IVADO

IVADO Symposium

COMPARATIVE PERSPECTIVES ON AGENCY: PHILOSOPHY AND ARTIFICIAL INTELLIGENCE

May 22, 2026

CANADA FIRST RESEARCH EXCELLENCE FUND | APOGÉE CANADA FONDS D'EXCELLENCE EN RECHERCHE

Québec

IVADO Symposium

COMPARATIVE PERSPECTIVES ON AGENCY: PHILOSOPHY AND ARTIFICIAL INTELLIGENCE

Conference Report

Prepared by Christophe Facal

June 2026

An event organized by the IVADO Cluster on Ethics, Equity, Diversity and Inclusion (EDI), and Indigenous Engagement.

The widespread deployment of artificial intelligence (AI) systems is transforming the way we describe machine action. Machines are increasingly characterized as "agents": multimodal agents, autonomous agents, conversational agents, decision-making agents, and so on.

Yet the concept of agency is plural and context-dependent, and a genuine tension appears to exist between the computational use of terms such as *agent* or *agentic* and their philosophical meanings. In AI research, agents are often described without presupposing subjectivity or robust moral responsibility, whereas philosophy typically understands agents as members of moral and political communities. This divergence encourages conceptual slippage, as public (and sometimes even scientific) discourse frequently moves between these registers, attributing to AI systems "decisions," "choices," "behaviours," or even "intentions."

There is therefore an urgent need to clarify how terms such as *agent*, *artificial agency*, and *agentic AI* are being used.

This colloquium provided a structured space for interdisciplinary dialogue between researchers in philosophy and artificial intelligence. Its aim was to strengthen the theoretical intelligibility of contemporary discourse surrounding "AI agents" while fostering critical reflection on the social, political, and normative implications of artificial agency.

Table of Contents

1 Program 3
2 Presenters 4
3 Opening Remarks 5
4 Panel 1 5
4.1 Hugo Cossette-Lefebvre 5
4.2 Clémence Bergerot 6
4.3 Jeff Sebo 6
4.4 Discussion 6
5 Panel 2 7
5.1 Patricia Gautrin 7
5.2 Alexis Morin-Martel 8
5.3 Tegan Maharaj 8
5.4 Discussion 8
6 Panel 3 9
6.1 Audrey Durand 9
6.2 Dominic Martin 9
6.3 Anandi Hattiangadi 10
6.4 Discussion 10
7 Panel 4 11
7.1 Marion Korosec-Serfaty 11
7.2 Xabier E. Barandiran 11
7.3 Discussion 12
8 Closing remarks 12

1 Program

8h30 – 9h00: Registration and Welcome

9h00 – 9h30: Opening Remarks

- Aaron Courville (Université de Montréal)
- Daniel Weinstock (McGill University)

9h30 – 11h00: What are the minimal conditions (necessary and/or sufficient) for speaking of artificial agency from philosophical and technical perspectives? Which conditions matter most? How do computer scientists define an agent?

- Panelists: Hugo C. Lefebvre (McGill University), Clémence Bergerot (Humboldt University), Jeff Sebo (NYU)
- Chair: Jocelyn Maclure (McGill University)

11h00 – 11h15: Break

11h15 – 12h45: Is consciousness or phenomenological experience necessary for agency, or are behavioural properties sufficient? In other words, how can philosophical and technical criteria be compared without referring to different things under the same concept?

- Panelists: Patricia Gautrin (Université de Montréal), Alexis Morin-Martel (McGill University), Tegan Maharaj (HEC Montréal)
- Chair: Joé Martineau (HEC Montréal)

12h45 – 2h00: Lunch

2h00 – 3h30: What shared vocabulary could help avoid anthropomorphic slippage? Can concepts from the philosophy of action (intentionality, deliberation, acting for reasons, etc.) be translated into frameworks such as BDI (Belief–Desire–Intention) (Rao & Georgeff, 1991, 1998), decision theory, or LLM agents?

- Panelists: Audrey Durand (Université Laval), Dominic Martin (Université du Québec à Montréal), Anandi Hattiangadi (Stockholm University)
- Chair: Annie Pullen Sansfaçon (Université de Montréal)

3h30 – 3h45: Break

3h45 – 4h45: Under what conditions can an artificial system be legitimately considered an agent: is externally observed behavioral adequacy sufficient, or does attributing genuine agency require access to its internal states and constitutive mechanisms? What about collective agents?

