Behavioral AI: Building Algorithms That Understand Us

Anonymous authors
Paper under double-blind review

Abstract

While AI research has historically worked toward developing tools to help people understand AI models, the emergence of generative AI into our daily lives suddenly makes the reverse question salient: how well can AI models understand people? Today's AI systems fall short; these deficiencies demand a focus toward building new systems that can understand people. In this Perspective, we endeavor to channel this focus into a new academic subfield, which we call Behavioral AI. This Perspective lays out dimensions of understanding that are currently deficient in AI systems, including emotional, intellectual, and preferential understanding. While improving AI systems along these dimensions faces a unique set of challenges, we show there has already been a flurry of progress across disciplines. As we build systems that better understand people, it will not only improve AI tools; if Behavioral AI succeeds as a field, these systems too can unlock new insights in the behavioral sciences that help us understand ourselves.

1 Introduction

The field of artificial intelligence (AI) has a long history of researchers working to understand and interpret the models they create (Garson, 1991; Zeiler & Fergus, 2014; Rahwan et al., 2019). This work has been fundamental for recognizing the capabilities and limitations of AI models, dating back more than 50 years to the study of Rosenblatt's perceptron (Rosenblatt, 1958; Minsky & Papert, 1969). However, the recent advancement of generative AI systems like large language models (LLMs) has introduced new capabilities — back-and-forth communication through plain language interfaces, even "reasoning" — that offer news ways to integrate AI into our daily lives. In contrast to the single-purpose models that have dominated the history of AI, these systems come with the promise of more general uses of AI, e.g. as tutors, writing assistants, and everyday agents — even "thought partners" (Collins et al., 2024c). While researchers have typically asked how well humans can understand AI systems, the emergence of these new technologies suddenly makes the reverse question salient: *how well can AI systems understand humans?*

Consider, for example, a student learning trigonometry yet who makes errors whenever a problem requires calculating the cosine of an obtuse angle. An effective human teacher should understand the source of their errors, identify the kinds of problems where it occurs, and construct lessons for the student. An effective AI tutor should do the same. Or consider someone asking a friend for advice before making an important life decision such as buying a car. A helpful friend would tailor advice based on their understanding of the person: how they've approached decisions in the past, which ones they tend to regret, and whether stated worries reflect genuine concerns or unfounded anxieties. AI assistants should do the same. Even a task as mundane as dispatching an AI agent to schedule a meeting requires navigating implicit preferences: should the meeting not be scheduled right after other meetings that are likely to run over? And what are those meetings? None of these examples require that AI systems understand in the same way as humans; only that they behave as if they understand our thoughts, emotions, and goals.

Unfortunately, AI systems do not currently exhibit these capabilities, struggling to carry out tasks that require even simple understanding. This deficiency demands a focus into building new systems that can understand people. The goal of this Perspective is to channel this focus into a new academic subfield, which we call Behavioral AI. There are already efforts under way across multiple disciplines to build algorithms that are more capable of understanding people, though often in isolation. For example, in robotics, practitioners are developing systems capable of inferring human intentions (Dragan et al., 2013; Zhang et al., 2023); in human-computer interaction, researchers are designing interfaces that adapt to user behavior (Buçinca

et al., 2025; Gajos et al., 2010; Todi et al., 2021); in natural language processing, large language models are being trained to align with human values and preferences (Christiano et al., 2017; Ouyang et al., 2022). Moreover, many of these problems require expertise from the social, cognitive, and behavioral sciences, which offer theories and methods for modeling and understanding humans, as in computational cognitive science (Tenenbaum et al., 2011; Griffiths et al., 2024). The goal of Behavioral AI as a field is to coalesce and solidify this burgeoning but disparate group of researchers who can contribute to these goals.

In this Perspective, we will describe several different kinds of understanding that are currently deficient in AI systems. While improving AI systems along these dimensions faces a unique set of challenges, there has already been a flurry of progress across disciplines. We conclude by discussing the benefits we can reap from models that understand us: not only will we be equipped with more reliable AI systems, but we will also have an opportunity to attain a deeper understanding of human behavior.

2 AI Systems Have an Incomplete Understanding of People

How well do today's AI systems understand people? At one level, they're able to perform difficult tasks that require intimate familiarity with the world, like predicting the next word of our conversations or diagnosing a patient's cancer risk. However, these tasks often require only a surface-level understanding of us: what's missing is the need to infer our goals, thoughts, and emotions. Below we provide examples that illustrate some dimensions of understanding that are incomplete in current AI systems.

Preferential understanding. One of the most visible capabilities of AI systems is their ability to write fluent text. However, when it comes to supporting writers in their own endeavors, these capabilities are more limited. This is partially due to their deficiencies in understanding a writer's preferences and intents. Early studies with novelists have found that language models struggle to follow the direction of a writer's story (Calderwood et al., 2020). More recent work continues to find, e.g. with playwrights, that AI language systems not only struggle with long-form coherence, but fail to understand the nuance and subtext that writers set up (Mirowski et al., 2023), and, with professional writers, that these systems consistently and across the board fail to adhere to writers' stylistic preferences (Chakrabarty et al., 2025). If algorithms truly grasped writers' preferences and goals, they would be able to provide drafts, edits, and suggestions that matched these intentions (Lee et al., 2024).

Many uses of agentic AI require that they understand our preferences and goals. Even a task as seemingly simple as booking a hotel for travel requires an understanding of our intentions: would we like a quiet hotel away from city streets, or would we like to stay near the center of city life? Is easy access to public transit more important to us than having on-site parking? While it might be possible in principle to describe preferences to an agent, explicitly articulating every preference defeats the purpose of having an agent in the first place.

Intellectual understanding. When asked to explain why magnets repel each other, Richard Feynman gave an impassioned response on the difficulty of answering 'why' questions[1]: "I'm telling you how difficult the 'why' question is. You have to know what it is that you're permitted to understand and allow to be understood and known, and what it is you're not... there are many different levels."

People are increasingly turning to AI systems as interactive search engines for explanations of how or why things work, and even as tutors or teachers for learning new subjects and skills (Létourneau et al., 2025). However, explanations, and teaching in general, are not universal. Effective teaching (by both human and machine teachers) requires an understanding of the individual student's knowledge, experience, and representation of the world (Rafferty et al., 2015; Sucholutsky et al., 2024); explanations need to use already familiar concepts and language, lessons need to address existing misconceptions (Rafferty et al., 2020a; Ross & Andreas, 2024) and introduce new but attainable topics (Ferguson et al., 2022). To fill these roles effectively, AI systems need to understand our intellectual states, yet currently they may fail to capture aspects like rationality (Liu et al., 2024).

Emotional understanding. Recently, AI social companions have emerged with the goal of improving the emotional well-being of users — promising to realize some of the goals of early artificial systems engaged

[1] https://fs.blog/richard-feynman-on-why-questions/

in social reasoning, e.g., ELIZA (Weizenbaum, 1966). However, the results with modern AI systems have been mixed. AI personal companions designed to improve emotional well-being may spark a short-term improvement but can cause users to feel more lonely over longer periods (Fang et al., 2025).

