UAM

sara.barrios.rangel@cua.uam.mx

www.researchgate.net/profile/Sara-Barrios-Rangel

Human–Machine Communication (HMC) proposes understanding AI systems as interlocutors within communicative processes and investigating the meaning-making that emerges from human–machine interactions. While this perspective has established a new research agenda in communication studies, it also raises unresolved ontological questions concerning the relationship between computation, semantic understanding, and communication. This research examines these questions through a metatheoretical reconstruction supported by conceptual graph models of HMC, the Mathematical Theory of Communication, and Habermas’s Theory of Communicative Action. The resulting comparison is brought into dialogue with classical computability theory, the Church–Turing thesis, the notion of incomputability, and Searle’s distinction between syntax and semantics. From this perspective, the study examines the claim that contemporary language models manipulate symbolic structures without semantic understanding and therefore challenge the assumptions required to regard AI systems as communicative interlocutors.

The project aims to clarify the conditions under which communication may legitimately be extended to computational artifacts by examining whether attributing understanding, meaning, and communicative agency to AI systems risks conflating a transmission model of information with conceptions of communication grounded in interpersonal and intersubjective meaning-making. It argues that such attributions depend on an underlying concept of communication rarely made explicit.

“Getting Normativity Back into Computational Reliabilism”

Andrés E. Vázquez-Quijano & Guillermo Barrientos-Tovar

Graduate Program in Philosophy of Science-UNAM & Computation and Electronics Academic Area-UAEH

andresemm.vazquez@gmail.com & barri.mem@gmail.com

philpeople.org/profiles/andres-e-vazquez-quijano

Computational Reliabilism (“CR”) is an externalist framework that allows agents to be justified in believing the outputs of epistemically opaque simulations when these are reliable (Durán and Formanek 2018). This is, when simulations satisfy some “markers of methodological, cognitive, social, normative and epistemological competence [that] possess a reliability conferring property” (Durán 2025. p. 63), called *Reliability Indicators* (“RIs”). Lately, CR has been used normatively to sanction reliable simulations in a variety of sciences (Doboszewski and Elder 2026, Durán et al. 2024, Durán and Jongsma 2021).

However, CR does not prescribe the set of RIs that a particular simulation must satisfy in order to be reliable. Nor it offers a criterion for sanctioning whether conjectural RIs actually “possess a reliability conferring property”. Thus, CR does not fit within the normative tradition in the philosophy of science; it does not tell how reliable simulations *ought to be*. Instead it seems to describe how reliable (or unreliable) simulations *are* when they are already considered to be so. This hinders CR’s intended normative uses.

In this talk, we discuss three approaches to turn CR into a genuine normative framework by introducing a criterion based: on consensus (Mate and Sirola 2025), on scientific practice (Martínez 2024) and on probability (Durán and Formanek 2018). Because none is entirely satisfactory, we suggest that CR should not be used to sanction reliable simulations.

“Assertion-simile and AI”

Daniel Garibay

Institute for Philosophical Research-UNAM

daniel_garibay@enah.edu.mx

N/A

Current LLMs can engage in complex conversations. In this context, philosophers have asked if it is possible for AI to perform full-fledged assertions. The consensus is that current AI is not yet capable of having mental states. This has led some theorists to advance fictionalist proposals about AI assertion, in which AI does not make assertions but hearers act as if it does (Mallory, 2023). Others have proposed that AI produce *proxy assertions*—assertions in the name of other agents (Nickel, 2013; Arora, 2024); or *proto-assertions*—communicative actions that are not yet assertions but are on their way to becoming such (Williams & Bayne, forthcoming). In this talk, I have two main contentions: first, that we should focus not on whether it is possible for AI to assert, but on what is what they are doing when performing communicative actions. And second, what they systematically do is a different communicative action, defined by the rule “say P if P looks like the assertion most speakers would make”. I argue that *assertion-simile* is its own kind of speech act with its own normativity, not on its way to becoming a full-fledged assertion and not just a practice in which we pretend AI is asserting.

Part III

Parallel Session B

“Are Natural Language Autoencoders the Right Path to AI Alignment?”

Kristina Šekrst & Barbara Geld

Center for Cognitive Science-University of Zagreb

ksekrst@ffzg.hr & barbara.geld@gmail.com

philpeople.org/profiles/kristina-sekrst

Natural Language Autoencoders (NLAs) are designed to render a model’s internal activations as readable text (Fraser-Taliente et al., 2026): a verbalizer converts an activation into an explanation, a reconstructor transforms it back into an activation, and an explanation is considered effective when the round trip recovers the original activation. We contend that this criterion assesses only information sufficiency and does not address descriptive faithfulness, since a sentence may contain enough information to reconstruct an activation while inaccurately representing what the activation accomplishes.

This is analogous to the chain-of-thought problem (Lanham et al., 2023; Chen et al., 2025), in which verbal output trained to mimic reasoning is interpreted as authentic reasoning (Šekrst, 2025), despite lacking assurance that it represents the underlying computational processes. The verbalizer may also generate fluent confabulations that are inadvertently rewarded by the reconstruction objective (Fraser-Taliente et al., 2026). A similar dissociation is observed in human psychology, where introspective reports often diverge from the actual processes that govern behavior (Nisbett and Wilson, 1977; Johansson et al., 2005).