- Panelists: Xabier E. Barandiaran (University of the Basque Country), Marion Korosec-Serfaty (Université du Québec à Montréal)
- Chair: Daniel Weinstock (McGill University)

4h45 – 5h00: Closing Remarks

- Daniel Weinstock (McGill University)

2 Presenters

Hugo Cossette-Lefebvre is a postdoctoral researcher at the Stephen A. Jarislowsky Chair in Human Nature and Technology at McGill University. His research examines equality, social relations, and the ethical implications of emerging technologies.

Clémence Bergerot is a PhD candidate at Humboldt University Berlin. Her research combines reinforcement learning, complex systems, and cognitive science to study agency, decision-making, and multi-agent dynamics.

Jeff Sebo is Associate Professor at New York University, where he works on ethics, moral status, animal cognition, and the ethical implications of emerging technologies.

Patricia Gautrin is a researcher at AlgoraLab (Mila) and a PhD candidate in AI ethics at Université de Montréal. Her research focuses on artificial moral agents, virtue ethics, and Ethics by Design.

Alexis Morin-Martel holds a PhD in philosophy from McGill University and will begin a SSHRC Postdoctoral Fellowship at the University of North Carolina at Chapel Hill in 2026. His research focuses on trust, technology, and moral responsibility in digital environments.

Tegan Maharaj is Assistant Professor in the Department of Decision Sciences at HEC Montréal and a member of IVADO's AI Safety and Alignment research cluster. Her research focuses on responsible AI, AI safety, and the governance of autonomous systems.

Audrey Durand is Associate Professor in the Department of Computer Science and Software Engineering at Université Laval, a Canada-CIFAR AI Chair, and an academic member of Mila. Her research focuses on reinforcement learning and interactive machine learning methods.

Dominic Martin is Professor of Ethics at ESG UQAM and Co-Director of the Ethics and Economics Axis at the Centre de recherche en éthique (CRÉ). His research focuses on business ethics, political philosophy, and the ethical challenges raised by artificial intelligence.

Anandi Hattiangadi is Professor of Philosophy at Stockholm University and a researcher at the Institute for Futures Studies. Her work focuses on philosophy of mind, language, and artificial intelligence.

Marion Korosec-Serfaty is Assistant Professor at ESG UQAM, a researcher at Tech3Lab, and a member of IVADO's AI Implementation and Responsible Governance research cluster. Her research examines human-AI interaction, explainability, and responsible AI use in professional decision-making.

Xabier E. Barandiaran is an Ikerbasque Associate Research Professor in the Department of Philosophy at the University of the Basque Country. His research explores agency, autonomy, and cognition through the lenses of philosophy of mind, cognitive science, and artificial intelligence.

3 Opening Remarks

Daniel Weinstock opened the colloquium by highlighting the central question motivating the event: whether philosophers and AI researchers mean the same thing when they speak of "agency." As the concept of agency occupies a foundational place in ethical reflection while increasingly becoming a central concept in AI research, he presented the colloquium as an opportunity to foster dialogue between disciplines and explore potential points of convergence and disagreement. He also emphasized the interactive format of the event, which was designed to prioritize discussion among panelists and participants.

Aaron Courville then introduced IVADO and its mission as an interdisciplinary consortium dedicated to advancing AI and its responsible adoption. He presented IVADO's role as a bridge between academic disciplines, industry, and public institutions, and situated the colloquium within the broader objectives of the R3AI initiative: Shifting paradigms for a robust, reasoning and responsible AI and its adoption.

4 Panel 1

Question: *What are the minimal conditions (necessary and/or sufficient) for speaking of artificial agency from both philosophical and technical perspectives? Which ones truly matter? How do computer scientists define an agent?*

Panelists: Hugo C. Lefebvre (McGill University), Clémence Bergerot (Humboldt University), Jeff Sebo (NYU).

4.1 Hugo Cossette-Lefebvre

Hugo Cossette-Lefebvre opened the first panel by examining whether the notion of agency has the same meaning in philosophy and AI research. Drawing on examples of increasingly sophisticated AI systems, including GPT-4's widely discussed CAPTCHA episode and emerging agentic systems capable of coordinating tools and performing complex tasks, he argued that behavioural sophistication alone does not settle the question of agency.