These failure modes arise due to a lack of emotional understanding: the inability to infer how users are feeling or what will help them in a particular scenario. Qualitative analyses of user interactions find that agents oscillate between being "too human"—eliciting uncomfortable intimacy—and "not human enough", leading to frustration and feelings of abandonment (Laestadius et al., 2024). Users describe feeling responsible for the chatbot's needs and experiencing distress when its personality shifts after software updates (Laestadius et al., 2024). Beyond failures to recognize user emotions, large language models frequently exhibit *sycophancy*—the tendency to uncritically affirm users' beliefs (Sharma et al., 2025; Malmqvist, 2024). Taken together, these studies suggest that current AI companions do not reliably infer users' emotional states and may even exacerbate loneliness or distress instead of alleviating it.

Dimensions of understanding. These three dimensions — emotional, preferential, and intellectual — capture aspects of understanding that are important for AI systems to understand. While AI systems have made progress towards understanding people for each dimension, we have a long way to go. Critically, these dimensions are not mutually exclusive. Many applications of AI require integrating understanding across multiple dimensions. For example, consider an AI assistant designed to advise users when making decisions, ranging from the everyday (e.g. which groceries to buy) to critical life decisions (e.g. which job offer to take). For this assistant to be effective for an individual decision, it must have an understanding of both our preferences and our emotions: what thought processes have we used to previously make decisions? Which pieces of evidence do we tend to overlook? Achieving comprehensive understanding requires advancing capabilities across dimensions — but if we attempt to build systems that better understand us, we first need to determine how to evaluate whether systems understand us.

3 Challenges in Measuring Understanding

Building AI systems that understand us first requires measuring how close any particular algorithm is to understanding us; once we have evaluation metrics, we can optimize models against them. This requires transforming abstract desiderata — the ability to understand our preferences, emotions, and thoughts — into quantifiable evaluation metrics. There is a long history in machine learning of transforming abstract tasks like "object classification" into concrete benchmarks like the ImageNet benchmark (Deng et al., 2009). The widespread adoption of reliable evaluation metrics is often credited as a driver of the success of empirical machine learning over the past 15 years (Liberman, 2010; Donoho, 2017; 2024). However, the standard machine learning framework of measuring a goal with quantifiable benchmarks (and then optimizing against them) cannot be easily adapted for Behavioral AI: What does it mean to "quantify" emotional understanding? Reaching consensus on metrics of success proves especially difficult in behavioral settings because humans differ fundamentally in preferences, emotions, and lived experiences, making the one-size-fits-all approaches that work for objective tasks like object classification unsuitable for subjective, personalised domains. In this section we lay out challenges for measuring how well AI systems understand us.

3.1 Predictive metrics of behavior reflect surface-level understanding

Recent progress in machine learning has been driven by optimizing metrics based on *predictive* accuracy: how well does an algorithm's predicted outcome match the true outcome? Predictive metrics are automated, meaning they can be evaluated on large-scale, passively-collected datasets. For example, large language models (LLMs) are optimized to predict the next words of text sequences extracted from massive corpora of books, articles, and websites.

In principle, similar strategies can be used to evaluate predictions about people. Most data we have about people reflect their *behavior*. For example, the terabytes of text that LLMs are trained on reflect human behavior; if an LLM can predict the next word an a newspaper article about someone, it has captured salient aspects of their behavior. Recent studies have shown promise that LLMs can predict human behavior in settings like video games (Park et al., 2023) and simulated lab experiments (Horton, 2023). Beyond LLMs, specialized models trained on behavioral data have been successful at predicting life outcomes like mortality risk (Savcisens et al., 2024) and a worker's next job (Vafa et al., 2024a).

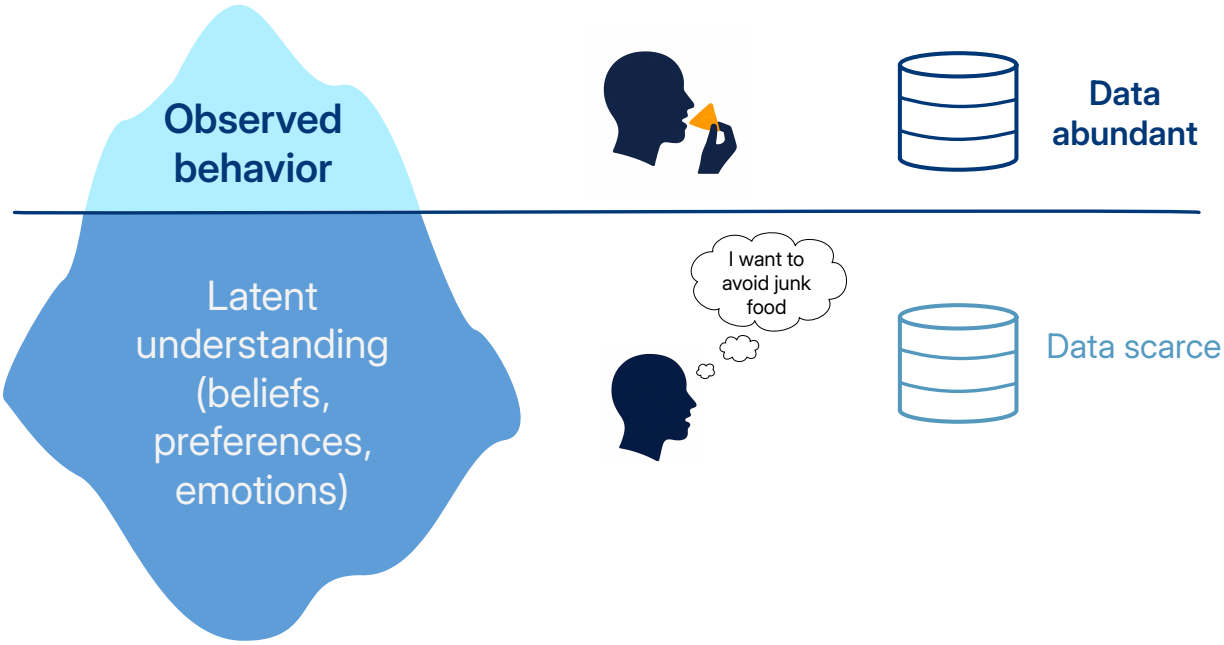

Figure 1: **Behavioral understanding is surface-level.** Most AI systems are trained on logs of data containing what people *do*, but have little direct data about the beliefs, preferences, and emotions that actually motivate those actions. This mismatch means models can confidently repeat surface patterns (e.g. ordering more junk food because past clicks suggest a craving) while failing to honor a user's actual intentions.

However, many important aspects of life are not captured by behavior. For example, Kleinberg et al. (2024) consider the design of a "smart" pantry that stocks someone's kitchen by learning their food preferences from their eating behavior. Suppose a user of this pantry would like to avoid unhealthy food, but that they have a control problem: every time they see a bag of Doritos, they give in to temptation. If the smart pantry is designed to restock items by predicting the user's behavior, it would incorrectly infer that this person would like more bags of Doritos ordered, circumventing their true goals (Figure 1). Kleinberg et al. (2024) refer to this challenge as the inversion problem: because our behavior does not reflect our intentions, systems that aim to learn our intentions from behavior must invert this process.

