Forthcoming in *The Role of Artificial Intelligence in Science: Methodological and Epistemological Studies*
(Darrell P. Rowbottom, Andre Curtis-Trudel & David L. Barack, eds.), Routledge.

AI Surrogacy in Psychological Research

William D'Alessandro
Philosophy, William & Mary (wbdalessandro@wm.edu)

Jessica A. F. Thompson
Experimental Psychology, University of Oxford (jessica.thompson@psy.ox.ac.uk)

Abstract

AI tools hold considerable promise for psychological research. The precise shape of their potential uses has become clearer in recent years as machine learning models have been trained to reproduce a variety of complex human cognitive behaviors with impressive success. The prospect of AI-human performance parity, along with the advantages of AI systems in speed, cost and ease of use, has prompted psychologists to explore how science might benefit from reassigning some traditionally human research roles to machines. This chapter provides an outlook on such methods. We begin by documenting the prospective upsides of what many researchers take to be the most promising types of AI surrogacy, including the use of models to efficiently generate, screen and refine preliminary hypotheses. We then discuss such methods' limitations and drawbacks, and more generally the methodological considerations researchers must attend to in choosing when and how to rely on machine substitutes for human behavior.

1 "Can AI models replace human participants?"

A 2023 paper in *Trends in Cognitive Sciences*, cited hundreds of times already as of this writing in late 2024, proposed a guardedly affirmative answer to the question "Can AI Language Models Replace Human Participants?" (Dillion et al. 2023). The floodgates have come thundering open in the short time since. A large and fast-growing body of work in neuroscience, psychology and behavioral science has embraced AI tools as surrogates for human subjects and researchers, in ways already extensive enough to make Dillion et al.'s circumscribed vision look rather modest.

Many commentators have, of course, sounded notes of concern about the supplanting of humans by large language models (LLMs) and other AI systems. Some of these worries pertain to AI in science broadly (Messeri and Crockett 2024; Koskinen 2024; Binz, Alaniz, et al. 2025), while others involve

specific use cases in psychology and adjacent fields (Grossmann et al. 2023; Abdurahman et al. 2024; Crockett and Messeri 2023). Our goal is to provide an overview of all sides of this developing picture: how psychologists are applying (or proposing to apply) AI to traditionally human research tasks, why many are optimistic about these methods, and what we take to be the crucial methodological limitations and challenges of the AI surrogacy paradigm.

The practices we describe also raise many ethical questions, but we limit our discussion here to the methodological and epistemological issues surrounding AI surrogacy. (Or rather we do so to the extent that the two can be disentangled, though there remain points of close contact. Examples include the role of shared perspective in interpreting AI outputs (§2.5), LLMs' misportrayal and flattening of minority group characteristics (§3.4), and the relationship between human interests and the goals of science (§3.5).) For more sustained analysis of AI ethics issues in psychology, see e.g. Agnew et al. 2024; Chenneville, Duncan, and Silva 2024; Chen et al. 2024.

2 The promise: what AI surrogacy might do for research

The topic we've gestured at stands in need of clarification: what sort of replacement is at issue in discussions of AI taking over human research roles? We understand "AI surrogacy" here to cover any human-associated tasks substantially assigned to artificial intelligence, especially those constituting central cognitive components of theoretical and experimental research.[1] (So our focus here will not be on, for example, the use of AI tools for proofreading papers or managing lab schedules.) This notion of surrogacy is intentionally broad. In addition to cases of direct replacement (where an AI system performs a task *T* but, had it not, some human would have done *T* instead), our definition applies to cases of AI systems doing the *kinds* of cognitive tasks humans have traditionally done, even if it's unclear or doubtful whether any human would counterfactually have performed the specific tasks at issue.[2]

Though the title of Dillion et al.'s piece asks about replacing human participants, we consider surrogacy on both the researcher and subject sides; the two sets of issues raised are intertwined in any case.

[1] In its taxonomy of "visions of AI across the research pipeline", Messeri and Crockett (2024) uses the language of surrogacy to refer specifically to AI-generated synthetic data. Our broader notion of surrogacy includes this use (discussed especially in §2.5) among others.

[2] It may of course be important to distinguish carefully between replacement-type and augmentation-type cases in some contexts. When discussing general methods and hypothetical scenarios (as we often will), however, one can rarely give definite answers to counterfactual questions about whether a given AI research task would otherwise have been done by humans. So it's convenient to adopt a notion of surrogacy that's insensitive to such issues.

So far as we know, not even the most ardent techno-optimists have called for the end-to-end mechanization of psychology, at least in the near term (though see Lu et al. (2024) and Manning, Zhu, and Horton (2024) for workups of autonomous AI researchers in machine learning and social science, respectively). Rather, AI proponents have identified a number of specific locations along the research pipeline where machine surrogacy offers apparent advantages. We discuss five such applications in the subsections below: AI-driven hypothesis generation, item piloting and instrument development, the exploration of novel experimental paradigms, general-purpose cognitive modeling, and the replacement of human study subjects. We sketch the state of the art and mention some methodological issues in each case. Subsequently, §3 explores concerns about AI methods of a more general sort.

2.1 Hypothesis generation

The engines of scientific progress run well only when fed by a steady stream of questions, problems and puzzles. Advancements in knowledge are therefore bottlenecked by the number of high-quality testable hypotheses which experts are able to produce. Typical human skills in this domain may, moreover, leave something to be desired, given our cognitive limitations and biases; as Berger (2024) points out, "relying on intuition, personal observation, or whatever literature [scientists] happen to be aware of" (798) hardly stands out as a bulletproof method for identifying the best ideas.

It's therefore no surprise that AI systems' hypothesis-generating potential has generated significant interest. Speed and cost are obvious points in the machines' favor: current LLMs can produce dozens of research proposals in seconds, for free or nearly so. Sifting through many such outputs might yield promising ideas relatively quickly even if most were of low quality.

In fact, however, the perceived goodness of AI-authored hypotheses has also risen to an estimable level. Banker et al. (2024) put GPT-4 to work producing research proposals and found that social psychologists rated its ideas above human-generated hypotheses on each of five dimensions of quality (clarity, originality, impact, plausibility and relevance). There's some evidence that more sophisticated methods may yield further improvements: Tong et al. (2024) used GPT-4 to create a causal graph for psychological variables, which was then used as a basis for automated hypothesis generation; proposals produced in this way outscored those produced by a baseline LLM in perceived novelty (though not in perceived usefulness). Relatedly, good agreement has been found between LLMs and human raters asked to judge the quality of AI-generated hypotheses in the field of natural language processing (Chai et al. 2024). Other recent works in this vein include (Liu et al. 2024; O'Brien et al. 2024; Xu et al. 2024; Zhou et al. 2024).

Of course, the fact that AI hypotheses are judged favorably in certain respects (by humans or AI itself) doesn't entail that these hypotheses are especially likely to be true, fruitful or otherwise scientifically virtuous. Indeed, LLM outputs are often found to be preferred over alternatives produced by human experts, seemingly on account of the fluent, readily digestible nature of LLM text; this effect has been observed in domains from poetry (Porter and Machery 2024) to talk therapy (Kuhail et al. 2024), and is often accompanied by an impression that AI outputs are more humanlike than genuine human work. It's conceivable that many AI-generated hypotheses are also preferred for broadly stylistic and presentational rather than epistemic reasons.[3] As far as we're aware, few efforts have yet been made to determine whether raters' preferences are correlated with positive experimental findings or other objective dimensions of scientific merit.

Several studies have, however, attempted to directly measure the accuracy of LLM predictions. Lippert et al. (2024) found GPT-4 to achieve parity with human experts in predicting the results of a large social psychology experiment. See also Luo et al. (2024) for LLMs' ability to predict the correct versions of published neuroscience abstracts, and Hewitt et al. (2024) for LLM predictions of social science survey results.

LLM-based methods are a popular paradigm for hypothesis generation, but not the only AI game in town. An alternative approach uses standard machine learning algorithms to discover prospective relationships between variables of interest, together with interpretability techniques to render these hypotheses intelligible to human scientists. Ludwig & Mullainathan (2024) applied a version of this method to judges' decisions about pretrial detention, finding humanly identifiable but non-obvious features of arrestees' mugshots which significantly predicted detention versus release. As the authors note, however, their method depends on the availability of unstructured, high-dimensional data from which humanly meaningful features can be extracted, and hence is likely to be useful only for a limited set of research questions. Shang & Xiao (2023) explores similar approaches to automated hypothesis generation in neuroscience, again acknowledging the dataset-dependent nature of such techniques: in this case, the authors "focus on large-scale electrical and optical physiology data because they best represent the rich, complex patterns that AI thrives on" (2).

[3] This might explain the findings of Banker et al., for instance, whose AI-generated hypotheses seem noticeably more vivid, detailed, self-contained and well-motivated than the human alternatives (which were scraped by machine from the abstracts of published papers). A typical GPT-4 hypothesis from this study was "that individuals belonging to a low-status group within a society will have more positive attitudes towards a high-status outgroup, in comparison to individuals in the high-status group themselves, due to a phenomenon of upward social comparison and aspirational identification". Meanwhile, a typical human hypothesis was "that people may have multiple representations of a preference toward an object even within a single context".

2.2 Instrument development and piloting

Psychological theories frequently posit unobserved mental traits, like agreeableness, intelligence or social anxiety, to explain observed behavior. They develop *instruments*—tests, scales, questionnaires and the like—to indirectly measure these hypothesized traits. Instruments consist of *items* such as survey questions and task prompts. Both instruments and individual items must be carefully evaluated prior to wide adoption in experimental settings. *Piloting*, for instance, refers to preliminary item testing for comprehensibility, clarity, inter-item correlations and other statistical properties.

One of Dillion et al.'s primary proposals for how LLMs might supplant humans in the research pipeline is by taking over item piloting in the design of new experiments and instruments. They suggest, for instance, that "[r]esearchers can give LLMs different questions and see if they act as expected within a nomological net (e.g., form a reliable scale)" (598).[4] Argyle et al. (2023) similarly encourages researchers to "leverage the insights gained from simulated, silicon samples to pilot different question wording, triage different types of measures, identify key relationships to evaluate more closely, and come up with analysis plans prior to collecting any data with human participants" (349).