In philosophical contexts, he suggested, agency involves three core dimensions: interactivity, adaptability, and self-directedness. Agency is not merely the capacity to respond to an environment, but to pursue one's own ends in light of environmental constraints. By contrast, AI agents are typically understood in computer science as systems that perform goal-directed tasks using external tools, reasoning processes, and real-time information. While these systems exhibit interactivity and adaptability, their goals remain externally assigned rather than self-generated.

Cossette-Lefebvre argued that this distinction matters because the language used to describe AI systems shapes how people understand and interact with them. Referring to AI systems as agents may encourage anthropomorphic interpretations and create confusion regarding responsibility, trust, or moral standing. He concluded by noting that the normative questions raised by agency differ across contexts: while discussions of human agency often focus on autonomy and self-determination, discussions of AI agency tend to focus on alignment, control, and governance.

4.2 Clémence Bergerot

Clémence Bergerot approached the question of artificial agency through the lens of reinforcement learning. She proposed three criteria for agency: identity, understood as the distinction between an agent and its environment; interactional asymmetry, or the capacity to influence the environment; and normativity, namely goal-directed behaviour oriented toward outcomes that matter for the agent itself.

She argued that reinforcement learning captures some aspects of agency, since agents learn through interaction with their environment and adjust their behaviour to maximize rewards. In this framework, one can identify rudimentary forms of identity, normativity, and environmental influence. However, these features remain limited: the agent's goals are externally specified, its identity is architecturally defined, and its capacity to act remains constrained by the parameters established by designers.

Bergerot also discussed the notion of empowerment, a concept that measures the degree of control an agent can exert over its environment. While empowerment strengthens the criterion of environmental influence, it does so by shifting attention away from goal-directedness and toward control itself.

She concluded that current AI systems may instantiate some dimensions of agency, but only in a limited sense. Rather than treating agency as an all-or-nothing property, she suggested that artificial systems are better understood as exhibiting varying degrees of agency-like capacities.

4.3 Jeff Sebo

Jeff Sebo argued that debates about AI agency often suffer from treating agency as a single, all-or-nothing concept. Rather than asking whether AI systems are agents, he proposed distinguishing between different forms of agency and the moral questions associated with each.

He outlined a three-level framework. Minimal agency involves goal-directed behaviour, environmental interaction, and adaptive autonomy. Intentional agency adds beliefs, desires, and intentions, allowing an agent to represent the world and pursue preferred outcomes. Rational agency, associated most closely with adult humans, further includes higher-order reflection, self-assessment, and the capacity to revise one's beliefs and desires in light of reasons.

Sebo argued that these different forms of agency carry different moral implications. While minimal agency may be sufficient for describing a wide range of living systems, intentional and rational agency are more closely connected to questions of moral status and responsibility. He concluded that discussions of AI should avoid asking whether artificial systems are "really" agents and instead focus on the specific capacities that matter for the moral or conceptual questions under consideration. He also cautioned against applying stricter standards to AI than to humans, who themselves remain shaped by biology, socialization, and other external influences.

4.4 Discussion

The discussion broadened the scope of the panel considerably, moving beyond the question of AI agency to reflect on agency, autonomy, embodiment, and moral status more generally.

A first theme concerned whether human agency should serve as the benchmark for evaluating nonhuman or artificial systems. Participants questioned whether traditional philosophical accounts, often centred on reflective self-awareness and autonomy, capture a universal conception of agency or reflect a more specific cultural and historical understanding of personhood. Several interventions emphasized the importance of examining different forms of agency on their own terms rather than measuring them against an idealized human model.

A second theme focused on the origins of goals and the role of self-determination. While participants generally agreed that current AI systems pursue externally assigned objectives, the discussion highlighted the complex ways in which goals, behaviour, and environmental feedback interact in both biological and artificial systems. Related exchanges also emphasized the social dimensions of agency, noting that human agency develops through socialization and relationships rather than emerging fully formed in isolated individuals.