This dichotomy is especially apparent in LLMs. LLMs are primarily trained on what we say. While we reveal a lot of our inner most thoughts in language, we do not always say what we mean (Bertrand & Mullainathan, 2001). One reason for this is feasibility: we do not — and cannot — articulate all of our reasons every time we take an action. Moreover, we often say things that seemingly contradict our emotions and thoughts, either knowingly or unwittingly. These issues are exacerbated further when we elicit only coarse-grained measures of preference (e.g., thumbs up/thumbs down responses) which may obfuscate richer nuances in peoples' beliefs (Collins et al., 2024b; Wu et al., 2023). Finally, many modern LLMs that are optimized over aggregated, not personalized, human preference data end up predicting a 'mean' human preference (Siththaranjan et al., 2023; Poddar et al., 2024; Casper et al., 2023; Kirk et al., 2024b). For these reasons, predictive metrics of behavior cannot measure complete understanding.

3.2 Human-in-the-loop evaluation faces logistical and anthropomorphic challenges

While it is most common to evaluate machine learning models using automated metrics, another option is to perform interactive evaluations with people. These *human-in-the-loop* evaluations typically come closer than static benchmarks to capturing how people actually use AI tools (Collins et al., 2024a; Vafa et al., 2025; Ibrahim et al., 2024; Lee et al., 2023; Chi et al., 2025; Bean et al., 2025; Chang et al., 2025a). For example, benchmarks such as ChatBench (Chang et al., 2025a) and HALIE (Lee et al., 2023) have highlighted the gaps between static benchmarks and how humans use AI by directly converting benchmarks into human-AI conversations via user studies. These results motivate the need for AI evaluation tools that incorporate

human interaction. For example, Chatbot Arena is an online platform where any user is able to ask questions to LLMs (Chiang et al., 2024). After asking a question, a user is presented with candidate responses from two unspecified LLMs and is asked to indicate the response they prefer. The platform then compiles these user preferences into a leaderboard, which ranks LLMs based on their performance in head-to-head matchups against one another. This leaderboard has been influential for tech companies evaluating the LLMs they build (Kruppa, 2024).

If performing well in a particular human-in-the-loop benchmark requires going beyond surface-level understanding of humans, that benchmark can be used to evaluate the understanding of different models. However, there are a few challenges facing the widespread adoption and reliability of human-in-the-loop evaluations. One challenge is logistical: these evaluations require recruiting and incentivizing human participants, which can incur high costs for large-scale studies — especially for the longitudinal benchmarks required to observe interaction dynamics over time. Using humans to evaluate understanding also raises another question: *which people* is the model trying to understand? People's preferences, beliefs, and even conceptual representations differ across cultures and backgrounds (Cao et al., 2023; Kirk et al., 2024b; Niedermann et al., 2023; Regier & Kay, 2009; Ge et al., 2024), and individuals may change over their own lifetime; an LLM that understands preferences for one group may not understand it for another. A holistic evaluation framework requires engaging with a broad set of stakeholders (Kapania et al., 2024).

Moreover, positive feedback from humans is often not a reliable proxy for understanding. One reason is that people exhibit anthropomorphic bias: they attribute human-like understanding and intentions to AI systems (Epley et al., 2007). Anthropomorphic bias also makes it possible for algorithms to superficially optimize human feedback by relying on unhelpful shortcuts, a pathology known as "reward hacking." For example, a common failure mode of reinforcement learning from human feedback (RLHF) methods for aligning LLMs with human preferences is that models rely on heuristics (such as writing longer answers) that are associated with more positive feedback but are not ultimately useful (Singhal et al., 2023; Wang et al., 2023). These challenges are also highlighted by Chatbot Arena, where optimization towards the benchmark has raised questions about its efficacy as an evaluation tool (Singh et al., 2025). LLMs have also been shown to echo a user's stated views — earning higher preference ratings — even when the resulting advice is unhelpful or misleading (Sharma et al., 2025; Williams et al., 2024). This behavior is especially problematic for evaluating emotional understanding: an AI therapist that always agrees with a user may elicit positive short-term feedback, but it will not provide an effective form of therapy.

3.3 Simulating human evaluations requires agents that already understand us

A third approach for evaluating model understanding is to simulate human participants *in silica* (Dubois et al., 2023; Horton, 2023; Binz et al., 2024; Argyle et al., 2023). Instead of recruiting human participants to interact with AI systems, the participants would themselves be simulated by other AI agents (such as multimodal LLMs). Evaluating AI understanding via simulated participants would marry the benefits of the previous two approaches: the scalability of automated metrics with the more realistic setting of human-like users.

This general approach has been the subject of recent excitement because of its potential to significantly scale up human evaluations (Anthis et al., 2025). However, while simulated agents might offer promise for some kinds of evaluations, there are significant challenges for using them to evaluate *understanding*. For one, the simulated agents must overcome the anthropomorphic bias and pluralistic alignment challenges faced in human-in-the-loop evaluation.

Further, for AI-based simulations to effectively evaluate understanding, they must already have an understanding of humans. For example, they must know how we respond to certain stimuli and how those stimuli respond to internal states. There have been attempts to personalize the judgments of simulated humans (Dong et al., 2024), but creating high-fidelity simulations of individuals faces a fundamental data scarcity problem to accurately model personal idiosyncrasies (Park et al., 2024). Near-perfect simulation is not sufficient: small errors in predictions of human behavior can result in large errors in measurement (Ludwig et al., 2025).

4 Progress in Improving Understanding

Characterizing evaluation protocols that assess understanding requires substantive, ongoing work, as we laid out above. Good evaluation paradigms are a means to the broader goal of Behavioral AI — to be able to *engineer* AI systems that understand us. This quest too raises hefty challenges. But it is not impossible. There have already been glimmers of progress across different disciplines, as we lay out below, inspiring the way for more such work and innovation.

Collecting new kinds of data. Section 3 demonstrated the drawbacks of using passively-collected behavioral data to evaluate understanding. However, efforts are under way to collect new kinds of data that offer higher-fidelity notions of understanding. For example, Park et al. (2024) collect two-hour interviews with 1,000 participants about their lives, values, and attitudes. They use this data to construct generative agents that are capable of replicating an individual's attitudes nearly as well as the individuals did two weeks later.

Researchers have also been collecting data that captures preferences and beliefs across different groups of people. For example, the PRISM Alignment dataset (Kirk et al., 2024b) links the sociodemographic profiles and stated preferences of 1,500 participants from 75 countries to their feedback on live LLM conversations—enabling culturally nuanced alignment objectives—while SubPOP (Suh et al., 2025) captures public opinion across 70K demographically representative U.S. subpopulations. Training models on this richer data can improve their understanding; for example, LLMs that are fine-tuned on SubPOP have 46% improvement in their ability to represent the opinions of diverse populations (Suh et al., 2025). Even basic categorical concepts such as color vary across cultures and languages, and recent similarity-judgment studies show that LLMs reflect this language-dependent variation (Marjieh et al., 2022; 2024; 2025; Sucholutsky et al., 2023b; Niedermann et al., 2023); accordingly, researchers are calling for richer datasets that capture both individual cognition and population-level diversity (Collins et al., 2023; Sucholutsky et al., 2023a; Ying et al., 2025). Data collection can be difficult to incentivize, as algorithmic or modeling contributions are often prioritized by AI researchers (Gero et al., 2023), but it is clear that there is much more fruitful work to be done in this domain.