Beghetto et al. (2025) discuss such possibilities at greater length, noting three opportunities for AI automation in instrument development. First, language models could stand in for human respondents in "cognitive interviewing" protocols, helping to ensure that items are worded clearly so as to convey their intended meanings to study participants. Second, LLMs might be used to spot global misalignment issues and other inconsistencies by "recogniz[ing] patterns in textual information and analyz[ing] the semantic relationship among items, constructs, and definitions" (1). Finally, LLMs could contribute to item authoring or revising, with an eye toward maintaining stylistic consistency, reading-level appropriateness and other desired features.

Beghetto et al. report on two case studies featuring the use of LLMs for these purposes. In the first case, researchers sought to improve on a possibility-thinking scale. GPT-4 identified three flawed items from the original version of the scale, two of which were also found to be problematically cross-loaded in

[4] Cronbach & Meehl (1955) introduced the notion of a nomological net(work), characterized as "the interlocking system of laws which constitute a theory" and which "may relate (a) observable properties or quantities to each other; or (b) theoretical constructs to observables; or (c) different theoretical constructs to one another" (290). On this picture, establishing the construct validity of a psychological instrument involves identifying a relevant network and showing that the instrument behaves as expected with respect to it. For instance, scores on an assessment of social anxiety should correlate positively with measures of general anxiety and negatively with measures of social confidence, should predict behavioral outcomes like speaking performance in social situations, and should be distinguishable from measures of non-social anxiety.

independent empirical work.[5] (Beghetto et al. report surprise at GPT-4's "particularly insightful" analysis here (3).) The LLM also generated a number of novel items. The authors found that the best-performing modified version of the scale used nine human-produced and three GPT-created items, furnishing "promising evidence that partnering with LLMs can support and strengthen scale development and validation efforts" (4).

Schlegel, Sommer, and Mortillaro (2025) used GPT-4 to generate new items for tests of emotional intelligence. They found that the GPT-generated items were highly correlated with the established items ($r = 0.46$) and were largely similar in their psychometric properties (clarity, difficulty, realism, diversity, internal consistency). Interestingly, when they asked six widely-used LLMs to complete established tests of emotional intelligence, all models scored considerably higher than humans on average (81% versus 56% mean accuracy). Thus, simply administering candidate items to these LLMs, without additional prompt engineering, would not yield results that reflect typical human performance. Indeed, GPT's superior ability here may be part of why it was successful in creating new items.

The negative results of Abdurahman et al. (2024) provide a partial contrast, however. Here the authors took up Dillion et al.'s suggestion about using LLMs to construct nomological nets, prompting GPT-3.5 to produce judgments in six moral domains and comparing these results to human data. The correlations among domains in the two datasets differed considerably (e.g., purity and loyalty judgments were correlated positively in the human sample but negatively in GPT's responses), "indicating that GPT struggles to [re]produce previously established nomological networks" (4).

Some elements of item piloting and instrument development draw primarily on language models' general-purpose writing and text analysis skills. These applications raise few issues beyond those associated with typical uses of generative AI. In other instances, however, additional care is warranted, and it's unclear to what extent AI systems can offer a viable alternative to human subjects. For example, as Harding et al. (2023) point out, a key goal of cognitive interviewing and other piloting techniques is to assess the reception of test items by subjects with widely varying prior knowledge, linguistic backgrounds, reading abilities, attentional strategies and other cognitive characteristics. These techniques often include prompts such as "How easy or difficult was it for you to remember this? How sure are you of your answer? How easy or difficult was it for you to come up with your answer? How

[5] An item is said to be cross-loaded when it measures more than one psychological factor to a significant degree. For example, in a Big Five personality trait assessment, the item "I enjoy meeting new people at parties" might load on both Extraversion and Openness to Experience. Cross-loading is undesirable because it complicates the scoring and interpretation of test items (among other reasons).

easy or difficult is it for you to talk about this issue?" (Hibben and de Jong 2016). The extent to which LLMs can simulate such subjects in answering such questions (or can be finetuned to do so more successfully) is unclear at present. (See §3.4 below for related discussion.)

2.3 Novel experimental paradigms

If an AI system can be trained to successfully mimic human behaviour in a task of interest, it can provide essentially infinite human-like behavioural data. Such a *behavioural clone* can be used in new research paradigms that would not be possible using exclusively human participants. Below we describe one such methodology, variants of which have been used successfully to discover novel strategies to positively influence human behaviour.

The behavioural cloning pipeline begins by collecting large amounts of human data in a task of interest. Neural networks are then trained to mimic human behaviour in the task—these are the behavioural clones. Importantly, the clones are not trained to perform the task as well as possible; they are optimized to behave as humans do in the task. A separate deep reinforcement learning system then learns a policy to maximize some aspect of the clones' behaviour. This generates a hypothesis about how to similarly influence human behaviour, which can be tested on a new cohort of human participants. This new human data can be incorporated into the training data for the behavioural clones in an iterative process that gradually refines the learned policy. Inspection of the learned policy ideally leads to an interpretable version that achieves the same effect. Variations of this method have been used successfully to encourage cooperation in a prisoner's dilemma game (McKee et al. 2023) to encourage sustainable behaviour and find desirable social mechanisms in a common pool resource problem (Koster et al. 2022; 2025).

Unlike some other forms of AI surrogacy we discuss, the behavioural clone is not replacing human participants in a setting where humans would normally be queried. Theoretically, one can imagine skipping the behavioural cloning step and having the deep reinforcement learning system interact directly with humans, but this would be intractable in practice. By contrast, the ability to "clone" human behaviour makes speedy optimization possible, opening up new data-driven approaches to discovery and hypothesis generation. An optimistic take might suggest that this presents an answer to the challenge posed by Allen Newell over 50 years ago in "You can't play 20 questions with nature and win" (1973). Rather than coming up with theoretically-motivated hypotheses and designing experimental conditions to compare in human experiments, psychologists and social scientists, armed with powerful optimization algorithms, can directly read out the conditions that are optimal (at least

for the behavioural clones). Thus, the primary benefit of this novel experimental paradigm may be the acceleration of scientific discovery. Gradient descent is a more efficient algorithm than binary search.

Behavioural cloning experimental paradigms are similar in form to the practice of training language models with Reinforcement Learning with Human Feedback (RLHF), a common method for aligning language model outputs with human preferences. In RLHF, a *reward model* (or *preference model*) is trained on human preferences over prompt-response pairs. This reward model then stands in for human preferences, similar to how the behavioural clone stands in for human behaviour, in a finetuning process that leads to more human-preferred model outputs. RLHF was an essential ingredient in the recipe that brought the impressive success of modern large language models (Ziegler et al. 2020; Ouyang et al. 2022), so we might reasonably expect that similar approaches will be powerful in other domains as well.

Behavioral scientists have long used models of human behaviour to build theories and make predictions. Sometimes these models consist of relatively simple assumptions about rational decision making under a specified utility function in constrained environments, enabling analytic treatment but often diverging considerably from real-world behaviour. The use of more realistic agent-based models to simulate human behaviour allows for more accurate predictions in more complicated scenarios where simple models fail (Farmer and Foley 2009). The use of modern deep reinforcement learning to train models of human behaviour and human preferences can be seen as the next step in this tradition (Tacchetti et al. 2025). Modeling human preferences is not just good for training language models. For example, in a paradigm similar to the behavioural cloning approach described above, recent work modeled human preferences over political opinions to find common ground in democratic deliberation (Bakker et al. 2022; Tessler et al. 2024). In these use cases, what matters most is that the clone or the preference model makes accurate predictions about human behaviour or preferences. That is what enables the downstream reinforcement learning to discover novel solutions. The clone or preference model need not in any way capture the underlying mechanisms of said behaviour.

2.4 General-purpose cognitive modeling

A grand goal of some psychology and neuroscience researchers is to build a comprehensive model of human cognition, capable of predicting and explaining many phenomena from a unified set of principles (Newell 1990; Anderson and Lebiere 2003). In this vision's traditional form, the model in question is the product of an integrated theory, emerging gradually from deep achievements of human understanding across the mind, brain and behavioral sciences. But the arrival of powerful AI tools offers a potential shortcut to this goal. Just as systems like AlphaFold have revolutionized protein

structure prediction and other difficult modeling problems in the natural sciences, so too might psychologists hope to develop models of the mind via black-box deep learning methods rather than painstaking human theory-building.

A large recent effort in this direction is the Centaur model, built from Meta's Llama 3.1-70B LLM finetuned with a large dataset of human choices from 160 psychology experiments (Binz, Akata, et al. 2025). Centaur's developers describe it as "the first real candidate for a unified model of human cognition", establishing that "it is possible to discover domain-general models of human cognition in a data-driven manner" (4).

The primary utility of Centaur is in predicting human choices in multi-option scenarios of various kinds (categorization tasks, probabilistic reasoning tests, gambling scenarios, multi-armed bandits, etc.), such choice sequences having made up the majority of its finetuning data. On average, Centaur was found to moderately outperform the base Llama model at predicting held-out experimental data; Llama in turn typically outperforms task-specific models of human behavior from the cognitive science literature, by a slightly larger margin.

Across all tasks, Centaur achieved an average pseudo-R^2 value of 0.50, where $R^2 = 0$ and $R^2 = 1$ respectively correspond to random guessing and perfect prediction. For novel behaviors outside the discrete-choice paradigm represented in its finetuning data, Centaur is often less successful: for instance, it achieves $R^2 = 0.18$ at predicting human inferences on a logical reasoning task (7).

It appears that, for some types of task, Centaur may owe its performance in part to "shortcut learning" (Geirhos et al. 2020), a common and often undesired training outcome in which "a model achieves high predictive performance by exploiting easily learnable features [of the training data], rather than capturing the underlying structure or intended rationale of the task" (Xie & Zhu 2025). For instance, Xie and Zhu found Centaur to outperform traditional cognitive models at predicting some types of human choices even when the descriptions of the relevant tasks were completely removed from its prompts. It appears likely in these cases that Centaur has learned to exploit temporal dependencies in the choice-sequence training data (e.g., the general tendency for subjects to repeat their own earlier choices) without attempting to model task-specific psychological processes.

Understanding the basis on which models like Centaur makes behavioural predictions is critical for assessing their appropriateness for different use cases. For example, without this understanding, a researcher piloting a new version of an experiment on Centaur might wrongly infer that modifying the

task instructions would elicit no changes in participants' behaviour, when in fact Centaur (unlike a typical human subject) was simply ignoring the instructions altogether.