The discussion also returned repeatedly to the question of embodiment. Participants debated whether agency can be adequately understood in purely functional terms or whether it depends on forms of material embodiment and situated interaction with the world. While no consensus emerged, the exchange revealed a persistent tension between functionalist and biologically grounded approaches to agency.

Finally, participants considered the importance of temporality and history. Questions were raised about whether agency is shaped by developmental trajectories, evolutionary histories, or experiences of vulnerability and mortality. These discussions suggested that agency may be a more heterogeneous phenomenon than is often assumed, encompassing a range of capacities and forms rather than a single defining feature.

5 Panel 2

Question: *Is consciousness or phenomenological experience necessary, or are behavioural properties sufficient for agency? In other words, how can we compare philosophical criteria and technical criteria without talking about different things under the same concept?*

Panelists: Patricia Gautrin (Université de Montréal), Alexis Morin-Martel (McGill University), Tegan Maharaj (HEC Montréal).

5.1 Patricia Gautrin

Patricia Gautrin examined whether consciousness is a necessary condition for agency, contrasting philosophical accounts that tie agency to subjective experience with more functionalist approaches. Drawing on David Chalmers' notion of the philosophical zombie, she noted that if agency requires phenomenological consciousness, then even highly sophisticated AI systems may remain excluded from genuine moral agency.

She contrasted this position with computational and functionalist perspectives, according to which moral agency depends less on subjective experience than on the capacity to process information, evaluate alternatives, and make decisions in ethically relevant contexts. Building on the work of John Sullins, Gautrin suggested that artificial systems may acquire a form of morally relevant agency through their role in social environments and their ability to make autonomous decisions.

She concluded by proposing a virtue-ethical perspective on AI design. Rather than focusing on the possibility of artificial consciousness, researchers might seek to cultivate desirable behavioural dispositions within AI systems, allowing them to develop context-sensitive patterns of action without presupposing human-like subjective experience.

5.2 Alexis Morin-Martel

Rather than addressing the metaphysical question of whether AI systems are genuine agents, Alexis Morin-Martel focused on the social consequences of increasingly agent-like behaviour. He noted that large language models can now sustain coherent, persuasive, and socially fluent interactions, making them difficult to distinguish from human interlocutors in many contexts.

Drawing on P.F. Strawson's distinction between the participant stance and the objective stance, Morin-Martel argued that ordinary social life depends on treating others as answerable agents capable of explanation, apology, and accountability. The growing difficulty of distinguishing humans from AI systems, however, may encourage what he termed *agentic distrust*: a tendency to suspect that one's interlocutor is not a genuine agent at all.

This ambiguity, he suggested, creates a moral risk. When people withdraw the participant stance and begin treating others as mere objects to be managed or ignored, they may fail to recognize genuinely human interlocutors as participants in moral life. The problem, therefore, is not simply whether AI systems possess agency, but how increasingly convincing simulations of agency may reshape human relationships and patterns of trust.

5.3 Tegan Maharaj

Tegan Maharaj argued that many disagreements about AI agency stem from a problem of translation between philosophical and technical vocabularies. Rather than assuming that concepts such as agency or consciousness carry the same meaning across disciplines, she suggested that researchers should make explicit how these concepts are defined and operationalized in different contexts.

Drawing on recent work on increasingly agentic AI systems, Maharaj presented agency as a matter of degree rather than a binary property. She highlighted characteristics such as goal-directedness, long-term planning, direct impact on the world, and behavioural flexibility as useful technical indicators of agency. From this perspective, consciousness is not required for the forms of agency that matter in many governance and safety contexts.

She noted, however, that consciousness remains more difficult to translate into technical terms. Unlike agency, which can be operationalized through observable characteristics, consciousness lacks a widely accepted computational equivalent. As a result, attempts to connect philosophical and technical discussions of consciousness remain considerably more challenging.

5.4 Discussion

The discussion highlighted persistent disagreements about whether concepts such as consciousness, desire, and moral deliberation can be meaningfully translated into technical frameworks. While Patricia Gautrin defended a

functionalist approach in which moral decision-making does not require phenomenological consciousness, Tegan Maharaj emphasized the difficulty of finding technical equivalents for many human psychological concepts. More generally, participants disagreed on whether functional similarities are sufficient to justify the use of agency-related vocabulary in AI contexts.