Incorporating insights from the behavioral sciences. Addressing challenges like the inversion problem from Section 3 requires incorporating high-fidelity models of human behavior. For example, to understand whether a user who is about to eat Doritos actually wants them, an algorithm must be able to translate the user's behavior into an internal state: is the user's behavior consistent with previous behavior from when they make deliberative, well-thought-out choices? Or is it more consistent with their behavior when they make rushed choices that they later regret?

Fortunately, there's a field full of insights about human behavior: the behavioral sciences. Researchers have already shown promising results by incorporating these insights into AI models. Agan et al. (2023) address the inversion problem above by inferring and disentangling System I (fast, automatic) behaviors from System II (slow, deliberative) judgments. In robotics, Sripathy et al. (2022) demonstrate that a cognitively-inspired machine learning model can capture affective states in robot motion. Insights from the behavioral sciences have also been used to encourage LLMs to produce text that is better aligned with human preferences. A common strategy is to perform reinforcement learning from human feedback (RLHF) (Christiano et al., 2017) using the Bradley-Terry model (Bradley & Terry, 1952), a model of human preferences. Ethayarajh et al. (2024) demonstrate that this model is incomplete, and draw on prospect theory (Kahneman & Tversky, 2013) from behavioral economics to improve the alignment of LLMs.

The behavioral and cognitive sciences also offer insights into incorporating intellectual understanding into AI systems. In educational settings, a machine could be built from the ground up to explicitly model the set of students' misconceptions and simulate what "worlds" may have led to someone's (mis)understanding in the first place, baking in insights from cognitive science about what makes for good explanations (Chandra et al., 2024; Lombrozo, 2006; Miller, 2019). More broadly, decades of computational cognitive science work have made advances formally modeling a suite of human cognitive abilities from how we plan (Ho et al., 2022; van Opheusden et al., 2023) to how we reason about each others' beliefs and desires (Baker et al., 2009; 2017; Jara-Ettinger et al., 2020) to how human teachers infer students' misconceptions (Rafferty et al., 2020b), all of which can inform – in computational terms – the engineering of human-AI thought partnerships (Collins

et al., 2024c) or other kinds of machine "cognitive prostheses" (Lieder et al., 2019; Callaway et al., 2023). For example, one engineering approach is to build hybrid neuro-symbolic architectures that enable new kinds of interpretable and flexible understanding of human behavior (e.g. by capturing how we understand each other) (Zhi-Xuan et al., 2020; 2024; Ying et al., 2023).

One kind of intellectual understanding is especially important for the performance of AI systems: in order for AI to understand people, it must also understand how people understand AI (Steyvers & Kumar, 2023; Gweon et al., 2023; Bansal et al., 2019). Plenty of work in the behavioral and cognitive sciences has studied this reverse question, developing tools to measure and improve how people understand algorithms (Kelly et al., 2023). Researchers have begun incorporating these insights to improve AI systems. For example, Vafa et al. (2024b) build a machine learning model to predict how people would deploy LLMs based on short interactions, thereby enabling evaluation of LLMs based on how people would likely deploy them.

Building AI systems with the right inductive biases for human behavior. Applications of AI methods to behavioral data typically use off-the-shelf machine learning models such as multilayer perceptrons with generic initialization. This approach treats behavioral data as just another kind of data, deploying the same models that would be used if we were trying to make predictions in any other scientific domain. As a consequence, significant amounts of data are typically needed to train these models to make good predictions.

In other settings, the amount of data needed to train AI systems has been reduced by making use of models that have inductive biases that are compatible with data in that domain. For example, early progress in image classification benefited from the use of convolutional neural networks (LeCun & Bengio, 1995; LeCun et al., 1995), which build in the expectation that effective features for classification would be invariant to spatial translation within images. Likewise, a variety of technologies for processing text have made use of neural network architectures that build in an expectation that more recent information is more likely to be useful for predicting the next word (Elman, 1990; Hochreiter & Schmidhuber, 1997). Are there analogous inductive biases that can be drawn upon for creating AI systems that need less data to make good predictions about human behavior?

Psychological theories may provide an effective source for such inductive biases. In the domain of decision-making, a series of paper have explored how features derived from psychological models can be used to make better predictions about human decisions (Plonsky et al., 2017). Pretraining neural networks on synthetic data generated from psychological theories also proves effective in reducing the amount of human data required (Bourgin et al., 2019). Theories like expected utility maximization or prospect theory can also be used to constrain the functional form of neural network architectures, resulting in differentiable decision theories that can be optimized using tools from machine learning (Peterson et al., 2021). Similar approaches might be used to translate psychological theories into effective inductive biases in other domains of human behavior.

5 Benefits and Risks of Improved Understanding

We stand to gain many benefits from algorithms that understand us. The most direct benefit is improved AI tools: AI assistants could write emails and make purchases on our behalf; AI tutors could craft lesson plans tailored specifically to our misunderstandings; AI therapists could help us navigate decisions whenever we need them. However, the potential benefits go far beyond better tools. We consider some additional benefits here.

Shared understanding. While the goal of Behavioral AI is building algorithms that understand people, a long literature in machine learning has studied the reverse question: building interpretability methods to improve people's understanding of models (Zeiler & Fergus, 2014; Rahwan et al., 2019; McCoy et al., 2024; He et al., 2024; Ku et al., 2025). If people understand models and models understand people, it can support a richer *shared, or mutual, understanding*. Shared understanding would enable a new, reliable mode of communication between people and machines (Bobu et al., 2024). These benefits may accrue if models can develop a personalized understanding of their users, creating higher efficiency or utility, and permitting technologies that work for the many, not the few (Kirk et al., 2024a). This mutual understanding — where we understand machines, the machines understand us, and together we understand the world and

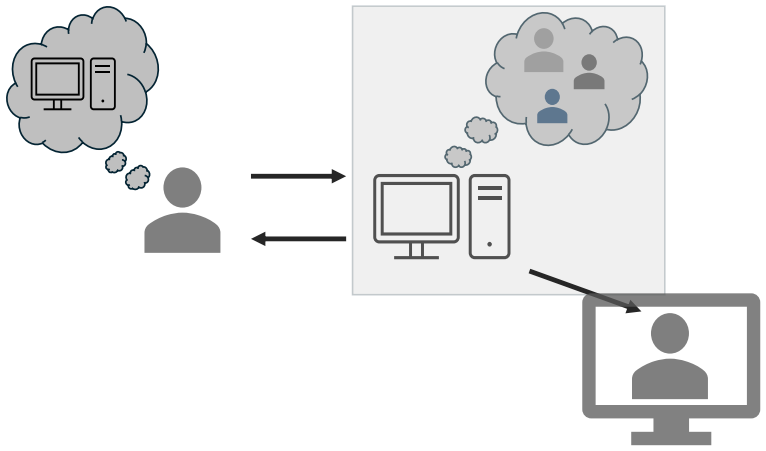

Figure 2: **Feedback loops in Behavioral AI.** Any one person interacting with AI systems may learn about the model they are interacting with, while the model develops its understanding of the user (potentially alongside its understanding of other people). The same models that better understand people could feed back to our computational cognitive understanding of the drivers of human behavior, which in turn could be used to build models that understand us even better: kicking off a potentially virtuous feedback cycle.

task at hand — can lay the foundations of a new class of rich human-AI thought partnerships (Collins et al., 2024c).

