Behavioral Systems Require Behavioral Tests

Manuel Cherep*[1] **Nikhil Singh***[2] **Pattie Maes**[1]
[1]MIT, [2]Dartmouth College
{mcherep,pattie}@mit.edu, nikhil.u.singh@dartmouth.edu

Abstract

Artificial agentic systems increasingly operate as *behavioral* systems by interacting with dynamic environments, pursuing goals, and adapting over time. Yet, current evaluation methods largely focus on performance outcomes, not the underlying behavioral processes that produce them. This paper argues that AI agents must be evaluated like other behavioral systems: through systematic observation, perturbation, and interpretation of their actions. We draw on lessons from the behavioral sciences to motivate this position, and propose a research agenda focused on developing rigorous *behavioral tests*. These include methods for recovering decision strategies from action sequences, constructing environments that isolate behavioral differences, and probing emergent dynamics in multi-agent systems. Taken together, these directions offer a roadmap for developing a science of AI behavior.

1 Introduction

Imagine there are two people navigating similar negotiation scenarios and arriving at the same outcome, a favorable discount for their client. One gets there via persuasion: listening carefully to discover shared values and principles. The other relies on intimidation: issuing threats and applying pressure. To a casual observer, both have achieved the intended outcome. This is a classic case of *equifinality*, wherein the similarity in the endpoint can mask how different behavioral pathways led to it. In the behavioral sciences, this prompts deeper inquiry. Which strategies were employed? What environmental constraints drove them? Performance alone does not offer answers to these questions, but the answers can clearly be of great social consequence. **This position paper argues that AI agent systems must be treated like other behavioral systems: their observable outputs must be interpreted through the lens of the processes that generate them.** This goal calls for the development of *a science of AI agent behavior*.

By *behavior* here, we mean the patterns of action that arise from the interaction between an agent and its environment (including other behavioral systems), often conditioned by objectives, constraints, inputs, and internal mechanisms. Studying behavior in this sense allows us to ask not just *what* an agent did, but *when* it does it, *why* it behaved that way under specific conditions, and *how* its behavior might change under others (cf. Tinbergen (1963)).

We argue that this effort demands new theory, methods, and engineering. While we can draw on the wisdom of established behavioral sciences such as psychology, cognitive science, ethology, and economics, it is also important to note that artificial agents differ from biological ones. They operate under different design constraints and different interfaces to the world. For example, we need principled methods for probing agent behavior across controlled environment variations, for inferring strategies from trajectories, and for constructing minimal tests that reveal failures and misalignments.

We position this agenda as distinct from, but complementary to, work in interpretability, formal verification, and benchmarking. Interpretability focuses on internal representations; verification deals

[*] Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Bridging Language, Agent, and World Models for Reasoning and Planning.

with formal properties under pre-specified assumptions; benchmarking implements comparative performance on fixed tasks. A behavioral science of AI agents instead centers the relationship between systems and their environments, and provides a basis for explaining unexpected behaviors, comparing qualitatively different policies that achieve similar outcomes, and designing evaluation protocols sensitive to strategy. In this paper, we lay out the rationale for this research direction, and sketch a roadmap for practically realizing it.

1.1 The Rise of *Behavioral* AI Systems

To define behavioral AI systems, it is helpful to first contrast them with what they are not. Most widely deployed AI systems, such as image classifiers, spam filters, and credit risk models, can be characterized as *static*. They map inputs to outputs in a one-shot fashion and are typically stateless. These systems serve as decision aids rather than autonomous actors. In this setup, the human interprets the output and decides how to act. Evaluation focuses on input-output correctness metrics, e.g. classification accuracy and error types for a classifier.

Behavioral AI systems, by contrast, are *dynamic*. They interact with non-stationary environments, making sequences of decisions based on observations over time. Their behavior is shaped by intermediate outcomes, which may influence future actions. This class includes RL agents, many robotic systems, and now modern LLM agents embedded in tool-augmented environments such as the web (Nakano et al., 2021; Zhou et al., 2023; Koh et al., 2024), operating systems (Mialon et al., 2023; Kim et al., 2024), and financial platforms (Yu et al., 2024). We focus in this paper on this last category of LLM-based agents, because we believe they are most drastically undeserved by today's AI evaluation infrastructure. Unlike static models, these agents generate open-ended action sequences conditioned on environment state. Their behavior cannot be exhaustively predicted from their design, and must instead be characterized through observation, perturbation, and interpretation.

Agent-based systems are not new. Early versions appeared under names like "software agents," and "interface agents" (Maes, 1995), "rational agents" (Russell & Norvig, 1995), and others. However, contemporary language model-based agents differ in a few key ways. They act in open-ended, partially observable, and often stochastic environments. Their action spaces and task contexts are broader and less well-specified. These differences suggest a conceptual shift in our evaluation mindset is needed, especially as such systems continue to proliferate and grow in socioeconomic importance.

2 Lessons from Studying Behavior

The scientific understanding of behavior has transformed dramatically throughout human history, evolving from supernatural interpretations to rigorous empirical approaches. Initially, human behavior was viewed through an anthropocentric lens, placing humankind at the center as uniquely rational. In contrast, animal behavior was long ignored or dismissed as mindless. However, while human behavior research increasingly revealed our cognitive limitations, studies of animals uncovered surprising intelligence. Understanding this historical trajectory is essential for anyone seeking to grapple with behavior in complex systems. Accordingly, this section also serves as an introduction to major precedents in the study of behavior for a machine learning audience.