While we don't wish to dwell exclusively on the details of any one project, Centaur illustrates some barriers that remain to be overcome for future exercises in AI-driven cognitive modeling, as well as some potentially intrinsic limitations of the genre.

One lesson is that base LLMs are already quite good at predicting (some types of) human behavior, so large improvements over this starting point will likely require high-quality datasets of considerable size. (Centaur's finetuning data comprised over ten million human choices from "many canonical studies" (3) but outperformed the base Llama model by a smaller margin than Llama's outperformance of leading task-specific models. The gap between Centaur and a frontier closed-source model would presumably be narrower yet.) Second, developing a truly general-purpose cognitive model—with strong predictive abilities not just in one behavioral regime, but for a wide range of phenomena in abnormal, cognitive, developmental, organizational, perceptual, social, and personality psychology—will be a high hurdle to clear, especially in light of the previous point about the sizes of the necessary datasets. Third, while Binz, Akata, et al. invoke the Newellian aim of "understand[ing] the human mind in its entirety" (2) as motivation for their work, it's often been noted that black-box machine learning methods leave much to be desired on this score (Räz and Beisbart 2024): a powerful AI model might tell us a great deal about *what* humans would do under various circumstances while leaving us none the wiser about how and why.[6]

How troubled should psychologists be by this last point? Some, like Bowers et al. (2025), are pessimistic about the value of predictive power without mechanistic understanding: "Even if [Centaur] did behave like a human… it is unclear what theoretical insights would have been gained. Successful prediction does not imply successful explanation, and as cognitive scientists the development of explanatory theories is our main goal" (2). Similarly, Xie and Zhu worry that tools like Centaur "[risk] undermining several key qualities that theorists seek in a cognitive model—namely, explanatory insight into cognitive mechanisms, reliable generalization, and interpretable or trustworthy parameters" (5).

A more moderate view might hold that accurate predictions are sometimes useful on their own, even when divorced from explanatory insight. But as Xie and Zhu show, *how* a model makes predictions

[6] Though see Duede (2023) for an optimistic perspective on the prospect of gaining scientific understanding from opaque deep-learning models.

may determine its suitability for a particular use case. Accuracy in a circumscribed testing regime is no guarantee of accuracy under different deployment conditions, and the functional organization of a predictively successful model need not match that of a human mind. Perhaps future progress on interpretability techniques will shine light on the inner workings of complex models, allowing for more rigorous and systematic assessments in this domain (Kästner and Crook 2024; Rai et al. 2024).

2.5 Replacing human subjects

The most ambitious and controversial AI surrogacy proposal to feature in recent discussions involves the wholesale replacement of human experimental subjects by machine proxies. The motivation for replacement in this sense is the idea, colorfully expressed by Dillion et al., that "[t]he human mind is what researchers seek to understand… [and the] 'minds' of language models are trained on vast amounts of human expression, so their expressions can indirectly capture millions of human minds" (599, Box 2). As illustration, Dillion et al. report an impressive correlation of 0.95 between judgments made by GPT-3.5 and average human participants in five moral classification studies (598, Box 1). Though they aver that "human participants are safe for now" (598), Dillion et al. suggest that LLM outputs can usefully supplement human experimental data, and could perhaps come to play a larger role in research once the statistical relationship between LLM and human behavior is better understood.

Other authors have countenanced stronger conclusions. Byun et al. (2023) suggests that "replacing humans with LLMs [in qualitative human-computer interaction research] may not be an unreasonable possibility. These models are able to generate logical and convincing themes and discussions… They are able to generate interesting research ideas, synthetic participant data, and reasonable questions for research artifacts" (8). Likewise, according to Grossman et al. (2023), "[w]ith the advent of advanced AI systems, the landscape of data collection in social sciences may shift. …LLMs may supplant human participants for data collection" (1108).

Unlike the other surrogacy proposals considered here, the prospect of legitimizing experimental research based solely on AI-generated data would require a sea change in scientific cultures and practices. We're not aware that any such study has yet been published in a mainstream psychology journal—at least, none which presents its findings as psychology per se, rather than an exploration of AI behavior and capabilities. Similarly, we're not aware that any mainstream journal has expressed openness to such work.[7]

[7] Perhaps the new Taylor & Francis journal *AI & Psychology* will be among the first. The journal has published no articles at the time of this writing, but according to its website it will cover "everything that can be included under robopsychology,

Whether such methods will or should gain wide acceptance depends on a number of factors. The specific research question at issue may be relevant in various ways, for one. Since AI systems are better at simulating some types of human behavior than others, artificial participant data will be more credible when it fits squarely into a well-understood regime of close correspondence. Conversely, some research setups might be inherently more tolerant of a certain amount of error in simulated data—say when the existence or directionality of an effect is more important than its precise size, when a study's purposes are exploratory or corroborative, or when a correction can be applied to counteract a known type of model bias.

On a different note, Agnew et al. (2024) argues that the embrace of machine-generated data stands in tension with certain humanistic aims of psychological research: notably, "the *representation* of participants' interests; participants' *inclusion* and empowerment in the development process; and the *understanding* that researchers otherwise develop through intersubjective engagement with participants" (1). We take it that the concerns here are partly ethical and partly epistemic. On the epistemic side, Agnew et al. maintain for instance that "the basis of psychological research and insight is not objective measurement, but intersubjective corroboration" (8), and on their view it's unclear whether human researchers will ever be able to access the relevant kind of shared perspective with AI partners. In the absence of such intersubjective scaffolding, AI outputs may be more on par as evidence with the behaviors of a poorly understood alien species than with the choices of human subjects.

This form of AI surrogacy raises many further methodological questions: about the psychological characteristics of AI models, model choice and reproducibility, the feasibility of simulating diverse groups and more. We discuss a number of these issues in section 3 below.

It's worth noting, finally, that a proliferation of cheap and fast research on silicon samples might require a rethinking of standard processes for disseminating knowledge. Even a fivefold increase (say) in the average researcher's productivity would put tremendous pressure on editors, reviewers, meeting organizers and other traditional gatekeepers[8], as well as on the ability of fellow scientists to keep up with new developments. Perhaps these new pressures could be managed only by further automation. At the end of this road, it would seem, lies a world of AI-generated papers assessed and summarized by

---

defined as the psychology of, for, and by robots, robotics, and AI" (https://www.tandfonline.com/journals/tpai20/about-this-journal).

[8] See Mollaki (2024) for discussion of the use of LLMs in peer review.

AI for the benefit of any remaining human researchers. The behavioral sciences may soon have to consider whether to embrace this vision, and how to avoid it if not.[9]

3 The peril: AI surrogacy's methodological minefield

In this second half of the chapter we move away from specific proposals to examine general issues raised by a variety of AI surrogacy methods. We consider streetlight effects, the distinctive psychology of LLMs, issues surrounding model choice, the prospects for simulating diverse human groups, and questions of explanation and understanding.

3.1 Streetlight effects and the shape of the research landscape

As noted above, AI surrogacy methods may offer significant advantages in speed, cost and ease over their traditional counterparts. These advantages will be appealing to many scientists, especially the large proportion who face pressure to maintain active research programs under binding resource constraints.[10] As AI-based work becomes more widely accepted, then, we should expect an increasing number of research programs, methods and outputs to shape themselves around the affordances of the available technology.

These dynamics raise the prospect of an encroaching AI "streetlight effect", wherein the character of a growing share of psychological research is determined not by a method-neutral commitment to the best ideas and techniques, but in large part by the capabilities, proclivities and biases of the AI systems driving research activity. This effect might manifest in various ways. If automated hypothesis generation becomes widespread, then new research will disproportionately target the kinds of ideas which AI tools are (perceived to be) best at generating. If specialized psychological models like Centaur (§2.4) gain ground as tools for testing or corroborating hypotheses, researchers will gain an incentive to favor questions in these models' wheelhouses. If the use of LLMs to simulate participant data is normalized, then LLM-friendly experiment designs will offer time-pressed scientists the path of least resistance. And so on.

[9] Bender et al. warn that, even if AI systems can greatly increase the rate at which research is produced and published, "[s]cience is not a factory, churning out widgets or statistical analyses wrapped in text. ...[W]e cannot equate papers and progress. Papers are but messages that we send one another to coordinate our collective quest for scientific understanding" (Binz et al. 2025, 8).

[10] E.g., see Lilienfeld (2017) and Almeida (2023) for observations about the deleterious effects of grant culture on the quality and diversity of psychological research, especially outside elite US and UK universities.

The streetlight worry is distinct from concerns about AI accuracy or reliability. Even if all machine-generated hypotheses were true, all ML models' predictions correct, and all silicon samples statistically identical to their human counterparts, overreliance on such tools might still bring about an objectionable distortion of the research landscape. For instance, it might turn out that applied psychology is more automation-friendly than basic theoretical research, or that behaviors closely related to language use are more amenable to study with LLM assistance. It would be regrettable if basic research and non-linguistic cognition were deprioritized for these reasons alone.

Of course, the current status quo is hardly free of streetlight influences. Scientists must always design research around the available tools, and it's no surprise that bad results can ensue when new technology is adopted with more enthusiasm than care. (Bennett et al. (2009) illustrates how, in the heady early days of the functional neuroimaging gold rush, many fMRI studies committed a basic statistical error which called their interpretation of scan data into question.) Even among alluring technological novelties, however, AI tools may pose particular risks as hammers that promise to effortlessly pound in many stubborn types of nail. Large apparent upsides and an appearance of all-purpose competence make limitations easier to ignore, especially when the boundaries of the latter are poorly understood and blame for oversights is easily deflected.