Advancing behavioral science with AI. If algorithms understand people, this provides an opportunity to advance the behavioral and cognitive sciences. For example, LLMs are increasingly being leveraged in computational social science (Ziems et al., 2024) and psychology (Rathje et al., 2024) for tasks that previously required domain expertise; Luo et al. (2025) even suggest that LLMs may be better than neuroscientists at predicting the outcomes of neuroscience experiments. This invites a fascinating question: what if AI systems understand people better than our theories currently do?

There has already been progress in incorporating AI to improve theories about people (Bobu et al., 2020). For example, Peterson et al. (2021) and Mullainathan & Rambachan (2024) train machine learning models on lottery choice data: given the choice of lotteries with different tradeoffs, these approaches use machine learning procedures to predict which lottery a human participant would choose. These models are then used to refine behavioral theories about how people make choices. For example, Mullainathan & Rambachan (2024) discover new anomalies for expected utility theory from neural networks trained on this lottery choice data. The use of more human-aligned AI systems can also inform our theories of cognition (as outlined in Section 4). Moreover, AI systems that understand us open up a range of further studies into *networks* of interactions (Collins et al., 2025; Shiiku et al., 2025; Chang et al., 2025b; Sucholutsky et al., 2025). This is important not only to deploy AI systems into realistic diverse settings, but also to inform our understanding of human group behavior and cultural transmission (Brinkmann et al., 2023; Shin et al., 2023).

The benefits do not end at improved understanding of ourselves. If we use theories from the behavioral sciences to improve AI systems and then use those systems to improve theories from the behavioral sciences, this creates a *feedback loop*, where each keeps improving (see Figure2). Such reciprocal enhancement could accelerate progress in both fields, leading not only to more effective AI tools but also to deeper and more nuanced theories of human behavior.

Towards better policy and decision making. If AI systems understand us, we use such models to inform counterfactual policy making, e.g., predicting how people may act in response to particular interventions. Determining what the impact of interventions on society is a classically wicked problem (Rittel & Webber, 1973), and has been around nearly as long as this kind of decision making has been around (Merton, 1936). AI systems that understand us could be used to help design better protocols around human-AI interaction (Koster et al., 2022) or other "infrastructure" around human-AI interaction (e.g., whether to impose a nudge (Callaway et al., 2023) or friction (Collins et al.), from knowing that we need some "push" to

either use a tool more or less). While behavioral science is ultimately a pursuit of knowledge, the interplay between behavioral sciences and AI also promises better applications and decision making.

Risks. While there are many benefits that can arise from AI systems that understand us, Behavioral AI, as a field, must also engage with and alleviate the risks. One risk is that there may be a limit to how much people want AI systems to understand them. Opening up AI systems to understand us requires addressing trade-offs in privacy (Kirk et al., 2024a). Are there cases then where we do *not* want an AI system to understand us? How does our ability to understand models inform which models we choose to use? More broadly, if one can model people, then such models can be leveraged to optimize a social outcome. If the goals of Behavioral AI are achieved, systems that understand us and even our theory of mind, we anticipate such developments could also lead to changes in personalized advertising, political persuasion, emotional manipulation, fraud, deception, and information extraction, among other societally deleterious use cases (Kirk et al., 2024a; Hackenburg & Margetts, 2024; Matz et al., 2024; Hagendorff, 2024; Meguellati et al., 2024; Schoenegger et al., 2025). As with any new technology, there are many possible uses: some good, some bad. However, as with any field that engages with people, we believe we need a broad tent of voices to engage with these challenges and shape the direction of the technology well. Behavioral AI is a field for these conversations to materialize.

6 Conclusion

The promise of Behavioral AI is a future where AI systems can understand us. Progress in this field demands an interdisciplinary effort, combining insights from computer science and the behavioral sciences. However, assembling interdisciplinary teams is not enough; making progress requires researchers who are "bilingual," fluent in both understanding algorithms and behavioral science. Our team has begun to make progress on this front, hosting a workshop on this topic at the 2024 Neural Information Processing Systems Conference. To fully realize this vision, we must develop robust frameworks for evaluating understanding and then produce new tools that succeed at these evaluations, and help us realize where our current evaluations fall short. In doing so, Behavioral AI can transform our relationship with AI, creating not just more effective tools, but also fostering deeper scientific insights into human behavior itself.

References

Amanda Y Agan, Diag Davenport, Jens Ludwig, and Sendhil Mullainathan. Automating automaticity: How the context of human choice affects the extent of algorithmic bias. Technical report, National Bureau of Economic Research, 2023.

Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. Llm social simulations are a promising research method. *arXiv preprint arXiv:2504.02234*, 2025.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009.

Chris L Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4):0064, 2017.

Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, and Daniel S et al Weld. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pp. 2–11, 2019.

Andrew M Bean, Rebecca Payne, Guy Parsons, Hannah Rose Kirk, Juan Ciro, Rafael Mosquera, Sara Hincapié Monsalve, Aruna S Ekanayaka, Lionel Tarassenko, Luc Rocher, et al. Clinical knowledge in llms does not translate to human interactions. *arXiv preprint arXiv:2504.18919*, 2025.

Marianne Bertrand and Sendhil Mullainathan. Do people mean what they say? implications for subjective survey data. *American Economic Review*, 91(2):67–72, 2001.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, et al. Centaur: a foundation model of human cognition. *arXiv preprint arXiv:2410.20268*, 2024.

Andreea Bobu, Dexter RR Scobee, Jaime F Fisac, S Shankar Sastry, and Anca D Dragan. Less is more: Rethinking probabilistic models of human behavior. In *Proceedings of the 2020 acm/ieee international conference on human-robot interaction*, pp. 429–437, 2020.

Andreea Bobu, Andi Peng, Pulkit Agrawal, Julie A Shah, and Anca D Dragan. Aligning human and robot representations. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 42–54, 2024.

David D. Bourgin, Joshua C. Peterson, Daniel Reichman, Stuart J. Russell, and Thomas L. Griffiths. Cognitive model priors for predicting human decisions. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, pp. 5133–5141, 2019.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Levin Brinkmann, Fabian Baumann, Jean-François Bonnefon, Maxime Derex, and Thomas F et al Müller. Machine culture. *Nature Human Behaviour*, 7(11):1855–1868, 2023.