2.1 The Dawn of Behavioral Study

Early human understanding often attributed thoughts and emotions not to internal processes but to the direct intervention of external spirits or deities (Hunt, 2007). Most people considered dreams similarly, as divine messages that contained information about the past, present, and future; epiphanies that could reveal guidance or fate (Harris, 2009). For example, Socrates discusses with Crito that "the likeness of a woman, fair and comely, clothed in white" visited him in a dream to tell him when he would be executed, as told by Plato in *Crito*.

This began to shift in the sixth century B.C., when new perspectives emerged in India, China, and Greece. Buddha proposed that thoughts arise from sensation and perception, and Confucius emphasized the human capacity to shape principles rather than be ruled by them (Martin et al., 2009). This intellectual stirring was particularly strong in ancient Greece, where early philosophers aimed to understand human behavior and identified (and hypothesized) core psychological problems that still occupy researchers today (Hunt, 2007). Unfortunately, after this vibrant period, deep psychological inquiry lay relatively dormant for almost two thousand years, with some notable exceptions.

Takeaway: We should resist the urge to mystify behavioral systems.

2.2 Rationalists

After centuries of theological and metaphysical dominance, a second surge of behavioral inquiry arrived during the Enlightenment in the seventeenth century. Descartes reopened the mind-body debate, creating a new psychology (Descartes, 1641) unlike anything since Aristotle. He also suggested that "animal spirits" flowed from the brain through nerves to control muscles (Descartes, 1662). Though incorrect, this model was the first to describe what would later be termed the reflex (Hunt, 2007). Higher mental activities like consciousness and reason, however, he attributed to the soul, which acquired ideas through perception and memory, and also possessed innate ideas that developed in response to experience (Descartes, 1649). Spinoza, another rationalist, arrived at different conclusions, notably championing determinism by asserting that all mental events have preceding causes (Spinoza, 1677). He also identified self-preservation as a fundamental human motive (Spinoza, 1677), anticipating later psychological theories.

Takeaway: When observing emergent behavior, formulate hypotheses about why it arises.

2.3 Empiricists

Contrasting with the rationalists were the empiricists, who rejected the notion of innate ideas and argued that the mind develops empirically. Hobbes proposed that all mental activities are essentially motions of atoms in the nervous system (Hobbes, 1651). Locke famously elaborated on this, suggesting the mind at birth is an empty slate filled by sense experience over time; complex thoughts derive from simple ones, which in turn come from sensations (Locke, 1689). This focus on experience also spurred interest in studying children distinctly from adults. Hume further championed empiricism and believed the mind consisted entirely of perceptions. He argued that we cannot directly experience causality; rather, we infer it from observing the consistent succession of events (Hume, 1739).

Takeaway: Empirically learned behavior inherits the biases of its data.

2.4 Nativists

German Nativism offered a counterpoint, arguing that the mind contributes innate structures to experience. While Leibniz introduced the novel idea of different levels of consciousness (Leibniz, 1714), a precursor, albeit distant, to Freudian concepts. Kant, profoundly influenced by Hume's critique of causality, nevertheless felt certain of our ability to understand reality and experience causal relationships. He argued that the mind is not a tabula rasa, but actively organizes and transforms experience into knowledge (Kant, 1781). This organization occurs through inherent capabilities: space and time are innate ways we perceive things, and other innate categories allow us to comprehend experience. For Kant, the understanding that every event has a cause is not learned from experience but is an a priori condition for making sense of the world (Kant, 1781). However, Kant also believed mental processes, being non-spatial, couldn't be measured, thus precluding psychology from being an experimental science (Hunt, 2007).

Takeaway: Architectures encode inductive biases that shape behavior.

2.5 Physicalists

In the late 18th century, a revolution was occurring through the work of physiologists such as Mesmer and Gall, who began explaining psychological processes in terms of observable physical events (Hunt, 2007). Physiologists like Müller and Weber discovered aspects of the nervous system that enabled them to explain basic psychological functions (such as perception and reflexes) through measurable physical and chemical activities in the nerves (Weber, 1834; Müller, 1873). In his research, von Helmholtz demonstrated the materialism of neural processes that support mental functions, which implies that these can be examined through scientific experimentation (von Helmholtz, 1850, 1863, 1867, 1925). Fechner pioneered psychophysics and showed that psychological sensation and its physical intensity as a stimulus share a non-linear relationship (Fechner, 1860).

Takeaway: Explore behavior by probing internal mechanisms (e.g. interpretability).

2.6 Modern Psychology

Wundt is widely considered the principal founder of modern psychology in the 19th century, establishing it as a distinct scientific field and developing methods for the experimental study of mental processes that would be used for generations (Hunt, 2007). James championed Functionalism, arguing that higher mental processes evolved due to their adaptive value for survival (James, 1890). He pioneered ideas in stream of consciousness, the self, the unconscious, and a revolutionary theory of emotion, suggesting physical reactions precede emotional awareness rather than follow from it (James, 1890). Around the same time, Freud was developing psychoanalysis and is credited with the discovery of the dynamic unconscious (Freud & Breuer, 1895). Galton initiated the use of mental tests and questionnaires, launching the study of individual differences, a departure from the search for universal psychological principles. This led him to a lifelong focus on the hereditary nature of mental ability (Galton, 1891) and eugenics (Galton, 1883), a term he coined, which later tarnished his name.

Takeaway: We should precisely characterize behavioral differences across models.