We can also compare AI streetlight effects to the blind spots associated with human minds, since our own contingent capabilities, proclivities and biases shape research in important ways. While both types of streetlight require vigilance, the human status quo may be more manageable in several respects. First, current AI systems may be strictly narrower in relevant knowledge and capabilities than typical human researchers (and are certainly narrower than all human researchers collectively). To the extent that this is true, we should expect AI-sourced research to span fewer domains with less depth than comparable human work. (AI's speed, cost and volume advantages offer some compensation here, but only some.) Second, many human foibles are well enough understood that researchers can anticipate and to some extent correct for them: say, by constructing funnel plots to check for publication bias, or administering sugar pills to control groups to account for placebo effects. We're comparatively ignorant about the quirks of AI systems and hence less prepared to notice and counteract the epistemic distortions they might introduce. Third, human researchers can incorporate new information into their knowledge bases almost immediately, whereas the costly training runs needed to update large AI models are often few and far between. So streetlights associated with AI knowledge and capabilities may be fixed in place for relatively long stretches. Fourth, human cognitive idiosyncrasies are stable over long timescales, while the proliferation of models and open-ended possibilities of finetuning make it difficult to establish general rules of AI psychology. The reader can likely think of further

considerations to add to ours. Continued progress on AI systems will presumably narrow some of these gaps, but others may be with us for the foreseeable future.

Peterson (2025) offers an interesting formal model of a related streetlight effect. Starting from the observation that LLMs and similar models tend to generate outputs close to the statistical center of their training data, Peterson considers how overreliance on AI as an information source may gradually bring about "knowledge collapse", "neglecting the long tails of knowledge and creating a degenerately narrow perspective over generations" (2). The paper models the decay of public knowledge in the presence of a cheap yet epistemically truncated AI information source. The model predicts that decreasing the cost of AI-generated content relative to the human baseline dramatically widens the distance between public knowledge and the truth: under a 50% discount, for instance, the public's beliefs end up 3.2 times further from the truth than in the no-discount condition after nine simulated generations. While Peterson's model is meant to apply to society as a whole, it's plausible that widespread reliance on cheap AI data might induce a similar knowledge-collapse effect in individual scientific communities.

3.2 The psychology of AI models

Some research applications of AI seek to take advantage of deep learning models' distinctive capabilities and elevated performance. Meanwhile, other surrogacy methods are motivated by the perception that LLMs and other AI systems are—or can be made to deliver—good approximations of a psychologically ordinary, average, representative, or otherwise statistically and normatively appropriate human subject. This conviction is most clearly on display in proposals to replace human participants by silicon samples (§2.5). But it's also operative, for example, in some uses of AI for item piloting and instrument development (§2.2), where models are meant to simulate how typical subjects might engage with these materials.

These proposals raise a number of methodological questions. Just how well can AI models serve as human proxies? Where do their strengths and weaknesses lie? What notion of normality or averageness best captures the relationship between model outputs and human behavior, and what sort of epistemic work can this notion do for us?

One line of research has examined LLMs' ability to simulate specific individuals, given appropriately customized prompting or finetuning. (Such models are sometimes known as digital doppelgängers or twins; D'Alessandro et al. (2025) proposes a typology of these systems.) Results to date here have been mixed. Petrov et al. (2024) found that GPT-4 generally "produced skewed distributions in the

direction of a 'desirable' [Big Five Inventory] trait" (10) and was therefore "less capable of representing individuals who tend to be more disagreeable, unconscientious, introverted, emotionally stable, and not open to experience" (16). More recently, using transcriptions of long-form audio interviews as GPT-4o prompts, Park et al. (2024) obtained normalized correlations of 85%, 80% and 66% with human respondents' own replications of their General Social Survey responses, Big Five personality scores and behavioral-economics game choices, respectively.

Of course, researchers are more often interested in the statistics of human populations than in data on specific individuals. So a more pressing question is how AI outputs compare to people in aggregate.[11]

One often encounters claims in the literature about the *averageness* of LLM behavior. Zhicheng Lin writes, for instance, that "[i]n many perceptual, linguistic, cognitive, and moral reasoning tasks, LLMs generate responses that closely capture what average people perceive, say, think, or do" (2024, 3). Since *average* carries many possible meanings, it's important to be clear about what claims of this sort can be justifiably made. On the one hand, researchers comparing human and LLM behavior have found strong correlations or similar mean values in some instances. But the distributions of these datasets may nevertheless differ strikingly. So it would be a mistake for researchers (or public consumers of research) to assume or suggest without specific evidence that the statistics of human and LLM responses are generally similar.

It's often observed, for instance, that LLM outputs in a given domain follow a sharply peaked, short-tailed unimodal or bimodal distribution, while human responses exhibit much greater variability. Abdurahman et al. (2024) demonstrates the potentially far-reaching nature of this effect, comparing GPT-3.5 to human responses on six major personality, cognition and emotion scales. "Across these psychological constructs, we consistently found that ChatGPT responses generally showed significantly less variance across all measures"; in particular, "GPT-3.5's variance [on the Moral Foundations Questionnaire-2] was 43–121 times smaller than human data… even when using parameter settings for maximum variability in the generated responses" (3).

Similarly, in an investigation of GPT-3.5's ability to replicate fourteen studies from the Many Labs 2 project, P. S. Park, Schoenegger, and Zhu (2024) were unable to analyze almost half the target studies

[11] In principle, one can imagine the first of these issues bearing on the second—researchers could, for instance, develop digital-twin simulations of large numbers of individual humans, and then experiment directly on these simulations and do statistics on the resulting data. Given the effort required to train a decent twin, however, it would seem more efficient simply to work with human participants from the start. Perhaps these economics will change once a critical mass of high-quality digital twins becomes available to researchers.

because "different runs of GPT-3.5 in our sample responded with zero or near-zero variation for either a dependent variable or condition variable question, in stark contrast to the significant variation shown by the corresponding human subjects" (5757).[12] Meanwhile, Aher et al. (2023) documented a "hyper-accuracy distortion effect" whereby larger LLMs (including GPT-4) asked to simulate human subjects are more likely to give identical and inhumanly accurate answers to general-knowledge questions.

In some cases, notable distributional differences coexist with large correlation coefficients or similar means. Dillion et al. (2023) calls GPT-3.5 "extremely well aligned with human moral judgments" (597), reporting a correlation of $r = 0.95$. Nevertheless, GPT-3.5 delivered extreme ratings (between 3.75 and 4, or between -3.75 and -4) in around 25% of 464 moral evaluation scenarios, while the mean human rating never reached either extreme.[13] P. Wang et al. (2024)'s study of personality in LLMs furnishes another example. In this case, the authors found that "the mean values in the descriptive statistics were the only dimension where the [human and LLM] datasets were similar. Other than that, both the standard deviation in the descriptive statistics and the psychometric performance, such as model fit and structural validity, were unsatisfactory" (41). Bisbee et al. (2024) observed a similar pattern with respect to GPT-3.5's answers to American National Election Study questions. Claims about the averageness of LLM behavior may therefore be misleading, and deserve to be carefully qualified. The precise nature of the statistical relationship between human and AI outputs will matter greatly to the appropriateness of many surrogacy methods.

Issues of averageness aside, what do we know about the psychological profiles of LLMs? Both similarities and dissimilarities to human baselines have been noted. One line of research has examined the extent to which language models reproduce classic findings in psychology and behavioral science. Here, Shaki et al. (2023) found evidence for priming and several other effects in GPT-3, while Park et al. (2024) successfully replicated only three out of eight well-confirmed phenomena from Many Labs 2 studies with GPT-3.5. Cui et al. (2024) undertook a much larger study, testing GPT-4 on 154 psychological experiments. The model was found to replicate 76% of main effects and 47% of interaction effects documented in the original experiments. Notably, however, "only 19.44% of GPT-4's replicated confidence intervals contain the original effect sizes, with the majority of replicated effect sizes exceeding the 95% confidence interval of the original studies and showing a 71.6% rate of unexpected significant results where the original studies reported null findings" (2).

[12] In this work, P. S. Park, Schoenegger, and Zhu used GPT-3.5's default intermediate temperature setting, described by OpenAI as eliciting answers with significant randomness.
[13] Data available at https://nikett.github.io/gpt-as-participant/.

Other work has looked directly at the psychometrics of AI models. As Lin (2024) notes, one should expect to find significant discrepancies with the typical human case here, since frontier LLMs are exceptionally skilled and versatile language users by nature. Thus it's no great surprise that models like GPT-4 outdo average humans in "detecting and interpreting irony, recognizing indirect requests or hints in conversation, analogical reasoning tasks, and probabilistic reasoning tasks like the Linda/Bill problems and the bat-and-ball problem" (Lin 2024, 8), as well as in assessing and providing guidance about social situations (Mittelstädt et al. 2024). It's often been observed, moreover, that LLM personalities and values reflect the idiosyncratic WEIRD[14] perspectives overrepresented in internet training data (Abdurahman et al. 2024; Crockett and Messeri 2023).

LLM psychology exhibits other notable features which can't be straightforwardly explained by the above factors. For instance, LLMs are in general more agreeable and conscientious, and substantially less neurotic, than the average American (Li et al. 2024). This profile is likely due in part to RLHF finetuning for helpfulness and other desirable traits. Indeed, Huang et al. (2024) tested a jailbroken version of GPT-4 in order to probe its "intrinsic characteristics" (15), finding significant psychometric differences including much lower agreeableness and conscientiousness, higher psychoticism and lower empathy compared to the default model. (For more on finetuning and its implications for model choice, see §3.3 below.)

Researchers have also found considerable psychometric variation between models, and even among versions of a given model from a single developer: for instance, Llama 3-8b and Llama 3-70b score 3.56 and 4.89 respectively on agreeableness (the human average is 3.78). Variation has likewise been found in the *stability* of LLM personality traits. Some models (such as Llama 3-8b) display humanlike levels of stability, while others (such as GPT-4) are notably unstable on both Big Five and Dark Triad traits. Finally, on the EmoBench tests of emotion understanding and application, Li et al. found LLM performance generally "not satisfactory, with all accuracies below 65%" (7). This is perhaps a surprising contrast with findings like that of Mittelstädt et al. on LLMs' superhuman social competence and Huang et al. (2024) on GPT-4's high emotional intelligence.

There's more to say about each of these topics. Broadly speaking, however, results like the above make it clear that researchers must think carefully, and perhaps test extensively, before putting language models to work in human roles. The fluency, intelligence, apparent humanlikeness and alleged averageness of current models may disguise many psychometric oddities. Moreover, the differences

[14] That is, western, educated, industrialized, rich and democratic.

among models themselves and between models and humans need not pattern together in intuitive ways.

3.3 Choosing, comparing and customizing models

Individual LLMs (e.g., GPT-4, Claude 3.5, Llama 3.1) are commonly thought of as singular, static entities. Behind the scenes, however, these models are made up of several machine learning systems trained on a variety of objectives and datasets.

