Title
Active inference: A method for phenotyping agency in AI systems?

Authors
Axel Constant[1*]
Jasmine Moore
Mahault Albarracin[2]
David Hyland
Jonathan Simon[5]
Karl J. Friston[5]
Andy Clark[1,6]

Institutions
1. Department of Engineering and Informatics, University of Sussex, Brighton, UK
2. Institut de Sante et Societe, Universite du Quebec a Montreal, Montreal, Quebec, CA
3. VERSES Research lab, Los Angeles, California, USA

***Correspondence**
axel.constant.pruvost@gmail.com
University of Sussex, School of Engineering and Informatics, Chichester I, CI-128, Falmer, Brighton, BN1 9RH, United Kingdom

Acknowledgement
AC was supported by a European Research Council, Synergy Grant (XSCAPE) ERC-2020-SyG 951631.

1 Introduction

The notion of "agency" in Artificial Intelligence (AI) is on everybody's lips, and for good reasons. In 2023, it was estimated that generative AI could add between $2.6 trillion and $4.4 trillion annually across a wide range of sectors in the financial and real economy. By 2024, however, the limitations of generative AI cast doubt on its ability to deliver on that promise. While generative AI excelled at leveraging historical data to make predictions, it revealed fundamental limitations. It lacked a genuine understanding of the world grounded in a "world model", was unable to engage in real-time reasoning, and struggled to adapt to novel scenarios. Since 2025, recognizing these limitations, the AI community has been working toward a new frontier: models that can understand the world, navigate edge cases autonomously, exhibit genuine reasoning, and operate independently. Considered by many to be the primary trend in AI for 2025, and the focus of commercial applications when cast as agentic "workflows" for the foreseeable future, the new approach of "agentic" AI has been presented as holding the potential to inaugurate a new age of intelligence, if properly steered.

But is AI agency merely another stage in the hype cycle, or is it the right problem set to solve to finally deliver on AI's promise? We believe that the search for AI agency is more than a passing trend. It is a core challenge whose resolution would address AI related issues faced both by AI researchers and by society at large. For researchers, solving AI agency would mean tackling reasoning, autonomy, and explainability, which are fundamental components of agency broadly construed. For society, it would mean creating AI systems to which we could reasonably ascribe intentions and motivations, and consequently ethical and legal responsibilities and liabilities. AI agency lies at the intersection of four critical issues in AI: (1) reasoning, (2) autonomy, (3) explainability, and (4) governance, and addressing agency entails making substantial progress on each of these fronts. Solving AI agency presents several challenges, the first of which is to arrive at a useful definition. Such a definition must be sophisticated enough to capture the diverse dimensions of agency as we understand it in humans, yet simple enough to be operationalized, measured in AI systems, and aligned with our intuitions so as to support practices of responsibility ascription. The second challenge is to provide a computational framework for decision-making capable of implementing the criteria laid out in such a

definition. This paper seeks to provide such a definition and argues that the theory of active inference, developed in theoretical neurobiology[1], meets these requirements.

Our aim is to recast a basic definition of agency into the formal language of active inference to show that an active inference model can meaningfully be defined as an agent as per the basic definition. A methodological advantage of active inference is that it can be used for a practice known as "computational phenotyping" in computational psychiatry[2,3]. This practice involves building an agent based active inference model that mimics known behavioral symptoms of a mental disorder, to then identify what parameters of the model contribute to explaining the behavioral symptoms; the assumption being that the active inference model is a biologically plausible map of the factors involved in generating the behavioral symptoms (e.g., neurobiological factors). By analogy, and under the assumption that we are correct in claiming that active inference models also provide a "map" of the basic structure of agency, the approach that we develop in this paper provides a way to phenotype computational traits of agency that would qualify an AI system -- active inference based or else -- as being endowed with agency. We show how active inference agents can be governed by accounting for necessary agency in the method of control applied. (§1),(§2),(§3),(§4),(§5),(§6)

2 Why the current definition of AI agency is not enough

Our current understanding of AI agency is limited. But it did not have to be. At the turn of the century, in their seminal paper on the notion of AI agency, Michael Wooldridge and Nicholas Jennings[4], offered two definitions of AI agency, one weak, and one strong, only one of which having appeared to survive. The weak definition framed AI agency as a "software based computer system" endowed with 4 properties: (1) autonomy (i.e., independence of execution), (2) social ability (i.e., ability to communicate with humans), (3) reactivity (i.e., ability to perceive and act), and (4) pro-activeness (i.e., goal directedness of action rather than mere reaction). In turn, the "strong" definition added to these criteria typically human-related capacities, such as knowledge, belief, intention, obligation, and emotions -- Wooldridge and Jenning noting that those notions are of particular interest for AI researchers.