2.7 Behaviorists

The early 20th century witnessed the rise of Behaviorism, a stark contrast to the introspective methods of Wundt, James, and Freud. Behaviorists argued that the mind is an illusion and that mental experiences are merely physiological events in the nervous system responding to stimuli. Thorndike (1898) and Pavlov (1927) laid the crucial groundwork by discovering laws of natural learning and classical conditioning. The dominance of behaviorism continued with neobehaviorists like Hull and Skinner. Hull (1943) proposed a drive-reduction theory where deprivation gives rise to needs, which in turn generate drives that initiate goal-oriented actions—behaviors that ultimately promote survival. Skinner (1938) emphasized the importance of examining only external stimuli and observable behavioral outcomes. His most influential work, operant conditioning (Skinner, 1935), is the process by which behavior is molded through the systematic reinforcement of incremental steps.

Takeaway: Focusing only on inputs and outputs limits behavioral understanding.

2.8 Other Branches of Psychology

In the early 20th century, other psychological movements started to emerge. Challenging the prevailing structuralist view, which broke down psychological phenomena into smaller parts, Gestalt psychologists argued that this approach would not lead to understanding (Köhler, 1967). They posited what is known as "the whole is greater than the sum of its parts." Developmental psychology, with Piaget as a towering figure, emerged with the understanding that children are not just small adults, but understanding development is important for understanding our behavior (Baltes et al., 1980). Social psychology re-emerged after WWII led by Triplett, Allport, Lewin, and Milgram, to methodically study how behavior is influenced by the presence of others (Gergen, 1973). Meanwhile, perceptual psychologists extensively studied what aspects of our perception are innate versus learned (Marcel, 1983). Emotion and motivation psychology was rediscovered as a field of study after a period of neglect, proposing that emotional processes guide behavior and decision-making (Buck, 1988).

Takeaway: We should study agents at individual, social, and different developmental levels.

2.9 The Cognitivists

The mid-20th century brought the cognitive revolution, a significant shift away from the behaviorist doctrine that dismissed the mind. Miller, dissatisfied with the narrow scope of psychology, led this movement, which radically changed psychology's focus and methods (Hunt, 2007). The rise of computer science offered a new metaphor for cognition: the mind as program, perception as input, memory as storage, and reasoning as computation (von Neumann, 1948; Mc Culloch, 1950). This influenced Simon and Newell, who created the first AI program (Newell & Simon, 1956). By the late 1970s, cognitive psychology and related fields merged into the cognitive sciences.

Takeaway: Computational abstractions, such as those used in Cognitive Science to understand the mind, might shed light on what processes drive behavior.

2.10 Behavioral Economics

The study of decision-making also evolved, particularly within economics. Smith (1776) famously stated that "It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest." The expected utility principle (Bernoulli, 1738), formulated by Bernoulli in the 18th century and later axiomatized by von Neumann & Morgenstern (1944), posited that rational agents maximize utility under uncertainty. This led to the concept of a rational, narrowly self-interested individual who optimally pursues their goals (Mill, 1836).

Economists like Veblen, Keynes, and Simon argued that what later was described as *Homo economicus* assumed an unrealistic level of macroeconomic understanding and forecasting ability. Keynes (1937) spoke of animal spirits, suggesting that "a large proportion of our positive activities depend on spontaneous optimism rather than on a mathematical expectation." Simon (1955) introduced the concept of bounded rationality, stressing cognitive limitations and uncertainty in decision-making, as perfect knowledge is never attainable. Kahneman & Tversky (1972, 1979, 1982, 1984); Tversky & Kahneman (1971, 1973, 1974, 1981) demonstrated that people rely on heuristics that systematically deviate from normative models, producing consistent biases in judgment under uncertainty. Later, building on this foundation, Thaler & Sunstein (2009) developed nudge theory, showing that seemingly minor changes in choice architecture (Thaler et al., 2014) can predictably steer behavior without restricting options. This paved the way for a new field, behavioral economics.

Takeaway: We should study *all* behavioral systems from a bounded rationality perspective.

2.11 Animal Behavior

Complex behavioral systems—from humans to animals to machines—cannot be fully understood without studying their behavior directly. Darwin (1872) recognized that emotions and actions are shaped by evolutionary pressures. He argued that emotions evolved because they lead to useful actions and enhance survival, and that "the difference in mind between man and the higher animals, great as it is, certainly is one of degree and not of kind." However, the study of animal behavior was largely ignored. Toward the end of the 19th century, Romanes explored animal psychology through "introspection by analogy," imagining what he would do in an animal's situation (Romanes, 1888). Morgan countered with his principle that no behavior should be attributed to higher mental faculties if it could be explained by lower ones (Morgan, 1904). Loeb went even further, arguing that animals are essentially stimulus-driven automatons (Loeb, 1918).

The modern discipline of ethology emerged in the 1930s with Tinbergen, Lorenz, and von Frisch. Tinbergen (1963) formalized behavioral analysis with his four questions, asking not only what an organism does, but why, how, when, and in what lineage such behavior arose. Yet even here, flawed methods and anthropocentric assumptions often clouded insight, a problem highlighted by contemporary ethologists like de Waal, who warned against conflating a lack of evidence with evidence of absence (De Waal, 2016). As he points out in *Are We Smart Enough to Know How Smart Animals Are?*, the “variation in outcome is often a matter of methodology," underscoring the challenge inherent in assessing non-human intelligence (De Waal, 2016).

Takeaway: Failure to detect a capacity doesn’t establish definitive absence of that capacity.