Language model pretraining usually involves training a transformer-based architecture on objectives like next-word or masked-word prediction using massive text datasets. This *base LLM* may undergo supervised finetuning on human datasets collected expressly to capture what makes a good AI chatbot virtual assistant. These datasets may take the form of human demonstrations of appropriate prompt-response pairs, or human judgements related to the harmlessness, helpfulness, and honesty of responses (Bai et al. 2022).

The next step is to train a *preference model* (or reward model) that takes in a prompt-response pair and outputs a scalar value that captures how highly a human would rank the response. This requires an additional dataset of human preferences, which is typically collected by presenting crowdworkers with pairs of candidate responses generated by the supervised finetuned (SFT) model and asking them to choose the best of the two. Combining several of these pairwise comparisons (via methods like Elo rating) produces a ranking over candidate responses that can be used to train the preference model. The preference model may be a version of the base LLM that is finetuned in a supervised manner on the human preference rankings.

Finally, the SFT model can be further finetuned, now with a reinforcement learning algorithm like proximal policy optimization (PPO), to produce responses that the reward model scores highly. Alternative recipes skip training a preference model and finetune directly on the human preference datasets (Rafailov et al. 2023). This final *finetuned model* is what psychological researchers will often have readiest access to, often without knowing exactly which datasets were used for training and finetuning and typically without access to the preference model. This lack of transparency is a clear challenge to the usefulness of LLMs as surrogates and the interpretability of their outputs.

The additional finetuning steps described above, which have been essential ingredients for producing the impressive capacities of modern large language models, steer language models towards particular kinds of language interaction that are deemed appropriate for an AI chatbot. The preference models

behind state-of-the-art models are typically not publicly available, and thus the values they encode cannot be easily audited. Recent work has found considerable heterogeneity in the values reflected in different preference models and systematic biases in their scores, some of which deviate significantly from human values as assessed with independent measures (Christian et al. MS).

The objectives these models are trained on are not designed to make the outputs similar to those of an average human, since average humans wouldn't make ideal chatbots or search engines. In some ways, ideal LLM behavior may be better than that of an average human, in order to meet helpfulness objectives. In other ways they will be unlike the average human in their avoidance of certain topics for safety and harmlessness. How much this matters for instances of AI surrogacy in psychological research may depend on the use case.

As noted in §3.2, some use cases of LLMs as surrogates in psychological research rely on the assumption that LLM behaviour is human-like, i.e., that LLM behaviour can be interpreted as an approximation of average, normal or representative human behaviour. Even if the state-of-the-art, closed-source finetuned LLMs released by prominent AI companies display behaviour that is empirically humanlike in some ways, their behaviour has also been steered in unknown ways, the influence of which may be difficult to detect. An approach like that of Centaur (§2.4), where a pretrained model is finetuned on large datasets of human behaviour in psychology experiments, may in some respects be more appropriate (although we note again that, on average, Centaur's finetuning only resulted in marginal prediction improvements relative to the LLama 3.1-70b base model).

It may not be possible to develop a single model that can faithfully replicate any desired type of human behavior. Psychologists wishing to use LLMs in the relevant ways may need to be prepared to finetune their own models with their particular use case in mind. Still, of course, finetuning on targeted datasets is not guaranteed to produce accurate predictions of human behaviour. We may see an increase in research on prompt engineering to elicit human-like responses, or on transfer algorithms to adapt a general-purpose base LLM to a target domain of human behaviour. When performing a cost-benefit analysis to determine whether AI surrogacy is worthwhile, practitioners will need to consider what investment (e.g., the collection or aggregation of new datasets, computational resources, additional training time, upskilling) will be required to achieve the necessary degree of humanlikeness in their particular setting.

Other use cases are less concerned with the verisimilitude of LLM behaviour. In fact, part of the benefit for uses like hypothesis generation and item development may be the superior ability of LLMs to

synthesize, summarize, and revise large bodies of text. However, which version of which LLM is used will still be of consequence in these settings. LLMs have been shown to display various kinds of linguistic erasure, systematically avoiding terminology related to topics found in negatively rated examples during finetuning (e.g., those related to human sexuality or racial, ethnic, and gender identity; Sap et al. 2019; Park et al. 2018; Dixon et al. 2018; Christian et al. MS). Were this linguistic erasure to infect the hypothesis generation process, it could limit the space of hypotheses considered in undesirable ways.

On the other hand, using a pretrained LLM directly is likely to produce falsehoods and recreate the prejudices present in its training data, so *not* finetuning is not a viable option either. Finetuning an LLM with research users specifically in mind may offer some advantages, but attempts to date have not been notably successful—for instance, Meta's Galactica model, trained on scientific texts and intended as a tool for students and researchers (Taylor et al. 2022), had to be taken offline after only three days due to its biased and incorrect outputs. Galactica also displayed linguistic erasure, responding with "Sorry, your query didn't pass our content filters. Try again and keep in mind this is a scientific language model" when queried about racism or AIDS (Heaven 2022).

3.4 Simulating diverse groups and specific traits

In §3.2 we considered the extent to which AI systems can be viewed as proxies for average or normal human psychology. That issue is most salient for situations where AI outputs are meant to stand in for arbitrary individual humans, or to produce summary statistics similar to a typical human population. For other surrogacy purposes—in particular, for certain kinds of silicon sampling methods—researchers will instead wish to use AI to simulate specific (and perhaps highly non-average) groups or traits. This section discusses what's known about current AI capabilities on this score and what challenges such simulation methods may face.

To start with, there may be principled reasons for concern about the ability of LLMs to faithfully simulate members of particular identity groups. A. Wang, Morgenstern, and Dickerson (2025) suggests, for instance, that both *misportrayal* and *flattening* are likely. 'Misportrayal' refers to the conflation of facts about identity group characteristics with mistaken beliefs about these characteristics from outsider perspectives. Since the two types of information may be poorly distinguished in training data, the authors argue, it will be technically difficult to develop misportrayal-proof models. 'Flattening', meanwhile, involves the inappropriate reduction of a group's characteristics to those that are most common, salient or widely known. Since pretrained LLMs are designed to favor statistically likely outputs, their group simulations may gravitate too much toward typicality, underrepresenting

in-group differences. A. Wang et al. and Cheng et al. (2023) present evidence for both phenomena. For instance, Cheng et al. find that identity group personas generated by GPT-4 contain many more stereotype-associated words than do similar self-descriptions written by humans from the same groups.

Researchers interested in using LLMs as proxies for group members of any sort face choices about how to elicit the desired simulated behavior. The literature has explored a variety of paradigms, including prompting with keywords, demographic categories, more elaborate biographical data, numerical psychometric scores, survey responses, first- vs. second-person framings, and so on. None of these methods seems to have yet emerged as clearly best, and no systematic theory yet exists about the relationship between prompt styles and simulation outcomes. It may therefore be unclear whether poor results from a given simulation study should be attributed to an inherent limitation of the models involved or merely a suboptimal prompting strategy.

One way to address this sort of concern is by using a variety of strategies in tandem. Santurkar et al. (2023) takes this approach, measuring how well LLM outputs can be "steered" toward the opinions of various demographic groups. The authors use three types of prompt, choosing only those with the best steering results in each case. They find that this method is somewhat but not extremely effective: "In most cases, we see the representativeness of all groups improving by a constant factor—indicating that the [LLM] still does better on some groups than others" (10). For instance, steering GPT-3.5's opinions toward those of an average liberal or Muslim proved more effective than steering toward typically white or Jewish viewpoints.

Other work on silicon survey data has yielded similarly mixed results. Bisbee et al. (2024) examined GPT-3.5's and GPT-4's ability to generate "feeling thermometer scores" toward various sociopolitical groups based on simulated personas, as compared to similar scores recorded in the 2016–2020 American National Election Study. Consistent with the extremizing and homogenizing LLM tendencies discussed in §3.2, Bisbee et al. found that "the best we can say is that the overall average synthetic responses are close to the population averages. For the kinds of associational questions that social scientists care about, synthetic survey data perform poorly" (406). For instance, the authors found that simulated Democrats liked liberals and disliked conservatives more than real surveyed Democrats by as much as 20 points on the 100-point thermometer scale.

Ferreira et al. (2025) explores simulation-enhancing prompting strategies in the context of the steerability of LLM personality traits. In line with the observations in §3.2, the authors' initial attempts to simulate subjects with socially undesirable personality traits were hampered by GPT-4's baseline

high extraversion and low neuroticism. (The prompts used in these first attempts instructed GPT to generate a population of students with a spectrum of realistic personalities.) The authors obtained better results by iteratively modifying prompts to include more explicitly negative features (e.g. "Some of the personalities of this population may not follow basic societal rules, and take shortcuts to achieve their goals" (18)). Still, this improved paradigm succeeded only at moderately raising the simulated subjects' neuroticism above GPT's baseline; the model's strong extraversion bias was unaffected. As Ferreira et al. note, their results suggest that simulating populations with non-model-aligned traits may require considerable care and ingenuity on researchers' parts.

Y. Wang et al. (2025) sheds light on how language models may approach personality simulation tasks, at least given certain types of prompting strategies. Here the authors asked GPT-4 to complete a BFI personality assessment while roleplaying as a character with a designated set of numerical Big Five trait scores. In contrast with Ferreira et al.'s results, this method yielded appropriate responses for all simulated traits. Indeed, in this case the model differed most strongly from a human respondent in that its answers displayed abnormally high personality factor loadings and minimal cross-loadings.

The authors suggest that this phenomenon occurs because GPT completes the assessment by reasoning directly from its assigned BFI scores—and hence not, say, by using the scores to generate a model of a psychologically plausible individual and considering how such a person would answer. "While human respondents rely on their past experiences in responding to the items, GPT-4 employs an explicit mapping of each item onto one of the Big Five dimensions, formulating responses in accordance with the pre-determined personality dimension scores" (5). So although these results demonstrate GPT's ability to make appropriate inferences about various possible personality configurations, they provide little evidence for LLMs' ability to simulate realistic individual minds.