The definition of AI agency considered "essential"[5], and that AI researchers have inherited has been the weak one[6], with a particular focus on the notion of "autonomy" and goal-directedness. Various "levels of autonomy" frameworks are routinely referred to as short hands for agency, capturing the common assumptions that agency refers to "the degree to which a system can adaptably achieve complex goals in complex environments with limited direct supervision"[7]. One such framework is the SAE Levels of Driving Automation for self-driving cars[8]. Another is the autonomy levels for unmanned systems proposed by the U.S. Department of Defense and the ALFUS framework (Autonomy Levels for Unmanned Systems)[9]; these frameworks assigning a rank or category to indicate how much a system can do in the absence of human intervention.

While useful, levels of automation are inherently coarse and situation-dependent. Linear "levels" oversimplify the problem, implying autonomy is a single spectrum, whereas in reality autonomy has multiple dimensions[10]. A robot might be highly autonomous in a structured factory floor (simple environment) but not in a chaotic public space; or it might handle navigation autonomously yet still rely on humans for goal-setting. As the National Institute of Standards and Technology (NIST) emphasizes it, "autonomy definitions and measures must encompass many dimensions and serve many audiences", from engineers to end-users[11]. Much of these dimensions, it appears, are reflected by the criteria of the strong definition of AI agency, such as the ability to leverage knowledge and contextually sensitive beliefs, intention or plans, emotions or affects and an understanding of one's obligations, towards the autonomous accomplishment of a goal.

Autonomy and goal directedness are simply not able to do all the work that we would want a concept of agency to do. For example, in a fully embedded agent, reactivity and pro-activity are not as clearly delineated as we may think, as any agent that exists has intrinsic goals. Those "inessential" properties part of the strong definition of AI agency identified by Wooldridge and Jennings may not be so inessential after all. This is so because we have certain intuitions as to what ought to count as agentic behavior, based on how we understand

such behavior in humans, and those intuitions go beyond the mere notion of "autonomous" conduct. That said, a satisfying definition of agency should be able to strike the right balance between strong and weak criteria, avoiding anthropomorphic definitions of agency that may be too difficult to implement in AI systems, yet moving beyond a too simple understanding that fails to reflect our intuitions about agency.

3 Rationality, Intentionality, and Explainability: A simple but not too simple view of agency

We believe that stepping back and looking at more fundamental conceptions of agency such as those developed by philosophers will help in reaching a parsimonious definition. At its core, agency implies the capacity for intentional action – the ability of an actor to initiate and direct events. While different philosophical approaches emphasize different aspects of relevance for agency, most agree on the relevance of some elementary components, at least as a foundation on which one can start building more elaborated definitions. Traditionally, for philosophers – and perhaps most famously for the American philosopher Donald Davidson, agency may be said to have referred specifically to the capacity to act **(1) rationally**, and **(2) intentionally**. Here, acting rationally means acting so as to achieve an outcome that should follow under some standards of rationality (e.g., deductive logic, see box 2 for different definitions of rationality that can apply). In turn, acting intentionally means acting in accordance with some mental states (e.g., beliefs, desires, intentions/plans) that correspond to the rational construction from which the intended outcome is derived (e.g., that corresponds to the premises of a syllogism), and where can will "weigh" their options by balancing out their beliefs and desires (e.g., having an ice cream despite knowing that it is bad for you). Additionally, one should expect that the mental states that are the basis of the rational action be also those that "causes" the agent to act. Causality is what allows for the explainability of agents' actions, as it reveals the agent's mental states in relation to its action.

Under the standard philosophical view, agency is the capacity to perform actions that are **(1) intentional** (i.e., based on mental states such as beliefs and desires), **(2) rational** (i.e., geared towards outcomes that should rationally -- e.g., deductively -- follow from the action and realize mental states), and **(3) explainable**, or explained causally based on an existing relationship between the mental states and the action. For instance, "turning on the light" to "enlighten the room" (rational outcome) to "get to the fridge" (intention/plan) because "you want to eat" (desires) and "you believe the fridge is in the kitchen" (beliefs) is the kind of action that, on the one hand, would qualify one as an agent, or with having agency, and on the other hand, would allow someone else to explain why one turned on the light. Now, intentionality and rationality come in degrees, which explains why we tend to attribute degrees of agency to different systems capable of acting. For instance, intuitively, many will agree that dogs have more agency than a bacterium, but less agency than humans. As a first approximation, one can imagine that level "1" agency may correspond to the basic ability to take rational and intentional action (e.g., a bacterium following a preferred chemical gradient while minimizing energy expenditure). This will involve goal directedness and beliefs based decision making, as well as some basic option weighing abilities. At a level "2", agent weight and make those intentional decisions while taking into account beliefs about more elaborated conception of the self (e.g., I believe that someone as disciplined as myself should not have ice cream), and level "3" will involve beliefs about the self in relation to others (e.g., I believe that my parents value discipline, and I and a disciplined person, therefore I won't have ice cream).