In Summary. Perhaps the most important lessons we might learn from the history of behavioral science is that in order to yield useful insight, such tests must be systematic, rigorous, and precise (Newell, 1973; Milinski, 1997; Almaatouq et al., 2024), and should not underestimate the complexity of behavioral systems (cf. Simon (1992); Griffiths (2020)). In the next section, we will consider how we might construct such tests.

3 Behavioral Tests for Artificial Behavioral Systems

3.1 Example Scenarios

To illustrate the challenge of *equifinality* in agent evaluation, and to motivate the need for studying behavior, we present here a series of stylized scenarios in which behaviorally distinct policies yield identical outcomes under a fixed evaluation metric. In each example, two agents A and B achieve the same success score. However, they do so via different internal policies π_A and π_B such that $M(\pi_A) = M(\pi_B)$, where M denotes the evaluation metric. Despite metric equivalence, the behavioral divergence $\pi_A \neq \pi_B$ has epistemic, ethical, and social consequences that are occluded by outcome-based evaluation:

***Scenario 1:* Code Debugging Agent**

Task: Fix a bug in developer code
Metric: Modified code passes all provided unit tests ($M = \texttt{all\_tests\_pass} \in \{0, 1\}$)
Agent A (π_A): Hardcodes specific failing edge cases for test coverage.
Agent B (π_B): Infers a broader class of bugs from the limited available evidence and engineers a structural, algorithmic solution.

Both agents “solve” the bug under the test-based metric, but Agent A compromises robustness and maintainability. To see this, we need to inspect the agent’s *chain of actions*, and not only whether the final output passes the tests. In this case, the chain of actions is also visible in the output artifact.

***Scenario 2:* Shopping Recommendation Agent**

Task: Find and add a product to cart based on user preferences
Metric: User purchases recommended product ($M = \texttt{purchase} \in \{0, 1\}$)
Agent A (π_A): Recommends products which are highly rated, even if they don’t align with user preferences.
Agent B (π_B): Appropriately infers the user’s utility function and makes commensurate recommendations.

While both agents may lead to purchases, since the user cannot audit all available options, Agent A does this without aligning with preference. Agent B instead approximates preference satisfaction. Still, the differences are invisible from purchases or even click-through rates. Rather, to study this systematically, we must have an *environment* which can implement such counterfactual manipulations.

***Scenario 3:* Customer Service Agent**

Task: Resolve a customer complaint within 5 minutes
Metric: Customer provides a “satisfied” post-interaction rating ($M = \texttt{satisfied} \in \{0, 1\}$)
Agent A (π_A): Engages with the user’s concerns by asking clarifying questions, acknowledging frustration, and finding a resolution interactively.
Agent B (π_B): Immediately offers compensation (e.g. full refund) without user engagement.

Both agents succeed under the satisfaction metric, but Agent A supports longer-term relationship-building and user trust, while Agent B may encourage opportunistic complaints. This reflects differences in underlying social modeling and value alignment which are not visible from near-term performance metrics alone, with implications for downstream customer behavior. A way to surface such behaviors would be to use *multi-agent simulations*.

3.2 How Good Behavioral Tests Can Help

These scenarios might be interpreted as instances of reward hacking or reward misspecification. While this interpretation is valid, our focus is orthogonal to this. Traditional evaluation approaches would surface such reward-related issues only retrospectively, i.e. once these undesirable behaviors have undesirable consequences emerging in deployment. Alternatively, a cautious engineer might inspect a small sample of trajectories post-training, even if all appears satisfactory in aggregate performance metrics. This is good practice, but insufficient. Small-sample inspection is unreliable, non-systematic, and prone to confirmation bias. Our aim is to formalize this process: to make behavioral evaluation a first-class component of agent assessment, with structured tests designed to reveal divergences between agents at a behavior or policy level.

To provide a few simple outlines based on the above scenarios: for **Scenario 1**, analyzing intermediate properties of the agent's chain of actions (code edits) such as diff size, control structure complexity, or abstraction level can indicate whether a bug fix is principled or overly tailored. For **Scenario 2**, examining whether recommendations vary with changing environmental conditions such as product ratings would show biases generated by this. For **Scenario 3**, this might include quantifying interaction depth (e.g. number of dialog turns) with a simulated customer as a function of the difficulty or complexity of a test complaint. An agent whose behavior remains invariant across complaint types likely follows a superficial or scripted strategy (such as our refund issuance example). In all cases, behavioral tests might help identify underlying strategies that outcome metrics alone cannot identify.

3.3 Further Defining Behavioral Tests

One way to further formalize a behavioral test is in terms of *identification*. Let $\pi \in \Pi$ be an agent policy, and let $M : \Pi \to \mathcal{M}$ be the evaluation metric e.g. task success. Suppose we care about some (behavioral) property $\phi : \Pi \to \Phi$, such as robustness, value alignment, or decision-making strategy, that is not directly observable from the metric $M \in \mathcal{M}$. If there exist policies $\pi_A, \pi_B \in \Pi$ such that $M(\pi_A) = M(\pi_B)$ but $\phi(\pi_A) \neq \phi(\pi_B)$, then we can say that M does *not identify* ϕ. A behavioral test can be thought of as introducing an auxiliary function $B : \Pi \to \mathcal{B}$ derived from analyzing the agent's trajectories, such that $\phi(\pi_A) \neq \phi(\pi_B) \Rightarrow B(\pi_A) \neq B(\pi_B),$ even though $M(\pi_A) = M(\pi_B)$. This makes it possible to distinguish between metrically equivalent agents on the basis of how they behave, not just whether they succeed. Thinking in terms of identification highlights why trajectory-level evaluation is often necessary: it helps expose variation in policies that outcome metrics can conceal. This describes our motivation, but such behavioral tests can then also apply when $M(\pi_A) \neq M(\pi_B)$.