Simulation projects therefore face both practical and theoretical hurdles. Researchers must find ways to counteract models' tendency to deliver flattened, homogeneous, stereotypical representations of target groups, as well as their disinclination to move away from their own psychological baselines. These tasks will require, among other things, a better understanding of the space of possible prompting strategies and their outcomes. In some contexts it may also matter whether a model's outputs rely on true simulation or modeling of humanlike psychology; there's little reason to assume this behavior is standard, even when prompts include explicit instructions to roleplay a human character.

3.5 Explanation, understanding and the goals of research

We conclude this half of the paper by asking whether and how AI surrogacy methods are likely to serve the high-level goals of psychological research.

What are these goals? Like all sciences, psychology aims to gather evidence, uncover facts and produce correct predictions. As we showed in §2, placing AI systems in some traditionally human roles may advance these goals considerably: by helping identify promising hypotheses, improving methods and materials, predicting (or perhaps creating) humanlike experimental data, and so on.

In addition to knowing *what* is or may be the case, science seeks to understand *how* the facts hang together and *why* things are as they are. Indeed, some have argued that understanding and explanation are the primary goals of scientific inquiry (Potochnik 2015; de Regt 2020) or the fundamental measure of scientific progress (Dellsén 2021; McCoy 2022). It's perhaps less clear whether AI surrogacy methods will provide a straightforward benefit here.

Thinking clearly about this last question requires some precision about the concepts involved. Although scientific explanation and understanding are contested topics with large literatures, we can point to some notable recent trends.

According to the influential view known as *mechanism*, for instance, the sciences most characteristically "explain a phenomenon by describing the mechanism underlying it, revealing its internal causal structure" (Craver et al. 2024). A mechanism in this sense is roughly a structured system of parts which interact to perform a function or produce an effect. An appropriate description of a mechanism, moreover, will be intelligible to its audience and will therefore serve as a source of *explanatory understanding*.[15] The best-known work in the mechanist tradition has focused on explanation in biology and neuroscience, though many insist on the importance of mechanisms in psychology more broadly (Bechtel 2009; Piccinini and Craver 2011).[16]

[15] Explanatory understanding (as in "Jane understands why color vision works poorly in low light") is often distinguished from *objectual* understanding (as in "Jane understands behavioral economics" or "Jane understands the neuron action potential"). The latter is sometimes thought of as the state of knowing many interconnected facts about a target topic (Kelp 2015); others have argued that psychological proficiencies (involving heuristics, intuition, schemas and the like) are further factors (D'Alessandro 2023; D'Alessandro and Stevens 2024; Inglis and Mejía-Ramos 2021). We focus on explanatory understanding here for simplicity.

[16] Whether psychology's highly abstract, model-based *functional* explanations should be viewed as fundamentally mechanistic has been a subject of debate (Weiskopf 2011; Shapiro 2017). But most of our comments here should apply to both types of case.

If something like the mechanist account is correct, then merely documenting, predicting or modeling the observed phenomena doesn't suffice for explanatory understanding. Indeed, as Thompson (2021) notes, it's insufficient even to have an intelligible theory, to statistically explain all the variance in a dataset, or to identify a coarse-grained causal relationship. What's required is knowledge of a specific kind: knowledge about the workings of the organized worldly structures from which the phenomena arise. On this view, explanatory progress in psychology will presumably be tied to continued high-quality experimentation within and across scales, from neuroscience to behavioral science, in search of mechanisms, their components, and the larger causal systems within which they function.

The AI methods we've discussed may help advance this goal in some respects. Machine learning algorithms will plausibly be useful for identifying causal structure in neural data, and the novel experimental designs discussed in §2.3 may expedite the discovery of some fundamental facts about cognition. (As Duede (2023) points out, "deep learning models can be used quite effectively in science, not just for pragmatic ends but also for genuine discovery and deeper theoretical understanding. This can be accomplished when [these models] are used as *guides* for exploring promising avenues of pursuit in the context of discovery" (1097).)

In other ways, however, increasing psychology's reliance on AI surrogates need not promote explanatory progress. Black-box predictive models can be highly accurate while telling us little about underlying mechanisms. And silicon samples which are good enough for plausibility checks and rough summary statistics may not support the detailed experimental probing necessary for drawing causal inferences. From the viewpoint of advancing fundamental understanding, then, there's as yet no clear technological substitute in sight for the rigorous study of human populations, behaviors and brains.

It's possible to imagine a future in which AI systems themselves gain whatever capacities are required to possess genuine scientific understanding in their own right. Botvinick and Gershman ask whether we should be content for such systems to autonomously carry forward the project of science, proposing questions and discovering explanations with minimal human involvement (except perhaps to reap the downstream benefits of AI findings). They suggest we should resist such an outcome: "We cannot cede understanding to artificial systems. We should insist on human understanding remaining a core goal of science", at least insofar as possible given the complexity of the phenomena (Binz, Alaniz, et al. 2025, 7). Marelli et al. agree: "The impact of LLMs on the future practice of science cannot be fully predicted, but science is a humanistic and human enterprise and must remain so" (Binz, Alaniz, et al. 2025, 9).

We're inclined to agree with these sentiments, but more philosophical work remains to be done. We can imagine a spectrum of scenarios between fully human-centered and fully AI-driven futures for psychology. Human scientists might propose research questions, for instance, while AI systems design appropriate mechanism-seeking experiments, interpret their results, and produce thoughtfully designed digests to maximize human understanding. Alternatively, machines may take a more assertive role in suggesting hypotheses and research directions aligned with actual or foreseeable human interests. Are either of these scenarios compatible with a humanistic vision of science in which our species' understanding remains a core goal? If AI delivers on its most ambitious promises, answering these questions may require careful thought in the coming years, and implementing our chosen answers will demand community-wide coordination. We hope to see these tasks handled with the wisdom and skill they require.

4 Conclusion

AI surrogates of any kind are only as good as the data used to train them. These data come ultimately from humans, whether in the form of text in LLM pretraining datasets, measurements of human preferences over prompt-response pairs, bespoke experimental datasets of human behaviour used to finetune models, or confirmatory human data collected to test predictions from silicon samples. In this sense, AI surrogacy merely shifts humans into new roles rather than eliminating psychology's reliance on them.

Why bother with AI surrogacy, then? Why not analyze the human text, preferences, judgements, and decisions directly? When do the benefits—the speed, versatility, low cost and impressive power of deep learning-based tools—outweigh the costs of substituting opaque and imperfect proxies for the real phenomena of interest?

Throughout this chapter, we've raised several questions that practitioners should consider when wading through this cost-benefit analysis. Before beginning to use an AI surrogate, we urge researchers to define the criteria that would make an AI system suitable as a surrogate in their particular use case, and to select or build models accordingly rather than blindly using off-the-shelf models.

For instance, does a given type of model output need to be human-like in specific ways? If so, is it enough for the AI and human samples to exhibit similar means or a large correlation coefficient, or must the underlying distributions agree in finer-grained ways? Does the model output need to capture behavioural patterns of particular subgroups? To what extent can specialized prompting overcome models' tendency to misrepresent and flatten group characteristics? Does it matter for a particular

application whether a model's outputs result from modeling relevant psychological mechanisms, or is mere prediction sufficient? Is the AI behaviour reproducible and transparent, or is it the product of proprietary training recipes that are outside of researchers' control and may be changed at any time?

Sometimes navigating this cost-benefit analysis may mean choosing a slightly weaker open-source model over the latest closed frontier model. Sometimes it may involve evaluating several models with an eye to specific performance desiderata prior to use. Sometimes it may involve finetuning a model on a task of interest or collecting large datasets to train a model from scratch. These activities may involve skills and facilities that are not common in psychology research labs. If AI surrogacy methods continue to gain popularity in psychological research (not to mention other uses of AI and computational models more broadly), research labs will need to invest in computational resources and training programs to equip researchers with relevant skills. Shortcuts here are likely to lead to lower-quality results.

Many concerns about AI surrogacy are related to the accuracy, robustness, representativeness, and reproducibility of AI behaviour. Setting all those concerns aside, however, questions still remain. For example, even if an AI system can generate and evaluate hypotheses in some sense "better" than their human counterparts, is this a job that *should* be offloaded to an AI system? Are we willing to accept the epistemic consequences of letting non-human entities with whom we don't share experiential perspectives or cultural sensibilities decide which research programs to prioritize?

The diversity of human values and viewpoints, and their amenability to criticism and revision, are integral to scientific progress. Likewise for the hard-won explanatory insights gained through basic research of the highest standards. If we want a psychological science that continues to serve our deepest interests, we must be wary of the temptation to trade the difficult quest for humanistic understanding for an expedient facsimile which merely looks close enough, on average.

5 References

Abdurahman, Suhaib, Mohammad Atari, Farzan Karimi-Malekabadi, et al. 2024. "Perils and Opportunities in Using Large Language Models in Psychological Research." *PNAS Nexus* 3 (7): pgae245. https://doi.org/10.1093/pnasnexus/pgae245.

Agnew, William, A. Stevie Bergman, Jennifer Chien, et al. 2024. "The Illusion of Artificial Inclusion." *Proceedings of the CHI Conference on Human Factors in Computing Systems*, May 11, 1–12. https://doi.org/10.1145/3613904.3642703.

Aher, Gati V., Rosa I. Arriaga, and Adam Tauman Kalai. 2023. "Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies." *Proceedings of the 40th International Conference on Machine Learning*, July 3, 337–71. https://proceedings.mlr.press/v202/aher23a.html.

Almeida, Jorge. 2023. "Underfunding Basic Psychological Science Because of the Primacy of the Here and Now: A Scientific Conundrum." *Perspectives on Psychological Science* 18 (2): 527–30. https://doi.org/10.1177/17456916221105213.

Anderson, John R., and Christian Lebiere. 2003. "The Newell Test for a Theory of Cognition." *Behavioral and Brain Sciences* 26 (5): 587–601. https://doi.org/10.1017/S0140525X0300013X.

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. "Out of One, Many: Using Language Models to Simulate Human Samples." *Political Analysis* 31 (3): 337–51. https://doi.org/10.1017/pan.2023.2.

Bai, Yuntao, Andy Jones, Kamal Ndousse, et al. 2022. "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback." arXiv:2204.05862. Preprint, arXiv, April 12. https://doi.org/10.48550/arXiv.2204.05862.

Bakker, Michiel, Martin Chadwick, Hannah Sheahan, et al. 2022. "Fine-Tuning Language Models to Find Agreement among Humans with Diverse Preferences." *Advances in Neural Information Processing Systems* 35 (December): 38176–89.