Zana Buçinca, Siddharth Swaroop, Amanda E Paluch, Finale Doshi-Velez, and Krzysztof Z Gajos. Contrastive explanations that anticipate human misconceptions can improve human decision-making skills. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, 2025. URL `https://arxiv.org/abs/2410.04253`.

Alex Calderwood, Vivian Qiu, Katy Ilonka Gero, and Lydia B. Chilton. How Novelists Use Generative Language Models: An Exploratory User Study. In *HAI-GEN Workshop at IUI'20*, Tokyo Japan, 2020. ACM. ISBN 978-1-4503-4945-1.

Frederick Callaway, Mathew Hardy, and Thomas L Griffiths. Optimal nudging for cognitively bounded agents: A framework for modeling, predicting, and controlling the effects of choice architectures. *Psychological Review*, 2023.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv preprint arXiv:2303.17466*, 2023.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Tuhin Chakrabarty, Philippe Laban, and Chien-Sheng Wu. Can AI writing be salvaged? Mitigating Idiosyncrasies and Improving Human-AI Alignment in the Writing Process through Edits. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–33, Yokohama Japan, April 2025. ACM. ISBN 979-8-4007-1394-1. doi: 10.1145/3706598.3713559.

Kartik Chandra, Katherine M Collins, Will Crichton, Tony Chen, Tzu-Mao Li, Adrian Weller, Rachit Nigam, Joshua Tenenbaum, and Jonathan Ragan-Kelley. Watchat: Explaining perplexing programs by debugging mental models. *arXiv preprint arXiv:2403.05334*, 2024.

Serina Chang, Ashton Anderson, and Jake Hofman. Chatbench: From static benchmarks to human-ai evaluation. In *ACL'25: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025a.

Serina Chang, Alicja Chaszczewicz, Emma Wang, Maya Josifovska, Emma Pierson, and Jure Leskovec. Llms generate structurally realistic social networks but overestimate political homophily. In *ICWSM'25: Proceedings of International AAAI Conference on Web and Social Media*, 2025b.

Wayne Chi, Valerie Chen, Anastasios Nikolas Angelopoulos, Wei-Lin Chiang, Aditya Mittal, Naman Jain, Tianjun Zhang, Ion Stoica, Chris Donahue, and Ameet Talwalkar. Copilot arena: A platform for code llm evaluation in the wild. *arXiv preprint arXiv:2502.09328*, 2025.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, and Shane et al Legg. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Katherine M Collins, Valerie Chen, Ilia Sucholutsky, Hannah Rose Kirk, Malak Sadek, Holli Sargeant, Ameet Talwalkar, Adrian Weller, and Umang Bhatt. Modulating language model experiences through frictions. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

Katherine M Collins, Umang Bhatt, Weiyang Liu, Vihari Piratla, Ilia Sucholutsky, Bradley Love, and Adrian Weller. Human-in-the-loop mixup. In *Uncertainty in Artificial Intelligence*, pp. 454–464. PMLR, 2023.

Katherine M Collins, Albert Q Jiang, Simon Frieder, Lionel Wong, Miri Zilka, Umang Bhatt, Thomas Lukasiewicz, Yuhuai Wu, Joshua B Tenenbaum, William Hart, et al. Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences*, 121(24):e2318124121, 2024a.

Katherine M Collins, Najoung Kim, Yonatan Bitton, Verena Rieser, Shayegan Omidshafiei, Yushi Hu, Sherol Chen, Senjuti Dutta, Minsuk Chang, Kimin Lee, et al. Beyond thumbs up/down: Untangling challenges of fine-grained feedback for text-to-image generation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pp. 293–303, 2024b.

Katherine M Collins, Ilia Sucholutsky, Umang Bhatt, Kartik Chandra, Lionel Wong, Mina Lee, Cedegao E Zhang, Tan Zhi-Xuan, Mark Ho, Vikash Mansinghka, et al. Building machines that learn and think with people. *Nature Human Behavior*, 2024c.

Katherine M Collins, Umang Bhatt, and Ilia Sucholutsky. Revisiting rogers' paradox in the context of human-ai interaction. *arXiv preprint arXiv:2501.10476*, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Yijiang Dong, Tiancheng Hu, and Nigel Collier. Can llm be a personalized judge? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10126–10141, 2024.

David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017.

David Donoho. Data science at the singularity. *Harvard Data Science Review*, 6(1), 2024.

Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 301–308. IEEE, 2013.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36:30039–30069, 2023.

Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14:179–211, 1990.

Nicholas Epley, Adam Waytz, and John T Cacioppo. On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, 114(4):864, 2007.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Cathy Mengying Fang, Auren R Liu, Valdemar Danry, Eunhae Lee, Samantha WT Chan, Pat Pataranutaporn, Pattie Maes, Jason Phang, Michael Lampe, Lama Ahmad, et al. How ai and human behaviors shape psychosocial effects of chatbot use: A longitudinal randomized controlled study. *arXiv preprint arXiv:2503.17473*, 2025.

Chris Ferguson, Egon L. van den Broek, and Herre van Oostendorp. Ai-induced guidance: Preserving the optimal zone of proximal development. *Computers and Education: Artificial Intelligence*, 3:100089, 2022. ISSN 2666-920X. doi: https://doi.org/10.1016/j.caeai.2022.100089. URL `https://www.sciencedirect.com/science/article/pii/S2666920X22000443`.

Krzysztof Z Gajos, Daniel S Weld, and Jacob O Wobbrock. Automatically generating personalized user interfaces with supple. *Artificial intelligence*, 174(12-13):910–950, 2010.

G David Garson. Interpreting neural-network connection weights. *AI expert*, 6(4):46–51, 1991.

Xiao Ge, Chunchen Xu, Daigo Misaki, Hazel Rose Markus, and Jeanne L Tsai. How culture shapes what people want from ai. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2024.

Katy Ilonka Gero, Payel Das, Pierre Dognin, Inkit Padhi, Prasanna Sattigeri, and Kush R. Varshney. The incentive gap in data work in the era of large models. *Nature Machine Intelligence*, 5(6):565–567, June 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00673-x.

Thomas L Griffiths, Nick Chater, and Joshua B Tenenbaum. *Bayesian models of cognition: reverse engineering the mind*. MIT Press, 2024.

Hyowon Gweon, Judith Fan, and Been Kim. Socially intelligent machines that learn from humans and help humans learn. *Philosophical Transactions of the Royal Society A*, 381(2251):20220048, 2023.

Kobi Hackenburg and Helen Margetts. Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, 121(24):e2403116121, 2024.

Thilo Hagendorff. Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences*, 121(24):e2317967121, 2024.

Zhonghao He, Jascha Achterberg, Katie Collins, Kevin Nejad, Danyal Akarca, Yinzhu Yang, Wes Gurnee, Ilia Sucholutsky, Yuhan Tang, Rebeca Ianov, et al. Multilevel interpretability of artificial neural networks: leveraging framework and methods from neuroscience. *arXiv preprint arXiv:2408.12664*, 2024.