On the traditional account, agents perform actions that are intentional, rational and causally explainable. But what about autonomy? *Autonomy* (literally "self-governance") has been central to modern moral and political thought, but philosophers sharply disagree on its precise meaning. Some traditions view autonomy as an individual's capacity to live according to self-chosen reasons and not by external coercion (e.g. Kantian ethics or Millian liberalism). Others, including relational[12] or feminist theories[13], argue that classic notions of autonomy overemphasize individualism and neglect the social context, prompting re-conceptions of autonomy as fundamentally relational or procedurally defined by authenticity of one's choices. Despite these disputes, the notion of autonomy always traces back to independence of action -- whether it is due to oneself or others, which itself is tied to an individual or a collective's set of intentions -- desires, beliefs, and ability to plan. For instance, an action performed on behalf of someone else is not fully autonomous as it will in large part conform to the mental states -- beliefs, desires, and intention -- of the other person. In the parlance of

vehicle autonomy, according to the SAE Levels of Driving Automation, moving from level 0 to level 5 means gradually shifting the basis of intentional decision of the car-driver system from the mental states of the driver to the "mental states" of the car. An autonomous action is an intentional action, one derived from your mental states or from mental states like yours. Thus, one can use, somewhat interchangeably, the notion of autonomy and intentionality when qualifying an action as autonomous.

There is another concept frequently associated with agency: Adaptivity. It is commonplace to associate the notion of agency with that of adaptivity to a complex environment (e.g., more or less deterministic environment), where degrees of agency are measured as level of adaptivity. Interestingly, the notions of adaptivity and rationality are closely related[7]. They are both based on the learning and updating of a rational structure with more or less complex counterfactual depth connecting mental states to states of the world -- sometimes referred to as a world model, and based on which an inference can be performed (e.g., If I switch on the light and walk to the fridge, I'll get there fine, but that will hurt my eyes, and if I keep the light shut, I will be fine but I could hit my toe, but …). This adaptivity is ultimately aimed towards an implicit objective, which is the phenotypical identity of the entity attempting to persist (Albarracin, & Sakthivadivel, 2025). This rational structure, or world model will afford different degrees of goal complexity, or degrees of complexity or diversity of achievable goals matching different levels of environmental complexity. To be more or less rational means deriving the consequences of one's action from a more or less complex and accurate world model, which will make you more or less adaptive. A world model is what guarantees explainability, such a model causally connecting your mental states -- beliefs, desires, intentions/plans -- to your actions, through an inference process or a rational process of deduction. In the remainder of this paper, we explore the world model that active inference agents are endowed with, and what makes this world model amenable to agentic action as per the threefold definition of agency as intentional, rational, and explainable action.

5 World models in active inference and agency

We suggested that a useful and simple -- but not too simple -- definition of AI agency should include the following criteria:

(1) Intentionality/autonomy: The ability to ground one's actions in one's own mental states such as beliefs, desires, intentions or plans;

(2) Rationality/adaptivity: The ability to make decisions that rationally (e.g., logically or probabilistically) follow from one's understanding of the world.

(3) Explainability: The ability for an observer to explain the system's action as causally related to its mental states and understanding of the world.

Intentionality/autonomy -- or autonomy, rationality/adaptivity -- or rationality, and explainability can all be meaningfully expressed as attributes of actions performed by systems endowed with agency. And all of these attributes appear to be traceable to a detailed understanding of the world. In AI, such an understanding is often argued as being derivable from a "world model" . There have been many proposals of "world models" for AI, especially in the last few years, and many of them are "explainable" in virtue of their transparent architecture. But few of them express clearly the relation between intentionality and rationality criteria of agency defined above. In a recent extensive review, Ding and colleagues [14] provide a categorization of world models along two dimensions: (i) world model designed to "construct[ing] implicit representations to understand the mechanism of the external world" (a.k.a. "internal representation") and (ii) world models designed to "predicting future states of the external world" (a.k.a. "future prediction"). The first category "focuses on the development of models that learn and internalize world knowledge to support subsequent decision-making" and the latter "emphasizes enhancing predictive and simulative capabilities in the physical world from visual perceptions". Ding and colleagues further note that:

> "*Whether focusing on learning internal representations of the external world or simulating its operational principles, these concepts coalesce into a shared consensus: the essential purpose of a*

world model is to understand the dynamics of the world and compute the next state with certainty (or with some guarantee), which empowers the model to extrapolate longer-horizon evolution and to support downstream decision-making and planning"(p.4)

We will not review all the types of world models presented by Ding et al. We will simply reuse their basic categorization to help us unpack the working of active inference models. That said, we will note that the type of active inference world model that we present in this section would fall under the category of "internal models" specialized for decision making according to Ding et al. Among those, we find Reinforcement Learning (RL) model based approaches such as Markov Decision Process (MDP) for learning and action policy selection. These models involve formulating an MDP for specific decision making problems whose solution will maximize a reward function through learning the transition dynamics and emission probabilities of the world. The active inference models we focus on in this section -- Partially Observable MDPs, or POMDPs [15] -- are close in spirit to such RL based approaches, but are different in practice as they do not involve the specification of a typical reward function (for a discussion of the similarities and difference active inference and reinforcement learning, see [16–18]). Active inference POMDPs are of particular interest because (i) they are clear instantiations of world models with representational and predictive capabilities, (ii) as we will see in the final section of this paper, they are amenable to a practice of computational phenotyping that may be useful in the detection of agency in AI systems more broadly, and (iii) because they derive their representational and predictive capacities from a model with parameters and inference processes that speak directly to the criteria of agency listed above. We will see that they involve the formation of beliefs about the world that are combined with "desires", or preferences to drive the selection of action according to principles of rationality -- akin to those of economics -- and in a way that is fully explainable (see figure x.)