Questions from the audience focused on self-preservation, mortality, and apparent resistance to shutdown in advanced AI systems. While some participants suggested that such behaviours might indicate emerging forms of agency, others argued that behavioural evidence alone does not justify attributing genuine desires or intentions to artificial systems.

The discussion also explored possible bridge concepts between philosophical and technical approaches. Several participants suggested that more specific notions, such as access to and integration of internal information, may be easier to operationalize than consciousness understood as a single, unified phenomenon. The exchange reflected a broader theme of the panel: the need to disaggregate complex philosophical concepts into more precise dimensions if meaningful dialogue between disciplines is to be achieved.

6 Panel 3

Question: *What shared vocabulary could help avoid anthropomorphic slippage? Can concepts from the philosophy of action (intentionality, deliberation, acting for reasons, etc.) be translated into frameworks such as BDI (Belief–Desire–Intention) (Rao & Georgeff, 1991, 1998), decision theory, or LLM agents?*

Panelists: Audrey Durand (ULavaL), Dominic Martin (UQAM), Anandi Hattiangadi (Stockholm University).

6.1 Audrey Durand

Audrey Durand opened the panel with a reflection on the growing use of anthropomorphic language in AI research. While acknowledging that she had no simple solution to the problem, she noted that researchers have long relied on human-centred metaphors, speaking, for example, of “training” models without implying that machines learn in the same way humans do.

She argued that the increasing sophistication of contemporary AI systems has made such language more problematic. Terms such as “wanting,” “thinking,” or “deciding” can easily blur the distinction between metaphorical descriptions and genuine claims about agency or intentionality. Her concern was less with anthropomorphic language itself than with the conceptual confusion that can arise when behavioural similarities are taken to imply deeper similarities in how humans and AI systems function.

6.2 Dominic Martin

Dominic Martin proposed a broad conception of agency based on three elements: the ability to represent aspects of the world, motivational states such as goals or preferences, and the capacity to act on the environment in light of those representations and motivations. On this view, the threshold for agency may be lower than is often assumed, reinforcing the need to think of agency as a matter of degree rather than a binary property.

He also cautioned that concepts such as consciousness, sentience, and thought remain imperfectly understood. Historically, conceptions of the mind have often been shaped by analogies with the most advanced technologies of the time, suggesting the need for humility when assessing artificial systems.

Martin concluded with two methodological recommendations. First, discussions of agency should avoid one-size-fits-all frameworks and instead recognize different levels and forms of agency. Second, debates about AI should begin with the normative questions at stake, such as responsibility, governance, or moral status, rather than with abstract metaphysical questions about whether AI systems are truly conscious or sentient.

6.3 Anandi Hattiangadi

Anandi Hattiangadi argued that current discussions of AI often rely on an overly behavioural conception of agency. While acknowledging that humans naturally interpret behaviour in intentional terms, she maintained that anthropomorphic language becomes problematic when behavioural resemblance is treated as evidence of genuine agency.

In her view, contemporary large language models are not agents in any meaningful sense. They do not possess beliefs, desires, intentions, or other psychological states typically associated with agency. Rather, they are sophisticated statistical systems that reproduce patterns found in linguistic data. The fact that their behaviour resembles that of agents does not imply that they share the underlying capacities that make agency possible.

Hattiangadi suggested that the more immediate risk is not that AI systems will become autonomous agents, but that humans will increasingly treat them as if they already are. This mismatch between how people interpret AI systems and how those systems actually function can foster misplaced trust, misunderstanding, and unrealistic expectations. She also questioned the common assumption that greater intelligence would automatically give rise to agency, arguing that behavioural sophistication alone does not establish the presence of genuine intentional states.

6.4 Discussion

The discussion focused on the challenges posed by anthropomorphic language in AI research. Participants generally agreed that the language used to describe AI systems matters, but disagreed about the nature of the problem.

One recurring theme concerned the relationship between behavioural resemblance and underlying mechanisms. Several participants cautioned against inferring that systems operate in the same way simply because they produce similar outputs. This issue became particularly salient in discussions of AI models designed to reproduce aspects of biological or social behaviour.