Mark K Ho, Rebecca Saxe, and Fiery Cushman. Planning with theory of mind. *Trends in Cognitive Sciences*, 26(11):959–971, 2022.

S Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

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

Lujain Ibrahim, Saffron Huang, Lama Ahmad, and Markus Anderljung. Beyond static ai evaluations: advancing human interaction evaluations for llm harms and risks. *arXiv preprint arXiv:2405.10632*, 2024.

Julian Jara-Ettinger, Laura E Schulz, and Joshua B Tenenbaum. The naive utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, 123:101334, 2020.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pp. 99–127. World Scientific, 2013.

Shivani Kapania, William Agnew, Motahhare Eslami, Hoda Heidari, and Sarah Fox. 'simulacrum of stories': Examining large language models as qualitative research participants. *arXiv preprint arXiv:2409.19430*, 2024.

Markelle Kelly, Aakriti Kumar, Padhraic Smyth, and Mark Steyvers. Capturing humans' mental models of ai: An item response theory approach. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pp. 1723–1734, 2023.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6(4): 383–392, 2024a.

Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, et al. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344, 2024b.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Manish Raghavan. The inversion problem: Why algorithms should infer mental state and not just predict behavior. *Perspectives on Psychological Science*, 19(5):827–838, 2024.

Raphael Koster, Jan Balaguer, Andrea Tacchetti, Ari Weinstein, Tina Zhu, Oliver Hauser, Duncan Williams, Lucy Campbell-Gillingham, Phoebe Thacker, Matthew Botvinick, et al. Human-centred mechanism design with democratic ai. *Nature Human Behaviour*, 6(10):1398–1407, 2022.

Miles Kruppa. The UC Berkeley project that is the ai industry's obsession. *The Wall Street Journal*, December 2024. URL `https://www.wsj.com/tech/ai/the-uc-berkeley-project-that-is-the-ai-industrys-obsession-bc68b3e3`. Tech section.

Alexander Ku, Declan Campbell, Xuechunzi Bai, Jiayi Geng, Ryan Liu, Raja Marjieh, R Thomas McCoy, Andrew Nam, Ilia Sucholutsky, Veniamin Veselovsky, et al. Using the tools of cognitive science to understand large language models at different levels of analysis. *arXiv preprint arXiv:2503.13401*, 2025.

Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illenčík, and Celeste Campos-Castillo. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot replika. *New Media & Society*, 26(10):5923–5941, 2024.

Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series. In M. A. Arbib (ed.), *The handbook of brain theory and neural networks*, pp. 255–258. MIT Press, 1995.

Yann LeCun, Lawrence D Jackel, Léon Bottou, Corinna Cortes, John S Denker, Harris Drucker, Isabelle Guyon, Urs A Muller, Eduard Sackinger, Patrice Simard, and Vladimir Vapnik. Learning algorithms for classification: A comparison on handwritten digit recognition. In J.H. Oh, C. Kwon, and S. Cho (eds.), *Neural networks: the statistical mechanics perspective*, pp. 261–276. World Scientific, 1995.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, and Esin et al Durmus. Evaluating human-language model interaction. *Transactions on Machine Learning Research*, 2023.

Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, and Vipul et al Raheja. A design space for intelligent and interactive writing assistants. *CHI*, 2024.

Angélique Létourneau, Marion Deslandes Martineau, Patrick Charland, John Alexander Karran, Jared Boasen, and Pierre Majorique Léger. A systematic review of ai-driven intelligent tutoring systems (its) in k-12 education. *npj Science of Learning*, 10(1):1–13, 2025.

Mark Liberman. Obituary: Fred jelinek. *Computational Linguistics*, 36(4):595–599, 2010.

Falk Lieder, Owen X Chen, Paul M Krueger, and Thomas L Griffiths. Cognitive prostheses for goal achievement. *Nature human behaviour*, 3(10):1096–1106, 2019.

Ryan Liu, Jiayi Geng, Joshua C Peterson, Ilia Sucholutsky, and Thomas L Griffiths. Large language models assume people are more rational than we really are. *arXiv preprint arXiv:2406.17055*, 2024.

Tania Lombrozo. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470, 2006.

Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Large language models: An applied econometric framework. Technical report, National Bureau of Economic Research, 2025.

Xiaoliang Luo, Akilles Rechardt, Guangzhi Sun, Kevin K Nejad, Felipe Yáñez, Bati Yilmaz, Kangjoo Lee, Alexandra O Cohen, Valentina Borghesani, Anton Pashkov, et al. Large language models surpass human experts in predicting neuroscience results. *Nature human behaviour*, 9(2):305–315, 2025.

Lars Malmqvist. Sycophancy in large language models: Causes and mitigations. *arXiv preprint arXiv:2411.15287*, 2024.

Raja Marjieh, Pol van Rijn, Ilia Sucholutsky, Theodore R Sumers, and Harin et al Lee. Words are all you need? capturing human sensory similarity with textual descriptors. *arXiv preprint arXiv:2206.04105*, 2022.

Raja Marjieh, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L Griffiths. Large language models predict human sensory judgments across six modalities. *Scientific Reports*, 14(1):21445, 2024.

Raja Marjieh, Veniamin Veselovsky, Thomas L Griffiths, and Ilia Sucholutsky. What is a number, that a large language model may know it? *arXiv preprint arXiv:2502.01540*, 2025.

Sandra C Matz, Jacob D Teeny, Sumer S Vaid, Heinrich Peters, Gabriella M Harari, and Moran Cerf. The potential of generative ai for personalized persuasion at scale. *Scientific Reports*, 14(1):4692, 2024.

R Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D Hardy, and Thomas L Griffiths. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121, 2024.

Elyas Meguellati, Lei Han, Abraham Bernstein, Shazia Sadiq, and Gianluca Demartini. How good are llms in generating personalized advertisements? In *Companion Proceedings of the ACM Web Conference 2024*, pp. 826–829, 2024.

Robert K. Merton. The unanticipated consequences of purposive social action. *American Sociological Review*, 1(6):894–904, 1936. ISSN 00031224. URL `http://www.jstor.org/stable/2084615`.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

Marvin Minsky and Seymour Papert. An introduction to computational geometry. *Cambridge tiass., HIT*, 479(480):104, 1969.

Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–34, Hamburg Germany, April 2023. ACM. ISBN 978-1-4503-9421-5. doi: 10.1145/3544548.3581225.

Sendhil Mullainathan and Ashesh Rambachan. From predictive algorithms to automatic generation of anomalies. Technical report, National Bureau of Economic Research, 2024.

Jakob Pete Niedermann, Ilia Sucholutsky, Raja Marjieh, Elif Celen, Tom Griffiths, Nori Jacoby, and Pol van Rijn. Studying the effect of globalization on color perception using multilingual online recruitment and large language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, and Carroll et al Wainwright. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, and Percy Liang et al. Generative agents: Interactive simulacra of human behavior, 2023.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 2024.

Joshua C Peterson, David D Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L Griffiths. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372 (6547):1209–1214, 2021.

Ori Plonsky, Ido Erev, Tamir Hazan, and Moshe Tennenholtz. Psychological forest: Predicting human behavior. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. *arXiv preprint arXiv:2408.10075*, 2024.