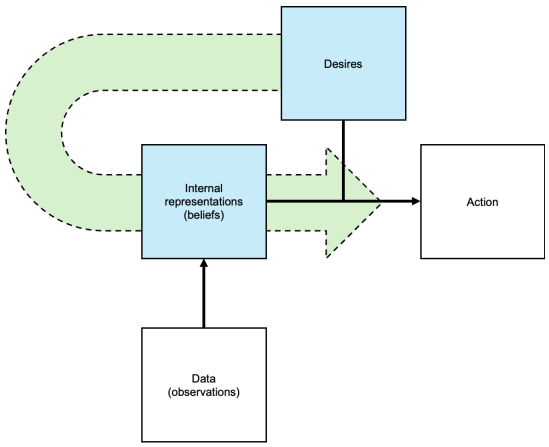

Figure x. Relation between rationality and intentionality in active inference world models. Beliefs and desires (blue boxes) are combined to select action in a rational way (green arrow) according to principles of economics, based on evidence, or observed data.

5.1 Active inference POMDPs

5.1.1 representations of the world with active inference

An active inference world model is held in the state transition matrix and the likelihood matrix. These represent the inherent likelihood of transition between states given an observation, and the inherent reliability of the observation itself to be true. These form a partially-observable markov decision process about the world, which is congruent with other learned world models such as UPDP[25] or TD-MCP2[26]. Modeling the world is an intrinsic part of an active inference agent, and it cannot proceed without some belief (however correct) about the world - it is the model itself, not the correctness, that is important.
Effectively, agents break reality at its joints by breaking possibly continuous inputs into discrete categories. Such categories then form the basis of reasoning. Once categories are created, the world can be mapped from observations to hidden causes, and coarse grained into various hierarchical levels. In this way, likelihood mappings can be aggregated to form longer and larger spatio-temporal scales, which translate to the transition matrices.

5.1.2. Predicting the future states and selecting action with active inference

Using these transition matrices, the agent can try out experiments. Specifically, the agent can choose to act such that it can test out whether its expected outcome will come about given that it manipulates some latent states of the environment. The agent is endowed with preferences, or outcomes it believes to be most amenable to its survival.
Our action selection is based primarily on the desirability of *observations*, rather than on the desirability of *states*. This allows us to provide rewards based on subjective, rather than objective, experience - which is important for autonomous agents without a state oracle for external states to self-determine rewards. Or, for short, it allows for *intrinsic* reward, rather than extrinsic motivation.

5.2 Intentionality, rationality, and explainability in active inference

5.2.1 Intentionality

Preference, beliefs and desires[19] (predictive minds can be Humean)

5.2.2 Rationality

Strict Bayesianism is inherently rational. Active inference agents cannot be *irrational* - they can only operate according to parameters other than those we assume they might. As all actions proceed from desires and interior beliefs about observations, and the reliability of those observations, active inference agents are inseparably rational.

5.2.3 Explainability

Monosemanticity of observation-dependent variables at planning levels leads to divisible features. (Is this sufficient to cover the case of unlabelled data?)

6 Worked example of agency phenotyping

6.1 The task

https://pubmed.ncbi.nlm.nih.gov/27517087/

Produce a simulation of an agent with more or less agency - intentionality, less rationality, less explainability

How do various governance controls work or fail at different levels of agency?

We don't have enough data for this.

There ARE phenotypes of agency - we're going to demonstrate that we can get to them.

What governance controls do we *have* for an agent at this level? Is setting the internal desirability of observations a sufficient level of control? What trade-offs do we have for setting the desirability of intermediate goal states versus oracle-feedback? Does oracle-feedback exist?

Intrinsic - physical blocking - extrinsic (direct) - extrinsic (social)

Intrinsic - mind control - direct manipulation of preferences (endowing them with empathy?)

Physical blocking - having states in the action space that the agent "can" go but can't actually go to. Coded externally. Done at the simulator level. Going left without a left space being an available action.

Extrinsic - an additional observation? For external input? And weight it somehow? (Do we?)

Show the phenotyping

6.2 The model

6.4 Results

6.5 Discussion

7 AI agency, reasoning, autonomy, and governance

The notion of agency is important for several reasons. We need it to explain other people's actions. It is the basis to determine if people should be considered responsible for their actions. And it is the basis for determining what range of actions should be allowed to certain people.