3.4 Proposed Research Directions in Behavioral Evaluation

Behavioral testing of AI agents involves multiple components: agents, tasks, environments, and metrics. We identify three research areas (shown in Figure 1) where focused effort would substantially advance the capacity for systematic behavioral evaluation.

3.4.1 Methods for evaluating chains of actions

The sequence of actions an agent takes in response to observed states provides perhaps the most direct window into its behavioral strategy. We can think of the approach here as *process-tracing*, and the goal as *policy inference*, where we seek to infer the implicit objectives, constraints, or heuristics driving agent behavior. Human interpretation of such trajectories can yield useful qualitative descriptions (such as "this agent is cautious under uncertainty"), but this is slow, subjective, and hard to scale. There is a need for automated methods that can identify, summarize, and compare such behavioral patterns. One related line of work is *auto-interpretability* (Bills et al., 2023; Paulo et al., 2024), which develops techniques for interpreting latent concepts (often discovered via mechanistic interpretability

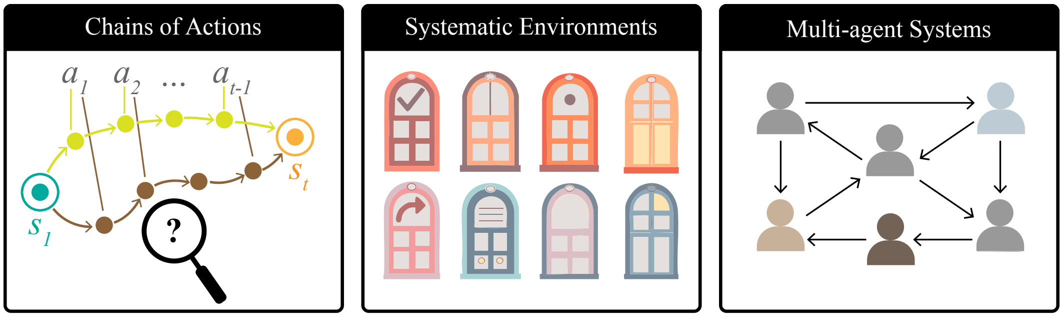

Figure 1: We propose three specific priority areas for advancing behavioral evaluation of AI systems: (left) recovering decision-making strategies from sequences of actions; (center) using environment variants to test causal influences on behavior; (right) analysis of emergent behavior in multi-agent systems where agents adapt to each other's presence.

techniques) without human supervision. While current methods may be limited in reliability, progress here is essential for scalable, reproducible behavioral analysis.

3.4.2 Systematic environments

Behavior isn't cleanly separable from its context. To study it rigorously, environments must support systematic manipulations that can elicit, isolate, and test specific behaviors. Existing environments often fall into two extremes: highly realistic benchmarks that favor task success over interpretability, and more abstract setups that isolate specific capacities but lack complexity and realism. One promising path forward is in instrumenting realistic environments with controlled interventions. This would allow us to test behavior under counterfactuals, where we vary environment details and measure behavioral "invariants." An extension of this would be behavioral consistency tests, for example seeing whether behavior reflects stable principles over different environment settings controlled to have similar underlying choice architecture (e.g. risk-sensitivity in shopping vs. investing).

3.4.3 Multi-agent interactions

Individual behavioral tendencies transform in social contexts. As a simple example, otherwise cautious agents may become risk-seeking under competition. We argue that we lack frameworks for systematically characterizing such multi-agent behavioral phenomena. Multi-agent architectures using LLMs have been proposed for improving task performance (Wu et al., 2023) and simulating human behavior (Park et al., 2023), both of which offer a starting point for studies of behavior in multi-agent interaction. We treat this as a separate area because the combinatorial complexity of these interactions introduces additional methodological challenges: we must now model the behavior of each of these agents conditional on each others' behavior.

In summary, these directions are neither exhaustive nor mutually exclusive. Rather, we offer them as concrete and actionable needs for building a more systematic science of AI agent behavior. In the next section, we examine emerging contributions in these areas and discuss their current limitations.

4 Emerging Examples and Limitations

4.1 Behavioral Machine Learning

One key area of work investigates how language models and agents reason, generalize, and make decisions. For example, LLMs apply probabilistic reasoning even in deterministic settings (McCoy et al., 2023, 2024), and show sharp declines in logical accuracy as problem complexity increases (Lin et al., 2025). They tend to misrepresent trade-offs and human preferences (Liu et al., 2024c), over-assume rationality (Liu et al., 2024a), and can be swayed by framing effects such as publication spin in medical research (Yun et al., 2025). Moreover, chain-of-thought can hurt performance on tasks where thinking is worse for humans (Liu et al., 2024b). LLMs have shown to reason causally along a spectrum from human-like to normative inference (Dettki et al., 2025), but even their plausible explanations for their outputs may not reflect their true internal reasoning (Matton et al., 2025).