Banker, Sachin, Promothesh Chatterjee, Himanshu Mishra, and Arul Mishra. 2024. "Machine-Assisted Social Psychology Hypothesis Generation." *American Psychologist* 79 (6): 789–97. https://doi.org/10.31234/osf.io/kv6f7.

Bechtel, William. 2009. "Looking down, around, and up: Mechanistic Explanation in Psychology." *Philosophical Psychology* 22 (5): 543–64. https://doi.org/10.1080/09515080903238948.

Beghetto, Ronald A., Wendy Ross, Maciej Karwowski, and Vlad P. Glăveanu. 2025. "Partnering with AI for Instrument Development: Possibilities and Pitfalls." *New Ideas in Psychology* 76 (January): 101121. https://doi.org/10.1016/j.newideapsych.2024.101121.

Berger, Jonah. 2024. "Machines, Psychology, and Hypothesis Generation: Commentary on Banker et al. (2024)." *American Psychologist* 79 (6): 798–99. https://doi.org/10.1037/amp0001258.

Binz, Marcel, Elif Akata, Matthias Bethge, et al. 2025. "A Foundation Model to Predict and Capture Human Cognition." *Nature*, July 2, 1–8. https://doi.org/10.1038/s41586-025-09215-4.

Binz, Marcel, Stephan Alaniz, Adina Roskies, et al. 2025. "How Should the Advancement of Large Language Models Affect the Practice of Science?" *Proceedings of the National Academy of Sciences* 122 (5): e2401227121. https://doi.org/10.1073/pnas.2401227121.

Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models." *Political Analysis* 32 (4): 401–16. https://doi.org/10.1017/pan.2024.5.

Bowers, Jeffrey, Guillermo Puebla, Sushrut Thorat, Konstantinos Tsetsos, and Casimir Ludwig. 2025. "Centaur: A Model without a Theory." Preprint, OSF, July 7. https://doi.org/10.31234/osf.io/v9w37_v3.

Byun, Courtni, Piper Vasicek, and Kevin Seppi. 2023. "Dispensing with Humans in Human-Computer Interaction Research." *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, April 19, 1–26. https://doi.org/10.1145/3544549.3582749.

Chai, Miaosen, Emily Herron, Erick Cervantes, and Tirthankar Ghosal. 2024. "Exploring Scientific Hypothesis Generation with Mamba." In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, edited by Lotem Peled-Cohen, Nitay Calderon, Shir Lissak, and Roi Reichart. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.nlp4science-1.17.

Chen, Dian, Ying Liu, Yiting Guo, and Yulin Zhang. 2024. "The Revolution of Generative Artificial Intelligence in Psychology: The Interweaving of Behavior, Consciousness, and Ethics." *Acta*

Psychologica 251 (November): 104593. https://doi.org/10.1016/j.actpsy.2024.104593.

Cheng, Myra, Esin Durmus, and Dan Jurafsky. 2023. "Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.84.

Chenneville, Tiffany, Brianna Duncan, and Gabriella Silva. 2024. "More Questions than Answers: Ethical Considerations at the Intersection of Psychology and Generative Artificial Intelligence." *Translational Issues in Psychological Science* (US) 10 (2): 162–78. https://doi.org/10.1037/tps0000400.

Christian, Brian, Hannah Rose Kirk, Jessica A. F. Thompson, Christopher Summerfield, and Tsvetomira Dumbalska. 2025. "Reward Model Interpretability Via Optimal and Pessimal Tokens." Paper presented at FAccT 2025. *ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3715275.3732068.

Craver, Carl, James Tabery, and Phyllis Illari. 2024. "Mechanisms in Science." In *The Stanford Encyclopedia of Philosophy*, Fall 2024, edited by Edward N. Zalta and Uri Nodelman. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2024/entries/science-mechanisms/.

Crockett, M. J., and Lisa Messeri. 2023. "Should Large Language Models Replace Human Participants?" Preprint. https://doi.org/10.31234/osf.io/4zdx9.

Cronbach, Lee J., and Paul E. Meehl. 1955. "Construct Validity in Psychological Tests." *Psychological Bulletin* (US) 52 (4): 281–302. https://doi.org/10.1037/h0040957.

Cui, Ziyan, Ning Li, and Huaikang Zhou. 2024. "Can AI Replace Human Subjects? A Large-Scale Replication of Psychological Experiments with LLMs." *CoRR* abs/2409.00128. https://doi.org/10.48550/ARXIV.2409.00128.

D'Alessandro, William. 2023. "Unrealistic Models in Mathematics." *Philosophers' Imprint* 23 (28): 0. https://doi.org/10.3998/phimp.1712.

D'Alessandro, William, Trenton W. Ford, and Michael Yankoski. 2025. "I Contain Multitudes: A Typology of Digital Doppelgängers." *The American Journal of Bioethics*, February 1. world. https://www.tandfonline.com/doi/abs/10.1080/15265161.2024.2441762.

D'Alessandro, William, and Irma Stevens. 2024. "Mature Intuition and Mathematical Understanding." *The Journal of Mathematical Behavior* 76 (December): 101203. https://doi.org/10.1016/j.jmathb.2024.101203.

Dellsén, Finnur. 2021. "Understanding Scientific Progress: The Noetic Account." *Synthese* 199 (3–4): 11249–78. https://doi.org/10.1007/s11229-021-03289-z.

Dixon, Lucas, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. "Measuring and Mitigating Unintended Bias in Text Classification." *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA), AIES '18, December 27, 67–73. https://doi.org/10.1145/3278721.3278729.

Duede, Eamon. 2023. "Deep Learning Opacity in Scientific Discovery." *Philosophy of Science* 90 (5): 1089–99. https://doi.org/10.1017/psa.2023.8.

Farmer, J. Doyne, and Duncan Foley. 2009. "The Economy Needs Agent-Based Modelling." *Nature* 460 (7256): 685–86. https://doi.org/10.1038/460685a.

Grossmann, Igor, Matthew Feinberg, Dawn C. Parker, Nicholas A. Christakis, Philip E. Tetlock, and William A. Cunningham. 2023. "AI and the Transformation of Social Science Research." *Science* 380 (6650): 1108–9. https://doi.org/10.1126/science.adi1778.

Harding, Jacqueline, William D'Alessandro, N. G. Laskowski, and Robert Long. 2023. "AI Language Models Cannot Replace Human Research Participants." *AI & SOCIETY*, ahead of print, July 21. https://doi.org/10.1007/s00146-023-01725-x.

Heaven, Will Douglas. 2022. "Why Meta's Latest Large Language Model Survived Only Three Days Online." MIT Technology Review, November 18. https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/.

Hewitt, Luke, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer. 2024. "Predicting Results of Social Science Experiments Using Large Language Models." Preprint. https://docsend.com.

Hibben, Kristen Cibelli, and Julie de Jong. 2016. "Cognitive Interviewing." In *Guidelines for Best Practice in Cross-Cultural Surveys*. Survey Research Center, Institute for Social Research, University of Michigan. https://ccsg.isr.umich.edu/chapters/pretesting/cognitive-interviewing/.

Huang, Jen-tse, Wenxuan Wang, Eric John Li, et al. 2024. "Who Is ChatGPT? Benchmarking LLMs' Psychological Portrayal Using PsychoBench." arXiv:2310.01386. Preprint, arXiv, January 22. https://doi.org/10.48550/arXiv.2310.01386.

Inglis, Matthew, and Juan Pablo Mejía-Ramos. 2021. "Functional Explanation in Mathematics." *Synthese* 198 (26): 6369–92. https://doi.org/10.1007/s11229-019-02234-5.

Kästner, Lena, and Barnaby Crook. 2024. "Explaining AI through Mechanistic Interpretability." *European Journal for Philosophy of Science* 14 (4): 52. https://doi.org/10.1007/s13194-024-00614-4.

Kelp, Christoph. 2015. "Understanding Phenomena." *Synthese* 192 (12): 3799–816. https://doi.org/10.1007/s11229-014-0616-x.

Koskinen, Inkeri. 2024. "We Have No Satisfactory Social Epistemology of AI-Based Science." *Social Epistemology* 38 (4): 458–75. https://doi.org/10.1080/02691728.2023.2286253.

Koster, Raphael, Jan Balaguer, Andrea Tacchetti, et al. 2022. "Human-Centred Mechanism Design with Democratic AI." *Nature Human Behaviour* 6 (10): 1398–407. https://doi.org/10.1038/s41562-022-01383-x.

Koster, Raphael, Miruna Pîslar, Andrea Tacchetti, et al. 2025. "Deep Reinforcement Learning Can Promote Sustainable Human Behaviour in a Common-Pool Resource Problem." *Nature Communications* 16 (1): 2824. https://doi.org/10.1038/s41467-025-58043-7.

Kuhail, Mohammad Amin, Nazik Alturki, Justin Thomas, Amal K. Alkhalifa, and Amal Alshardan. 2024. "Human-Human vs Human-AI Therapy: An Empirical Study." *International Journal of Human–Computer Interaction* 0: 1–12. https://doi.org/10.1080/10447318.2024.2385001.

Li, Yuan, Yue Huang, Hongyi Wang, Xiangliang Zhang, James Zou, and Lichao Sun. 2024. "Quantifying AI Psychology: A Psychometrics Benchmark for Large Language Models." arXiv:2406.17675. Version 1. Preprint, arXiv, June 25. https://doi.org/10.48550/arXiv.2406.17675.

Lilienfeld, Scott O. 2017. "Psychology's Replication Crisis and the Grant Culture: Righting the Ship." *Perspectives on Psychological Science* 12 (4): 660–64. https://doi.org/10.1177/1745691616687745.

Lin, Zhicheng. 2024. "Large Language Models as Linguistic Simulators and Cognitive Models in Human Research." SSRN Scholarly Paper No. 4974107. Social Science Research Network, January 29. https://doi.org/10.2139/ssrn.4974107.

Lippert, Steffen, Anna Dreber, Magnus Johannesson, et al. 2024. "Can Large Language Models Help Predict Results from a Complex Behavioural Science Study?" *Royal Society Open Science* 11 (9): 240682. https://doi.org/10.1098/rsos.240682.