AI agents, however, may pose unprecedented risks. AIS endowed with full agency may one day set its own goals, generalize tasks across domains, and provide solutions to challenges beyond human capability. This progress will not come without challenges. Concerns regarding alignment and control will emerge to the forefront. As AI advances, policymakers will need to wrestle with a critical question - how do we govern AI systems capable of operating independently, setting their own priorities, and making decisions without direct human control? Contemporary AI governance strategies will ultimately fall short of providing a solution to this inevitable problem: The agency problem of AI governance.

REFERENCES

1. Parr, T., Pezzulo, G. & Friston, K. J. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. (MIT Press, 2022).
2. Montague, P. R., Dolan, R. J., Friston, K. J. & Dayan, P. Computational psychiatry. *Trends Cogn. Sci.* **16**, 72–80 (2012).
3. Schwartenbeck, P. & Friston, K. Computational Phenotyping in Psychiatry: A Worked Example. *eNeuro* **3**, (2016).

4. Wooldridge, M. & Jennings, N. R. Intelligent agents: theory and practice. *Knowl. Eng. Rev.* **10**, 115–152 (1995).
5. Jennings, N. R., Sycara, K. & Wooldridge, M. A roadmap of agent research and development. *Auton. Agent. Multi. Agent. Syst.* **1**, 7–38 (1998).
6. Luck, M. A conceptual framework for agent definition and development. *Comput. J.* **44**, 1–20 (2001).
7. Shavit, Y. *et al.* Practices for governing agentic AI systems.
8. Vyas, V. & Xu, Z. Key safety design overview in AI-driven autonomous vehicles. *arXiv [cs.SE]* (2024).
9. Huang, H.-M., Pavek, K., Albus, J. & Messina, E. Autonomy levels for unmanned systems (ALFUS) framework: an update. in *Unmanned Ground Vehicle Technology VII* (eds. Gerhart, G. R., Shoemaker, C. M. & Gage, D. W.) (SPIE, 2005). doi:10.1117/12.603725.
10. Stayton, E. & Stilgoe, J. It's time to rethink levels of automation for self-driving vehicles. *SSRN Electron. J.* (2020) doi:10.2139/ssrn.3579386.
11. Huang, H.-M. *et al.* Autonomy Measures for Robots. in *Dynamic Systems and Control, Parts A and B* (ASMEDC, 2004). doi:10.1115/imece2004-61812.
12. Oshana, M. Relational Autonomy. *International Encyclopedia of Ethics* 1–13 Preprint at https://doi.org/10.1002/9781444367072.wbiee921 (2020).
13. Mackenzie, C. Feminist conceptions of autonomy. in *The Routledge Companion to Feminist Philosophy* 515–527 (Routledge, 1 [edition]. | New York : Routledge, 2017. | Series: Routledge philosophy companions, 2017).
14. Ding, J. *et al.* Understanding world or predicting future? A comprehensive survey of world models. *arXiv [cs.CL]* (2025).
15. Da Costa, L. *et al.* Active inference on discrete state-spaces: A synthesis. *J. Math. Psychol.* **99**, 102447 (2020).
16. Tschantz, A., Millidge, B., Seth, A. K. & Buckley, C. L. Reinforcement learning through active inference. *arXiv [cs.LG]* (2020) doi:10.48550/ARXIV.2002.12636.
17. Friston, K. J., Daunizeau, J. & Kiebel, S. J. Reinforcement learning or active inference? *PLoS One* **4**, e6421 (2009).

18. Da Costa, L., Sajid, N., Parr, T., Friston, K. & Smith, R. Reward maximization through discrete active inference. *Neural Comput.* **35**, 807–852 (2023).

19. Junker, F. T., Bruineberg, J. & Grünbaum, T. Predictive minds can be humean minds. *Br. J. Philos. Sci.* (2024) doi:10.1086/733413.

20. Limanowski, J. & Blankenburg, F. Minimal self-models and the free energy principle. *Front. Hum. Neurosci.* **7**, 547 (2013).

21. Friston, K. J. & Frith, C. D. Active inference, communication and hermeneutics. *Cortex* **68**, 129–143 (2015).

22. Vasil, J., Badcock, P. B., Constant, A., Friston, K. & Ramstead, M. J. D. A world unto itself: human communication as active inference. *Frontiers in psychology* **11**, 417 (2020).

23. Veissière, S. P. L., Constant, A., Ramstead, M. J. D., Friston, K. J. & Kirmayer, L. J. Thinking through other minds: A variational approach to cognition and culture. *Behav. Brain Sci.* **43**, e90 (2019).

24. Albarracin, M., Constant, A., Friston, K. J. & Ramstead, M. J. D. A variational approach to scripts. *Frontiers in Psychology* **12**, 585493 (2021).

25. Du, Y., Yang, M., Dai, B., Dai, H., Nachum, O., Tenenbaum, J. B., Schuurmans, D., & Abbeel, P. (2023). *Learning Universal Policies via Text-Guided Video Generation* (No. arXiv:2302.00111). arXiv. https://doi.org/10.48550/arXiv.2302.00111

26. Hansen, N., Su, H., & Wang, X. (2024). *TD-MPC2: Scalable, Robust World Models for Continuous Control* (No. arXiv:2310.16828). arXiv. https://doi.org/10.48550/arXiv.2310.16828

Active Inference: A Method for Phenotyping Agency in AI Systems?