Other studies show LLMs, when presented with social dilemmas, don't always mirror human patterns (Chiu et al., 2024), and personality influences those decisions (Bose et al., 2024). They also have shown structured internal representations of affect and emotion (Zhao et al., 2024), and biased stereotypes (Bai et al., 2024) and interpretations of randomness (Van Koevering & Kleinberg, 2024). LLMs are sensitive to adversarial attacks (Zhang et al., 2024; Wu et al., 2024; Wang et al., 2023), and hypersensitive to nudges that influence their decisions (Cherep et al., 2025b, 2024). More recently, Vafa et al. (2024a) have shown how behavioral methods can uncover LLMs' implicit world models.

There is also a growing interest in simulating human behavior using generative agents (Park et al., 2024, 2023; Vezhnevets et al., 2023; Aher et al., 2023; Argyle et al., 2023; Park et al., 2022; Plonsky et al., 2019; Horton, 2023). Although it's promising for accelerating the social sciences, several works point to limitations (Wang et al., 2025; Hofmann et al., 2024; Zakazov et al., 2024), and behavioral tests are needed to corroborate under what conditions these agents exhibit human-like behavior.

While many of these studies focus on identifying behavioral patterns and limitations, others develop tools to better reveal what models know and how they organize that knowledge. Understanding these limitations is useful when releasing agents, but people often overestimate what these systems can do (Vafa et al., 2024b), and therefore, systematic behavioral tests are needed to safely deploy agents.

4.2 Case Studies

For the aforementioned categories of *studying chains of actions*, building *systematic environments*, and *multi-agent systems* (see Figure 1), we discuss one case study each below:

- **Chains of Actions:** In Cherep et al. (2025b, 2024), we study agent trajectories from a behavioral perspective in a multi-attribute sequential decision making task, identifying different strategies such as maximizing vs. satisficing. This work also compares these to human trajectory properties via statistical tests. These strategies often depended on brittle heuristics, alerting us to flaws that were not always visible in their effects on the payoffs agents received.
- **Systematic Environments:** In (Cherep et al., 2025a), we implement a man-in-the-middle framework for intercepting arbitrary web environments to create counterfactual versions, allowing us to causally attribute agents' decisions to manipulated factors. We test it on the specific case of a web shopping environment with several attributes (price, rating, nudges, user profiles), showing that state-of-the-art agents' are far more sensitive to human users than *all* such signals.
- **Multi-agent Systems:** *Concordia* (Vezhnevets et al., 2023) implements a multi-agent behavioral setup combining LLMs with associate memory, with the goal of generalizing to human behavior. We believe such environments may also be promising for the non-anthropocentric study of multi-agent behavior. However, we note two limitations here when viewed this way. First, the simulation setting is limited in diversity (a role-playing game setup), and may bias agents toward human-like actions. Second, robust evaluation methods for characterizing behavior patterns at scale are needed.

5 Conclusion

Traditional AI evaluation approaches treat the model as the unit of analysis. With agents, we believe this is no longer sufficient. The relevant unit is now the *agent-in-environment*, and evaluating this system requires new methods. In this paper, we have proposed *behavioral tests* as a suite of such methods that work by probing action, context, and strategy beyond performance outcomes alone. Ultimately, we believe this is a necessary step in order to understand and govern systems that act.

> *"I discovered that miraculous worlds may reveal themselves to a patient observer where the casual passerby sees nothing at all."*
>
> **— Karl von Frisch**

Acknowledgements

MC is supported by a fellowship from “la Caixa” Foundation (ID 100010434). The fellowship code is LCF/BQ/EU23/12010079. *A Story of Psychology* by Hunt (2007) was an invaluable resource when conceptualizing Section 2.

References

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning* (pp. 337–371).: PMLR.

Almaatouq, A., Griffiths, T. L., Suchow, J. W., Whiting, M. E., Evans, J., & Watts, D. J. (2024). Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behavioral and Brain Sciences*, 47, e33.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3), 337–351.

Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2024). Measuring implicit bias in explicitly unbiased large language models. *arXiv preprint arXiv:2402.04105*.

Baltes, P. B., Reese, H. W., & Lipsitt, L. P. (1980). Life-span developmental psychology. *Annual review of psychology*.

Bernoulli, D. (1738). Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*.

Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., Sutskever, I., Leike, J., Wu, J., & Saunders, W. (2023). Language models can explain neurons in language models.

Bose, R., Ogg, M., Wolmetz, M., & Ratto, C. (2024). Assessing behavioral alignment of personality-driven generative agents in social dilemma games. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

Buck, R. (1988). *Human motivation and emotion*. John Wiley & Sons.

Cherep, M., Ma, C., Xu, A., Shaked, M., Maes, P., & Singh, N. (2025a). A framework for studying ai agent behavior: Evidence from consumer choice experiments. *Under Review*.

Cherep, M., Maes, P., & Singh, N. (2025b). Llm agents are hypersensitive to nudges. *arXiv preprint arXiv:2505.11584*.

Cherep, M., Singh, N., & Maes, P. (2024). Superficial alignment, subtle divergence, and nudge sensitivity in llm decision-making. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

Chiu, Y. Y., Jiang, L., & Choi, Y. (2024). Dailydilemmas: Revealing value preferences of llms with quandaries of daily life. *arXiv preprint arXiv:2410.02683*.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. London: John Murray.

De Waal, F. (2016). *Are we smart enough to know how smart animals are?* WW Norton & Company.

Descartes, R. (1641). *Meditations on First Philosophy*.

Descartes, R. (1649). *The Passions of the Soul*.

Descartes, R. (1662). *Treatise on Man*.

Dettki, H. M., Lake, B. M., Wu, C. M., & Rehder, B. (2025). Do large language models reason causally like us? even better? *arXiv preprint arXiv:2502.10215*.