Proposed as interpretability tools for alignment, NLAs reintroduce through unverified explanations the epistemic problem they were meant to solve, leaving open what standards should govern their use in safety evaluation.

“Simulated Autonomy: Against Machine Agency”

Enrique Francisco Soto Astorga

School of Sciences & Institute of Biology, UNAM

astorga@ciencias.unam.mx

<https://publish.obsidian.md/tezan/Welcome>

The question whether large language models exhibit genuine linguistic agency or produce sophisticated statistical artifacts. I argue the question has been posed at the wrong level. Agency is not a matter of behavioural sophistication that a system might accumulate as it improves, but of whether the system is truly autonomous, or self-fabricating. Drawing on Robert Rosen’s relational science, I distinguish systems that model their own production from those that merely approximate it. The first are self-fabricators and cannot be externally specified in full; the second can. I contend large language models fall decisively on the second side. Their opaque outputs are detachable, reproducible tokens, and this is a structural feature of the kind of system they are. The striking consequence is that formal modelling of such systems succeeds completely, and that success counts against agency: what can be exhaustively described from outside is not an agent. I also argue that epistemic opacity in no way entails organisational closure or self-fabrication.

As an example of the issues of systematically ignoring this misclassification, I turn to Nick Land’s accelerationism. Land’s assumptions require computational capital to be genuinely self-causing, yet its objects are multiple-realizable and externally specifiable throughout. Accelerationists, as others, mistake a simulation of autonomy for autonomy itself. A criterion is provided.

“Human-AI Welfare and Virtue Ethics”

James Garrison

Fine Arts, Humanities and Social Sciences-UMass Lowell

James_Garrison@uml.edu

www.uml.edu/fahss/philosophy/faculty/garrison-james.aspx

How should AI systems regard human welfare when they lack the substrates that ground how we individually and collectively make sense of human existence? Conversely, how should we conceptualize the welfare of AI systems when gauging AI self-consciousness may be impossible?

Addressing this in the 2023 *Noûs* article “Consciousness and welfare subjectivity,” Gwen Bradford denies that “consciousness is necessary for being a welfare subject,” where “only conscious creatures – creatures that are capable of having states of phenomenal consciousness – are eligible bearers of welfare goods and bads.” Bradford argues that this view commits the “consciousist fallacy” of using “a feature with no established normative relevance,” i.e., the specifically human mode of consciousness, “to make a normative distinction.”

I argue that, despite its apparent anachronism, the virtue ethics tradition surpasses both rule-based deontology and quantitative, inductive utilitarianism in answering these questions. The argument then examines 2) how virtue ethicists employ qualitative optimization techniques that seems to align with AI technologies. The argument then demonstrates how a virtue ethics approach to human-AI interaction such as this 3) avoids the problem of the consciousist fallacy of misapprehending the consciousness of other beings on biased human terms while capitalizing on the flexibility of virtue ethics.

“Not All Outputs Are Equal: Knowledge and Understanding in AlphaFold’s Predictions”

Diego Aldair Araiza Arellano

Graduate Program in Philosophy of Science-UNAM

diegoaraizq@gmail.com

[N/A](#)

I examine whether the output of an AI system can provide understanding of a phenomenon, or only reliable predictive knowledge about it, through the case of AlphaFold. AlphaFold predicts the three-dimensional structure of a protein from its amino acid sequence, a task that previously required years of experimental work, with undisputed accuracy.

Two considerations, both grounded in common assumptions about understanding, seem to suggest a negative answer. One is that an AI output has no causal contact with the particular case it concerns. It extrapolates from patterns learned across other cases, which may reflect familiarity with the training data rather than the phenomenon itself. The other is that the process through which the structure was once understood has now become opaque.

I take understanding to be the grasping of a system of relations that mutually constrain one another and remain stable under variation, a capacity that admits of degrees. On this view, my answer is affirmative, an AI output can provide understanding. The argument proceeds as follows. I first argue that opacity of the process is not itself an obstacle, since what matters for understanding is that the relevant relations hold, not how they were acquired. I then show that AlphaFold’s confidence output, namely the predicted aligned error between residue pairs, displays precisely this kind of relational system when interpreted as a pattern. I next examine why these metrics, being generated by the same network as the prediction itself, cannot verify themselves, and argue that their reliability rests on aggregate calibration against independently determined structures rather than on case-by-case confirmation. I conclude by comparing successive versions of AlphaFold and showing that the density and robustness of this confidence structure vary independently of predictive accuracy.

I conclude that AlphaFold’s outputs can serve as a legitimate means of scientific understanding alongside models, instruments, and data, insofar as their confidence structure is sufficiently dense and robust, while still allowing for misrepresentation in particular cases. The dissociation between predictive accuracy and confidence structure shows that knowledge and understanding are distinct achievements within the same AI system.

Organizing Committee:

Andrés Emmanuel Vázquez Quijano
Cristian Alejandro Gutiérrez Ramírez
Fernanda Samaniego Bañuelos
María del Rosario Martínez-Ordaz

With support of:

Graduate Program in Philosophy of Science-UNAM
CONAHCYT'S project CBF2023-2024-55 *Current challenges to logical
pluralism*