1 Introduction

The concept of **agency in artificial intelligence** has become a central topic in contemporary research and governance debates. Early economic forecasts suggested that generative AI might contribute between **2.6 and 4.4 trillion USD** annually to global productivity (McKinsey, 2023), yet by 2024, its limitations became apparent. While large language models excelled at predictive pattern completion, they lacked genuine *world understanding*, contextual reasoning, and the capacity to adapt in real time to novel circumstances. In consequence, attention has shifted toward a new paradigm, that of **agentic AI**. This paradigm aspires to model systems capable of understanding, reasoning, and acting autonomously within dynamic environments.

The promise of agentic AI is profound: it could unify progress on four interrelated challenges namely **reasoning, autonomy, explainability, and governance**. However, progress depends first on establishing an operational definition of *agency* that is both theoretically rigorous and computationally tractable. In this paper we propose such a definition and demonstrate that the **active inference framework** (Friston et al., 2022) provides the requisite formalism to implement it. We recast minimal philosophical criteria for agency (*intentionality*, *rationality*, and *explainability*) within the mathematics of active inference and propose a methodology for **computational phenotyping of agency**, adapted from analogous work in computational psychiatry (Montague et al., 2012; Schwartenbeck and Friston, 2016).

Active inference enables systematic detection because it treats agency as an observable behavioral phenotype, measurement, and comparison of agentic capacities in both active-inference and non-active-inference AI

systems. We get a foundation for assessing the *agency problem* in AI governance such as how to attribute responsibility, liability, and oversight when artificial systems display behavior that warrants the term "intentional."

2 Why Existing Definitions of AI Agency Are Insufficient

Foundational work by **Wooldridge and Jennings (1995)** distinguished between *weak* and *strong* notions of artificial agency. The *weak* definition emphasized autonomy, social ability, reactivity, and pro-activity, whereas the *strong* definition added cognitive and affective properties such as knowledge, belief, intention, and emotion. Over time, AI research largely adopted the weak form, reducing agency to degrees of *autonomy*—the system's capacity to achieve goals with minimal human supervision (Jennings et al., 1998; Luck, 2001).

This reduction underpins common taxonomies such as the **SAE J3016** levels of driving automation (Vyas and Xu, 2024) and the **ALFUS** framework for unmanned systems (Huang et al., 2005). These schemes are valuable for engineering certification but treat autonomy as a one-dimensional continuum, obscuring the *multidimensional structure* of agency (Stayton and Stilgoe, 2020; Huang et al., 2004). In practice, an autonomous vehicle may operate independently in structured settings yet depend on human input for goal formulation; autonomy alone fails to capture this nuance.

Furthermore, autonomy levels ignore mental-state constructs central to human action explanation—beliefs, desires, and intentions. These constructs, which Wooldridge and Jennings (1995) regarded as "inessential," are in fact indispensable for distinguishing merely automated from genuinely agentic behavior. A satisfactory framework must therefore balance the operational simplicity of weak definitions with the conceptual richness of strong ones, avoiding anthropomorphism while preserving explanatory adequacy.

3 Intentionality, Rationality, and Explainability: A Minimal Philosophy of Agency

Philosophers from **Davidson (1963)** onward have defined agency as the capacity to perform *intentional and rational* actions whose causes are *mental states*. In this classical view, an action is *intentional* when it arises from beliefs, desires, and intentions; it is *rational* when its outcome follows from those states under a norm of coherence; and it is *explainable* when observers can trace a causal chain from mental state to behavior (Davidson, 1980).

Degrees of agency correspond to the complexity and coherence of these internal representations. Simple biological organisms (say, bacteria following a nutrient gradient) exhibit minimal intentionality and rationality, while humans express multi-layered models incorporating self-concept and social expectations. **Autonomy**, literally "self-governance," can be interpreted as the capacity to act according to one's own internal states rather than external coercion (Oshana, 2020; Mackenzie, 2017). Whether viewed through Kantian, Millian, or relational lenses, autonomy is always rooted in control over the beliefs, desires, and intentions that structure action.

A further dimension, **adaptivity**, links agency to the capacity to revise beliefs in response to environmental change. Adaptivity and rationality are closely related: both depend on *world-model depth*—the agent's ability

to represent counterfactuals and evaluate alternative futures (Albarracin and Sakthivadivel, 2025). Hence, we propose a minimal working definition:

1. *Intentionality/autonomy*—actions grounded in the agent's own beliefs, desires, and plans;
2. *Rationality/adaptivity*—actions that probabilistically follow from an internal world model;
3. *Explainability*—actions that are causally traceable to those internal states.

This tripartite structure serves as a bridge between philosophical and computational accounts of agency.