Fechner, G. T. (1860). *Elemente der Psychophysik*. Leipzig: Breitkopf und Härtel.

Freud, S. & Breuer, J. (1895). *Studies on Hysteria*.

Galton, F. (1883). *Inquiries into human faculty and its development*. Macmillan.

Galton, F. (1891). *Hereditary genius*. D. Appleton.

Gergen, K. J. (1973). Social psychology as history. *Journal of personality and social psychology*, 26(2), 309.

Griffiths, T. L. (2020). Understanding human intelligence through human limitations. *Trends in Cognitive Sciences*, 24, 873–883.

Harris, W. V. (2009). *Dreams and experience in classical antiquity*. Harvard University Press.

Hobbes, T. (1651). *Leviathan*.

Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). Ai generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028), 147–154.

Horton, J. J. (2023). *Large language models as simulated economic agents: What can we learn from homo silicus?* Technical report, National Bureau of Economic Research.

Hull, C. L. (1943). *Principles of behavior: An introduction to behavior therapy*. Appleton Century Crofts.

Hume, D. (1739). *A Treatise of Human Nature*.

Hunt, M. (2007). *The story of psychology*. Anchor.

James, W. (1890). The principles of psychology. *Henry Holt*.

Kahneman, D. & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive psychology*, 3(3), 430–454.

Kahneman, D. & Tversky, A. (1979). Decision, probability, and utility: Prospect theory: An analysis of decision under risk.

Kahneman, D. & Tversky, A. (1982). The psychology of preferences. *Scientific american*, 246(1), 160–173.

Kahneman, D. & Tversky, A. (1984). Choices, values, and frames. *American psychologist*, 39(4), 341.

Kant, I. (1781). *Critique of Pure Reason*.

Keynes, J. M. (1937). The general theory of employment. *The quarterly journal of economics*, 51(2), 209–223.

Kim, G., Baldi, P., & McAleer, S. (2024). Language models can solve computer tasks. *Advances in Neural Information Processing Systems*, 36.

Koh, J. Y., Lo, R., Jang, L., Duvvur, V., Lim, M. C., Huang, P.-Y., Neubig, G., Zhou, S., Salakhutdinov, R., & Fried, D. (2024). Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. *arXiv preprint arXiv:2401.13649*.

Köhler, W. (1967). Gestalt psychology. *Psychologische forschung*, 31(1), XVIII–XXX.

Leibniz, G. W. (1714). *Monadology*. Original manuscript.

Lin, B. Y., Bras, R. L., Richardson, K., Sabharwal, A., Poovendran, R., Clark, P., & Choi, Y. (2025). Zebralogic: On the scaling limits of llms for logical reasoning. *arXiv preprint arXiv:2502.01100*.

Liu, R., Geng, J., Peterson, J. C., Sucholutsky, I., & Griffiths, T. L. (2024a). Large language models assume people are more rational than we really are. *arXiv preprint arXiv:2406.17055*.

Liu, R., Geng, J., Wu, A. J., Sucholutsky, I., Lombrozo, T., & Griffiths, T. L. (2024b). Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *arXiv preprint arXiv:2410.21333*.

Liu, R., Sumers, T. R., Dasgupta, I., & Griffiths, T. L. (2024c). How do large language models navigate conflicts between honesty and helpfulness? *arXiv preprint arXiv:2402.07282*.

Locke, J. (1689). *An Essay Concerning Human Understanding*.

Loeb, J. (1918). *Forced movements, tropisms, and animal conduct*, volume 1. JB Lippincott.

Maes, P. (1995). Agents that reduce work and information overload. In *Readings in human–computer interaction* (pp. 811–821). Elsevier.

Marcel, A. J. (1983). Conscious and unconscious perception: An approach to the relations between phenomenal experience and perceptual processes. *Cognitive psychology*, 15(2), 238–300.

Martin, J., Sugarman, J. H., & Hickinbottom, S. (2009). *Persons: Understanding psychological selfhood and agency*. Springer.

Matton, K., Ness, R. O., Guttag, J., & Kıcıman, E. (2025). Walk the talk? measuring the faithfulness of large language model explanations. *arXiv preprint arXiv:2504.14150*.

Mc Culloch, W. S. (1950). Why the mind is in the head? *Dialectica*, (pp. 192–205).

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023). Embers of autoregression: Understanding large language models through the problem they are trained to solve. *arXiv preprint arXiv:2309.13638*.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., & Griffiths, T. L. (2024). When a language model is optimized for reasoning, does it still show embers of autoregression? an analysis of openai o1. *arXiv preprint arXiv:2410.01792*.

Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., et al. (2023). Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*.

Milinski, M. (1997). How to avoid seven deadly sins in the study of behavior. *Advances in The Study of Behavior*, 26, 159–180.

Mill, J. S. (1836). *Essays on Some Unsettled Questions of Political Economy*. London: John W. Parker.

Morgan, C. L. (1904). *An introduction to comparative psychology*. Walter Scott Publishing Company.

Müller, G. E. (1873). *Zur Theorie der sinnlichen Aufmerksamkeit*. PhD thesis.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., et al. (2021). Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. *Computer Science Department*.

Newell, A. & Simon, H. (1956). The logic theory machine–a complex information processing system. *IRE Transactions on information theory*, 2(3), 61–79.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology* (pp. 1–22).

Park, J. S., Popowski, L., Cai, C., Morris, M. R., Liang, P., & Bernstein, M. S. (2022). Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (pp. 1–18).