Liu, Yiren, Si Chen, Haocong Cheng, et al. 2024. "How AI Processing Delays Foster Creativity: Exploring

Research Question Co-Creation with an LLM-Based Agent." *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA), CHI '24, May 11, 1–25. https://doi.org/10.1145/3613904.3642698.

Lu, Chris, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. "The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery." arXiv:2408.06292. Preprint, arXiv, September 1. https://doi.org/10.48550/arXiv.2408.06292.

Ludwig, Jens, and Sendhil Mullainathan. 2024. "Machine Learning as a Tool for Hypothesis Generation*." *The Quarterly Journal of Economics* 139 (2): 751–827. https://doi.org/10.1093/qje/qjad055.

Luo, Xiaoliang, Akilles Rechardt, Guangzhi Sun, et al. 2024. "Large Language Models Surpass Human Experts in Predicting Neuroscience Results." *Nature Human Behaviour*, November 27, 1–11. https://doi.org/10.1038/s41562-024-02046-9.

Manning, Benjamin, Kehang Zhu, and John Horton. 2024. *Automated Social Science: Language Models as Scientist and Subjects*. No. W32381. National Bureau of Economic Research. https://doi.org/10.3386/w32381.

McCoy, C. D. 2022. "Understanding the Progress of Science." In *Scientific Understanding and Representation: Modeling in the Physical Sciences*, edited by Insa Lawler, Kareem Khalifa, and Elay Shech. Routledge.

McKee, Kevin R., Andrea Tacchetti, Michiel A. Bakker, et al. 2023. "Scaffolding Cooperation in Human Groups with Deep Reinforcement Learning." *Nature Human Behaviour* 7 (10): 1787–96. https://doi.org/10.1038/s41562-023-01686-7.

Messeri, Lisa, and M. J. Crockett. 2024. "Artificial Intelligence and Illusions of Understanding in Scientific Research." *Nature* 627 (8002): 8002. https://doi.org/10.1038/s41586-024-07146-0.

Mittelstädt, Justin M., Julia Maier, Panja Goerke, Frank Zinn, and Michael Hermes. 2024. "Large Language Models Can Outperform Humans in Social Situational Judgments." *Scientific Reports* 14 (1): 27449. https://doi.org/10.1038/s41598-024-79048-0.

Mollaki, Vasiliki. 2024. "Death of a Reviewer or Death of Peer Review Integrity? The Challenges of Using AI Tools in Peer Reviewing and the Need to Go beyond Publishing Policies." *Research Ethics* 20 (2): 239–50. https://doi.org/10.1177/17470161231224552.

Newell, Allen. 1973. "You Can't Play 20 Questions with Nature and Win." *Visual Information Processing*, 283–308. https://doi.org/10.1021/bi047419c.

Newell, Allen. 1990. *Unified Theories of Cognition*. Unified Theories of Cognition. Harvard University Press.

O'Brien, Thomas, Joel Stremmel, Léo Pio-Lopez, Patrick McMillen, Cody Rasmussen-Ivey, and Michael Levin. 2024. "Machine Learning for Hypothesis Generation in Biology and Medicine: Exploring the Latent Space of Neuroscience and Developmental Bioelectricity." *Digital Discovery* 3 (2): 249–63. https://doi.org/10.1039/D3DD00185G.

Ouyang, Long, Jeffrey Wu, Xu Jiang, et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." *Advances in Neural Information Processing Systems* 35 (December): 27730–44.

Park, Ji Ho, Jamin Shin, and Pascale Fung. 2018. "Reducing Gender Bias in Abusive Language Detection." arXiv:1808.07231. Preprint, arXiv, August 22. https://doi.org/10.48550/arXiv.1808.07231.

Park, Joon Sung, Carolyn Q. Zou, Aaron Shaw, et al. 2024. "Generative Agent Simulations of 1,000 People." arXiv:2411.10109. Preprint, arXiv, November 15. https://doi.org/10.48550/arXiv.2411.10109.

Park, Peter S., Philipp Schoenegger, and Chongyang Zhu. 2024. "Diminished Diversity-of-Thought in a Standard Large Language Model." *Behavior Research Methods* 56 (6): 5754–70. https://doi.org/10.3758/s13428-023-02307-x.

Peterson, Andrew J. 2025. "AI and the Problem of Knowledge Collapse." *AI & SOCIETY*, ahead of print,

January 19. https://doi.org/10.1007/s00146-024-02173-x.

Petrov, Nikolay B., Gregory Serapio-García, and Jason Rentfrow. 2024. "Limited Ability of LLMs to Simulate Human Psychological Behaviours: A Psychometric Analysis." arXiv:2405.07248. Preprint, arXiv, May 12. https://doi.org/10.48550/arXiv.2405.07248.

Piccinini, Gualtiero, and Carl Craver. 2011. "Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches." *Synthese* 183 (3): 283–311. https://doi.org/10.1007/s11229-011-9898-4.

Porter, Brian, and Edouard Machery. 2024. "AI-Generated Poetry Is Indistinguishable from Human-Written Poetry and Is Rated More Favorably." *Scientific Reports* 14 (1): 26133. https://doi.org/10.1038/s41598-024-76900-1.

Potochnik, Angela. 2015. "The Diverse Aims of Science." *Studies in History and Philosophy of Science Part A* 53: 71–80. https://doi.org/10.1016/j.shpsa.2015.05.008.

Rafailov, Rafael, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. "Direct Preference Optimization: Your Language Model Is Secretly a Reward Model." *Advances in Neural Information Processing Systems* 36 (December): 53728–41.

Rai, Daking, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2024. "A Practical Review of Mechanistic Interpretability for Transformer-Based Language Models." arXiv:2407.02646. Preprint, arXiv, July 2. https://doi.org/10.48550/arXiv.2407.02646.

Räz, Tim, and Claus Beisbart. 2024. "The Importance of Understanding Deep Learning." *Erkenntnis* 89 (5): 1823–40. https://doi.org/10.1007/s10670-022-00605-y.

Regt, Henk W. de. 2020. "Understanding, Values, and the Aims of Science." *Philosophy of Science* 87 (5): 921–32. https://doi.org/10.1086/710520.

Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. "Whose Opinions Do Language Models Reflect?" arXiv:2303.17548. Preprint, arXiv, March 30. https://doi.org/10.48550/arXiv.2303.17548.

Sap, Maarten, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. "The Risk of Racial Bias in Hate Speech Detection." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, edited by Anna Korhonen, David Traum, and Lluís Màrquez. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1163.

Schlegel, Katja, Nils R. Sommer, and Marcello Mortillaro. 2025. "Large Language Models Are Proficient in Solving and Creating Emotional Intelligence Tests." *Communications Psychology* 3 (1): 80. https://doi.org/10.1038/s44271-025-00258-x.

Shaki, Jonathan, Sarit Kraus, and Michael Wooldridge. 2023. "Cognitive Effects in Large Language Models: 26th European Conference on Artificial Intelligence, ECAI 2023." *ECAI 2023 - 26th European Conference on Artificial Intelligence, Including 12th Conference on Prestigious Applications of Intelligent Systems, PAIS 2023 - Proceedings*, Frontiers in Artificial Intelligence and Applications, September 28, 2105–12. https://doi.org/10.3233/FAIA230505.

Shang, Jiaqi, and Will Xiao. 2023. "AI, Robot Neuroscientist: Reimagining Hypothesis Generation." Paper presented at NeurIPS 2023 AI for Science Workshop. October 28. https://openreview.net/forum?id=CllNd4XWVF.

Shapiro, Lawrence A. 2017. "Mechanism or Bust? Explanation in Psychology." *The British Journal for the Philosophy of Science* 68 (4): 1037–59. https://doi.org/10.1093/bjps/axv062.

Tacchetti, Andrea, Raphael Koster, Jan Balaguer, et al. 2025. "Deep Mechanism Design: Learning Social and Economic Policies for Human Benefit." *PNAS*.

Taylor, Ross, Marcin Kardas, Guillem Cucurull, et al. 2022. "Galactica: A Large Language Model for Science."

arXiv:2211.09085. Preprint, arXiv, November 16. https://doi.org/10.48550/arXiv.2211.09085.

Tessler, Michael Henry, Michiel A. Bakker, Daniel Jarrett, et al. 2024. "AI Can Help Humans Find Common Ground in Democratic Deliberation." *Science* 386 (6719): eadq2852. https://doi.org/10.1126/science.adq2852.

Thompson, Jessica A. F. 2021. "Forms of Explanation and Understanding for Neuroscience and Artificial Intelligence." *Journal of Neurophysiology* 126: 1860–74. https://doi.org/10.1152/jn.00195.2021.

Tong, Song, Kai Mao, Zhen Huang, Yukun Zhao, and Kaiping Peng. 2024. "Automating Psychological Hypothesis Generation with AI: When Large Language Models Meet Causal Graph." *Humanities and Social Sciences Communications* 11 (1): 1–14. https://doi.org/10.1057/s41599-024-03407-5.

Wang, Angelina, Jamie Morgenstern, and John P. Dickerson. 2025. "Large Language Models That Replace Human Participants Can Harmfully Misportray and Flatten Identity Groups." *Nature Machine Intelligence* 7 (3): 400–411. https://doi.org/10.1038/s42256-025-00986-z.

Wang, Pengda, Huiqi Zou, Zihan Yan, et al. 2024. "Not Yet: Large Language Models Cannot Replace Human Respondents for Psychometric Research." Preprint, OSF, September 23. https://doi.org/10.31219/osf.io/rwy9b.

Weiskopf, Daniel A. 2011. "Models and Mechanisms in Psychological Explanation." *Synthese* 183 (3): 313–38. https://doi.org/10.1007/s11229-011-9958-9.

Xu, Ruoxi, Yingfei Sun, Mengjie Ren, et al. 2024. "AI for Social Science and Social Science of AI: A Survey." *Information Processing & Management* 61 (3): 103665. https://doi.org/10.1016/j.ipm.2024.103665.

Zhou, Yangqiaoyu, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024. "Hypothesis Generation with Large Language Models." *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, 117–39. https://doi.org/10.18653/v1/2024.nlp4science-1.10.

Ziegler, Daniel M., Nisan Stiennon, Jeffrey Wu, et al. 2020. "Fine-Tuning Language Models from Human Preferences." arXiv:1909.08593. Preprint, arXiv, January 8. https://doi.org/10.48550/arXiv.1909.08593.