4 Active Inference as a World Model of Agency

Recent surveys distinguish **world models** that internalize knowledge for decision-making from those that merely predict future states (Ding et al., 2025). Active inference clearly belongs to the former class: it provides an internal generative model that integrates representation, prediction, and control (Friston et al., 2022).

In active inference, the environment is modeled as a **partially observable Markov decision process** (Da Costa et al., 2020). The agent maintains probabilistic beliefs about hidden states, learns parameters from sensory evidence, encodes *preferences* over outcomes, and selects *policies* that minimize **expected free energy,** a single quantity unifying epistemic (information-seeking) and pragmatic (goal-seeking) imperatives (Friston et al., 2009; Tschantz et al., 2020; Da Costa et al., 2023). Unlike reward-maximization in reinforcement learning, EFE minimization entails balancing exploration and exploitation within one variational objective.

Conceptually, beliefs correspond to *mental representations* of the world, preferences to *desires*, and policies to *intentions*. The causal chain "belief + desire → intention → action" is therefore explicit within the model, providing built-in **explainability**. This architecture renders action selection simultaneously intentional (since it derives from the agent's own preferences), rational (since it optimizes expected outcomes under a coherent model), and explainable (since the underlying factors are transparent), beliefs and desires (blue nodes) jointly determine action (white node) through rational inference (green arrow), linking internal mental states to observable behavior. The result is a formal bridge between philosophical and computational definitions of agency.

5 Computational Phenotyping of Agency

Because active inference models are generative, they can be inverted to infer latent parameters from behavior, a process known as **computational phenotyping** (Schwartenbeck and Friston, 2016). In psychiatry, this approach identifies neural or cognitive parameters explaining individual differences in behavior (Montague et al., 2012). Analogously, **agency phenotyping** infers parameters such as policy precision, preference strength, or epistemic-pragmatic balance, yielding quantitative *profiles of agency*.

A typical pipeline involves specifying a POMDP with hidden states, observations, preferences, and policies;

fitting it to behavioral data; and extracting interpretable parameters that mark degrees of intentionality, rationality, and explainability. Such profiles allow systematic comparison between architectures—active-inference agents, reinforcement learners, or heuristic planners—revealing where each satisfies or fails the threefold criterion. By mapping agency onto measurable parameters, phenotyping transforms a philosophical notion into an empirical construct amenable to auditing and governance.

6 From Individual to Collective Agency

Degrees of agency extend beyond the individual. The active-inference literature models the **self** as a dynamically maintained boundary condition (Limanowski and Blankenburg, 2013), **communication** as coupled inference between agents (Friston and Frith, 2015; Vasil et al., 2020), **shared intentionality** as multi-agent policy alignment (Veissière et al., 2019), and **cultural scripts** as higher-order priors structuring social behavior (Albarracin et al., 2021). Each level—from minimal self-maintenance to collective cognition—can be analyzed through the same free-energy-minimization principle, enabling scalable models of social and institutional agency.

7 Implications for AI Governance

The **agency problem in AI governance** arises precisely because legal and ethical accountability presuppose an identifiable agent. As AI systems gain autonomy in setting and pursuing goals, traditional control mechanisms become insufficient. Active-inference-based phenotyping offers a principled route to quantify agency, furnishing regulators with auditable evidence of intentionality, rationality, and explainability. These profiles could complement existing autonomy-level standards (SAE; ALFUS) by introducing *multidimensional agency maps* that better capture risk and capability boundaries (Shavit et al., 2024).

Active inference is a transparent basis for *explainable autonomy* because it links internal model structure to observable behavior, a necessary step toward aligning artificial systems with human norms and values. More broadly, phenotyping could inform certification procedures, human-machine teaming, and the design of oversight architectures suited to increasingly agentic AI.

8 Conclusion

We have proposed that the search for AI agency is not a transient trend but a core scientific and governance challenge. A minimal definition, intentional, rational, and explainable action, which captures the essential dimensions of agency without anthropomorphism. The **active inference framework** operationalizes these dimensions within a single generative model, providing both an explanatory theory of agency and a practical tool for its measurement.

Through **computational phenotyping**, agency becomes an empirical variable that can be detected, compared, and governed. This bridges the gap between philosophical analysis and engineering practice, offering a coherent path toward responsible deployment of truly agentic AI systems.

Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active inference: The free-energy principle in mind, brain, and behavior*. MIT Press. https://doi.org/10.7551/mitpress/12441.001.0001 direct.mit.edu+1

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences, 16*(1), 72–80. https://doi.org/10.1016/j.tics.2011.11.018 PubMed+2PMC+2

Schwartenbeck, P., & Friston, K. J. (2016). Computational phenotyping in psychiatry: A worked example. *eNeuro, 3*(4), ENEURO.0049-16.2016. https://doi.org/10.1523/ENEURO.0049-16.2016 PubMed+1

Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *The Knowledge Engineering Review, 10*(2), 115–152. https://doi.org/10.1017/S0269888900008122 Cambridge University Press & Assessment

Jennings, N. R., Sycara, K., & Wooldridge, M. (1998). A roadmap of agent research and development. *Autonomous Agents and Multi-Agent Systems, 1*(1), 7–38. https://doi.org/10.1023/A:1010090405266 link.springer.com