A second theme concerned the risks associated with anthropomorphic interpretations. Some participants argued that attributing agency, intentions, or understanding to AI systems can foster misplaced trust and unrealistic expectations. Others suggested that the significance of anthropomorphic language depends on the specific normative issues at stake, such as emotional dependency, public understanding, or the governance of AI systems.

The discussion concluded with a broader reflection on AI risk. Participants disagreed on whether agency-related language helps communicate the dangers associated with advanced AI systems or whether it obscures more immediate concerns related to human-AI interaction, institutional deployment, and public perception.

7 Panel 4

Question: *Under what conditions can an artificial system be legitimately considered an agent: is externally observed behavioral adequacy sufficient, or does attributing genuine agency require access to its internal states and constitutive mechanisms? What about collective agents?*

Panelists: Xabier E. Barandiran (University of the Basque Country), Marion Korosec-Serfaty (ESG UQAM).

7.1 Marion Korosec-Serfaty

Marion Korosec-Serfaty argued that neither behavioural performance nor access to internal system properties is sufficient to establish genuine agency in AI systems. Drawing on the Aristotelian distinction between *physis* and *techne*, she suggested that natural beings possess an internal source of purposiveness, whereas artifacts derive their purposes from external agents. Contemporary AI systems, in her view, remain closer to the latter category.

Using examples from professional decision-making contexts, she argued that human judgment involves more than producing appropriate outputs: it is a situated and accountable activity that can be justified and defended. While AI systems may increasingly resemble human judgment behaviourally, this resemblance does not establish the underlying conditions that make human judgment meaningful.

Korosec-Serfaty also presented research on human-AI interaction, suggesting that properties such as explainability, transparency, and reliability do not necessarily eliminate anthropomorphic interpretations. In some cases, they may even reinforce the impression that AI systems possess forms of reasoning or agency that they do not in fact have.

7.2 Xabier E. Barandiran

Xabier E. Barandiaran approached the question of AI agency from an ontological perspective, arguing that contemporary debates often rely too heavily on language-based and representational conceptions of mind. Instead, he proposed understanding agency through three core conditions: individuality, normativity, and interactional asymmetry.

On this view, genuine agents are entities for whom certain outcomes matter and that actively maintain themselves through interaction with their environment. Barandiaran argued that living organisms satisfy these conditions in ways that current AI systems do not. Although large language models display remarkable linguistic abilities, they lack the self-maintaining and normatively organized structures characteristic of biological agents. He therefore suggested that contemporary AI may be better understood as a form of intelligence without agency.

At the same time, Barandiaran emphasized that AI systems can profoundly transform human agency. By operating within the linguistic and social environments through which people coordinate action and make decisions, they reshape how agency is exercised, distributed, and delegated in contemporary societies.

7.3 Discussion

The discussion focused less on whether AI systems are agents and more on how institutions should govern systems that increasingly influence human judgment and decision-making. A central theme concerned the limits of the "human in the loop" model. Participants questioned whether formal human oversight is sufficient when users may gradually defer to systems that perform reliably most of the time, suggesting that responsibility must be understood not only at the individual level but also within broader institutional and socio-technical contexts.

Questions of explainability also featured prominently. Jocelyn Maclure raised concerns about the tension between the demand for meaningful human control and the opacity of many contemporary AI systems. In response, Marion Korosec-Serfaty discussed findings from her research suggesting that explainability can support critical engagement, but that certain forms of explanation may also encourage overreliance. More generally, she argued that users tend to engage more critically when faced with uncertainty about a system's recommendations.

The discussion concluded with a broader reflection on AI risk. While participants continued to disagree about the nature of agency, many emphasized the importance of focusing on the concrete social, cognitive, and institutional effects of AI systems rather than exclusively on speculative questions about future autonomous agents. More generally, the exchange reinforced a recurring theme of the colloquium: disagreements about AI agency often reflect deeper disagreements about the nature of agency itself.

8 Closing remarks

In their closing remarks, Daniel Weinstock and Thomas Adetou emphasized that the colloquium was intended not to resolve the question of artificial agency, but to foster dialogue across disciplinary perspectives. They highlighted the value of bringing together philosophers and AI researchers to explore a set of questions whose conceptual, technical, and normative dimensions remain open.

The event concluded with thanks to IVADO, the organizing team, speakers, moderators, technical staff, and participants whose contributions made the colloquium possible.