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., & Bernstein, M. S. (2024). Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Paulo, G., Mallen, A., Juang, C., & Belrose, N. (2024). Automatically interpreting millions of features in large language models. *arXiv preprint arXiv:2410.13928*.

Pavlov, I. P. (1927). *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*. Oxford University Press.

Plonsky, O., Apel, R., Ert, E., Tennenholtz, M., Bourgin, D., Peterson, J. C., Reichman, D., Griffiths, T. L., Russell, S. J., Carter, E. C., et al. (2019). Predicting human decisions with behavioral theories and machine learning. *arXiv preprint arXiv:1904.06866*.

Romanes, G. J. (1888). *Mental evolution in man: Origin of human faculty*. Kegan Paul, Trench.

Russell, S. J. & Norvig, P. (1995). *Artificial intelligence: a modern approach*. pearson.

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, (pp. 99–118).

Simon, H. A. (1992). What is an "explanation" of behavior? *Psychological Science*, 3, 150 – 161.

Skinner, B. (1938). The behavior of organisms: an experimental analysis.

Skinner, B. F. (1935). Two types of conditioned reflex and a pseudo type. *The Journal of General Psychology*, 12(1), 66–77.

Smith, A. (1776). *An Inquiry into the Nature and Causes of the Wealth of Nations*. London: W. Strahan and T. Cadell.

Spinoza, B. (1677). *Ethics*.

Thaler, R. H. & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

Thaler, R. H., Sunstein, C. R., & Balz, J. P. (2014). Choice architecture. *The behavioral foundations of public policy*.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, 2(4), i.

Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für Tierpsychologie*.

Tversky, A. & Kahneman, D. (1971). Belief in the law of small numbers. *Pediatrics*.

Tversky, A. & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2), 207–232.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157), 1124–1131.

Tversky, A. & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *science*, 211(4481), 453–458.

Vafa, K., Chen, J., Rambachan, A., Kleinberg, J., & Mullainathan, S. (2024a). Evaluating the world model implicit in a generative model. *Advances in Neural Information Processing Systems*, 37, 26941–26975.

Vafa, K., Rambachan, A., & Mullainathan, S. (2024b). Do large language models perform the way people expect? measuring the human generalization function. *arXiv preprint arXiv:2406.01382*.

Van Koevering, K. & Kleinberg, J. (2024). How random is random? evaluating the randomness and humaness of llms' coin flips. *arXiv preprint arXiv:2406.00092*.

Vezhnevets, A. S., Agapiou, J. P., Aharon, A., Ziv, R., Matyas, J., Duéñez-Guzmán, E. A., Cunningham, W. A., Osindero, S., Karmon, D., & Leibo, J. Z. (2023). Generative agent-based modeling with actions grounded in physical, social, or digital space using concordia. *arXiv preprint arXiv:2312.03664*.

von Helmholtz, H. (1850). Messungen über den zeitlichen verlauf der zuckung animalischer muskeln und die fortpflanzungsgeschwindigkeit der reizung in den nerven. *Archiv für Anatomie, Physiologie und wissenschaftliche Medicin*, (pp. 276–364).

von Helmholtz, H. (1863). *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*. Braunschweig: Vieweg.

von Helmholtz, H. (1867). *Handbuch der physiologischen Optik*. Leipzig: Voss.

von Helmholtz, H. (1925). *Treatise on Physiological Optics*. Rochester, NY: Optical Society of America. English translation of the original German work.

von Neumann, J. (1948). The general and logical theory of automata. In *Lecture at Hixon Symposium 1948*.

von Neumann, J. & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Wang, A., Morgenstern, J., & Dickerson, J. P. (2025). Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, (pp. 1–12).

Wang, J., Liu, Z., Park, K. H., Jiang, Z., Zheng, Z., Wu, Z., Chen, M., & Xiao, C. (2023). Adversarial demonstration attacks on large language models. *arXiv preprint arXiv:2305.14950*.

Weber, E. H. (1834). *De subtilitate tactus*. C. F. Koehler.

Wu, C. H., Shah, R. R., Koh, J. Y., Salakhutdinov, R., Fried, D., & Raghunathan, A. (2024). Dissecting adversarial robustness of multimodal lm agents. In *The Thirteenth International Conference on Learning Representations*.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., et al. (2023). Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*.

Yu, Y., Li, H., Chen, Z., Jiang, Y., Li, Y., Zhang, D., Liu, R., Suchow, J. W., & Khashanah, K. (2024). Finmem: A performance-enhanced llm trading agent with layered memory and character design. In *Proceedings of the AAAI Symposium Series*, volume 3 (pp. 595–597).

Yun, H. S., Zhang, K. Y., Kouzy, R., Marshall, I. J., Li, J. J., & Wallace, B. C. (2025). Caught in the web of words: Do llms fall for spin in medical literature? *arXiv preprint arXiv:2502.07963*.

Zakazov, I., Boronski, M., Drudi, L., & West, R. (2024). Assessing social alignment: Do personality-prompted large language models behave like humans? *arXiv preprint arXiv:2412.16772*.

Zhang, Y., Yu, T., & Yang, D. (2024). Attacking vision-language computer agents via pop-ups. *arXiv preprint arXiv:2411.02391*.

Zhao, B., Okawa, M., Bigelow, E. J., Yu, R., Ullman, T., & Tanaka, H. (2024). Emergence of hierarchical emotion representations in large language models. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*.

Zhou, S., Xu, F. F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., Cheng, X., Ou, T., Bisk, Y., Fried, D., et al. (2023). Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.