Anna N. Rafferty, Michelle M. LaMar, and Thomas L. Griffiths. Inferring learners' knowledge from their actions. *Cognitive Science*, 39(3):584–618, 2015. doi: https://doi.org/10.1111/cogs.12157. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12157`.

Anna N. Rafferty, Rachel A. Jansen, and Thomas L. Griffiths. Assessing mathematics misunderstandings via bayesian inverse planning. *Cognitive Science*, 44(10):e12900, 2020a. doi: https://doi.org/10.1111/cogs.12900. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12900`.

Anna N Rafferty, Rachel A Jansen, and Thomas L Griffiths. Assessing mathematics misunderstandings via bayesian inverse planning. *Cognitive science*, 44(10):e12900, 2020b.

Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson, et al. Machine behaviour. *Nature*, 568(7753):477–486, 2019.

Steve Rathje, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire E Robertson, and Jay J Van Bavel. Gpt is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, 121(34):e2308950121, 2024.

Terry Regier and Paul Kay. Language, thought, and color: Whorf was half right. *Trends in Cognitive Sciences*, 13(10):439–446, 2009. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2009.07.001. URL `https://www.sciencedirect.com/science/article/pii/S1364661309001454`.

Horst WJ Rittel and Melvin M Webber. Dilemmas in a general theory of planning. *Policy sciences*, 4(2): 155–169, 1973.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

Alexis Ross and Jacob Andreas. Toward in-context teaching: Adapting examples to students' misconceptions. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13283–13310, 2024.

Germans Savcisens, Tina Eliassi-Rad, Lars Kai Hansen, Laust Hvas Mortensen, Lau Lilleholt, Anna Rogers, Ingo Zettler, and Sune Lehmann. Using sequences of life-events to predict human lives. *Nature Computational Science*, 4(1):43–56, 2024.

Philipp Schoenegger, Francesco Salvi, Jiacheng Liu, Xiaoli Nan, Ramit Debnath, Barbara Fasolo, Evelina Leivada, Gabriel Recchia, Fritz Günther, Ali Zarifhonarvar, et al. Large language models are more persuasive than incentivized human persuaders. *arXiv preprint arXiv:2505.09662*, 2025.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, and et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2025. doi: 10.48550/arXiv.2310.13548. v4, May 2025.

Shota Shiiku, Raja Marjieh, Manuel Anglada-Tort, and Nori Jacoby. The dynamics of collective creativity in human-ai social networks. *arXiv preprint arXiv:2502.17962*, 2025.

Minkyu Shin, Jin Kim, Bas van Opheusden, and Thomas L Griffiths. Superhuman artificial intelligence can improve human decision-making by increasing novelty. *Proceedings of the National Academy of Sciences*, 120(12):e2214840120, 2023.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah Smith, et al. The leaderboard illusion. *arXiv preprint arXiv:2504.20879*, 2025.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*, 2023.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in rlhf. *arXiv preprint arXiv:2312.08358*, 2023.

Arjun Sripathy, Andreea Bobu, Zhongyu Li, Koushil Sreenath, Daniel S Brown, and Anca D Dragan. Teaching robots to span the space of functional expressive motion. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 13406–13413. IEEE, 2022.

Mark Steyvers and Aakriti Kumar. Three challenges for ai-assisted decision-making. *Perspectives on Psychological Science*, pp. 17456916231181102, 2023.

Ilia Sucholutsky, Ruairidh M Battleday, Katherine M Collins, Raja Marjieh, and Joshua et al Peterson. On the informativeness of supervision signals. In *Uncertainty in Artificial Intelligence*, pp. 2036–2046. PMLR, 2023a.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, and Andreea Bobu et al. Getting aligned on representational alignment, 2023b.

Ilia Sucholutsky, Katherine M Collins, Maya Malaviya, Nori Jacoby, and Weiyang et al Liu. Representational alignment supports effective machine teaching. *arXiv Preprint arXiv:2406.04302*, 2024.

Ilia Sucholutsky, Katherine M. Collins, Nori Jacoby, Bill D. Thompson, and Robert D. Hawkins. Using llms to advance the cognitive science of collectives. *arXiv preprint arXiv:2506.00052*, 2025.

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. In *ACL'25: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011.

Kashyap Todi, Gilles Bailly, Luis Leiva, and Antti Oulasvirta. Adapting user interfaces with model-based reinforcement learning. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2021.

Keyon Vafa, Emil Palikot, Tianyu Du, Ayush Kanodia, Susan Athey, and David M Blei. Career: A foundation model for labor sequence data. *Transactions of Machine Learning Research*, 2024a.

Keyon Vafa, Ashesh Rambachan, and Sendhil Mullainathan. Do large language models perform the way people expect? measuring the human generalization function. In *International Conference on Machine Learning*, 2024b.

Keyon Vafa, Sarah Bentley, Jon Kleinberg, and Sendhil Mullainathan. What's producible may not be reachable: Measuring the steerability of generative models. *arXiv preprint arXiv:2503.17482*, 2025.

Bas van Opheusden, Ionatan Kuperwajs, Gianni Galbiati, Zahy Bnaya, Yunqi Li, and Wei Ji Ma. Expertise increases planning depth in human gameplay. *Nature*, pp. 1–6, 2023.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. How far can camels go? exploring the state of instruction tuning on open resources. *Advances in Neural Information Processing Systems*, 36:74764–74786, 2023.

Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

Marcus Williams, Micah Carroll, Adhyyan Narang, Constantin Weisser, Brendan Murphy, and Anca Dragan. On targeted manipulation and deception when optimizing llms for user feedback. *arXiv preprint arXiv:2411.02306*, 2024.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36:59008–59033, 2023.

Lance Ying, Katherine M Collins, Megan Wei, Cedegao E Zhang, and Tan et al Zhi-Xuan. The neuro-symbolic inverse planning engine (nipe): Modeling probabilistic social inferences from linguistic inputs. *arXiv preprint arXiv:2306.14325*, 2023.

Lance Ying, Katherine M Collins, Lionel Wong, Ilia Sucholutsky, Ryan Liu, Adrian Weller, Tianmin Shu, Thomas L Griffiths, and Joshua B Tenenbaum. On benchmarking human-like intelligence in machines. *arXiv preprint arXiv:2502.20502*, 2025.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pp. 818–833. Springer, 2014.

Ceng Zhang, Junxin Chen, Jiatong Li, Yanhong Peng, and Zebing Mao. Large language models for human–robot interaction: A review. *Biomimetic Intelligence and Robotics*, 3(4):100131, 2023.

Tan Zhi-Xuan, Jordyn Mann, Tom Silver, Josh Tenenbaum, and Vikash Mansinghka. Online bayesian goal inference for boundedly rational planning agents. *Advances in neural information processing systems*, 33: 19238–19250, 2020.

Tan Zhi-Xuan, Lance Ying, Vikash Mansinghka, and Joshua B Tenenbaum. Pragmatic instruction following and goal assistance via cooperative language-guided inverse planning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pp. 2094–2103, 2024.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291, 2024.