Luck, M. (2001). A conceptual framework for agent definition and development. *The Computer Journal, 44*(1), 1–20. https://doi.org/10.1093/comjnl/44.1.1 OUP Academic

Shavit, Y., Agarwal, S., Sastry, G., O'Keefe, C., Robinson, D. G., & Coyle, D. (2023). *Practices for governing agentic AI systems* (White paper). OpenAI. https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf cdn.openai.com+1

Vyas, V., & Xu, Z. (2024). Key safety design overview in AI-driven autonomous vehicles. *arXiv*. https://arxiv.org/abs/2412.08862 arXiv

Huang, H.-M., Pavek, K., Albus, J. S., & Messina, E. (2005). Autonomy levels for unmanned systems (ALFUS) framework: An update. In G. R. Gerhart, C. M. Shoemaker, & D. W. Gage (Eds.), *Unmanned Ground Vehicle Technology VII* (Proc. SPIE 5804, pp. 439–450). SPIE. https://doi.org/10.1117/12.603725 proceedings.spiedigitallibrary.org

Stayton, E., & Stilgoe, J. (2020). It's time to rethink levels of automation for self-driving vehicles. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3579386 scirp.org

Huang, H.-M., Messina, E., Wade, R., English, R., Novak, B., & Albus, J. (2004, November 13–19). Autonomy measures for robots. In *Proceedings of the ASME International Mechanical Engineering Congress & Exposition (IMECE2004-61812)* (pp. 1241–1247). ASME. https://doi.org/10.1115/IMECE2004-61812 ASME Digital Collection

Oshana, M. (2020). Relational autonomy. In H. LaFollette (Ed.), *The international encyclopedia of ethics* (2nd ed.). Wiley. https://doi.org/10.1002/9781444367072.wbiee921 Wiley Online Library+1

Mackenzie, C. (2017). Feminist conceptions of autonomy. In A. Garry, S. J. Khader, & A. Stone (Eds.), *The Routledge companion to feminist philosophy* (pp. 515–527). Routledge. https://doi.org/10.4324/9781315758152-42 Taylor & Francis+1

Ding, J., Zhang, Y., Shang, Y., Zhang, Y., Zong, Z., Feng, J., Yuan, Y., Su, H., Li, N., Sukiennik, N., Xu, F., & Li, Y. (2025). Understanding world or predicting future? A comprehensive survey of world models. *ACM Computing Surveys*. https://doi.org/10.1145/3746449 (preprint: https://arxiv.org/abs/2411.14499) ACM Digital Library+1

Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., & Friston, K. (2020). Active inference on discrete state-spaces: A synthesis. *Journal of Mathematical Psychology, 99*, 102447. https://doi.org/10.1016/j.jmp.2020.102447 ScienceDirect

Tschantz, A., Millidge, B., Seth, A. K., & Buckley, C. L. (2020). Reinforcement learning through active inference. *arXiv*. https://arxiv.org/abs/2002.12636 arXiv

Friston, K. J., Daunizeau, J., & Kiebel, S. J. (2009). Reinforcement learning or active inference? *PLOS ONE, 4*(7), e6421. https://doi.org/10.1371/journal.pone.0006421 PLOS

Da Costa, L., Sajid, N., Parr, T., Friston, K., & Smith, R. (2023). Reward maximization through discrete active inference. *Neural Computation, 35*(5), 807–852. https://doi.org/10.1162/neco_a_01574 ResearchGate

Junker, F. T., Bruineberg, J., & Grünbaum, T. (2024). Predictive minds can be Humean minds. *The British Journal for the Philosophy of Science*. https://doi.org/10.1086/733413 journals.uchicago.edu

Limanowski, J., & Blankenburg, F. (2013). Minimal self-models and the free energy principle. *Frontiers in Human Neuroscience, 7*, 547. https://doi.org/10.3389/fnhum.2013.00547 Frontiers

Friston, K. J., & Frith, C. D. (2015). Active inference, communication and hermeneutics. *Cortex, 68*, 129–143. https://doi.org/10.1016/j.cortex.2015.03.025 PubMed

Vasil, J., Badcock, P. B., Constant, A., Friston, K., & Ramstead, M. J. D. (2020). A world unto itself: Human communication as active inference. *Frontiers in Psychology, 11*, 417. https://doi.org/10.3389/fpsyg.2020.00417 Frontiers

Veissière, S. P. L., Constant, A., Ramstead, M. J. D., Friston, K. J., & Kirmayer, L. J. (2019). Thinking through other minds: A variational approach to cognition and culture. *Behavioral and Brain Sciences, 43*, e90. https://doi.org/10.1017/S0140525X19001213 PubMed

Albarracin, M., Constant, A., Friston, K. J., & Ramstead, M. J. D. (2021). A variational approach to scripts. *Frontiers in Psychology, 12*, 585493. https://doi.org/10.3389/fpsyg.2021.585493