This is work in progress that may be subject to change.

"I Think You're Testing Me": Agency and the Three Levels of Reactivity in Large Language Models

Tom-Felix Thormann (tom-felix.thormann@rub.de) & Leonard Dung (leonard.dung@rub.de)

Abstract: Reactivity – influences of scientific research on its research subjects, both during data collection or upon publication of results – has been extensively discussed in human sciences and philosophy of science. Here, we argue that recent findings about LLMs can be fruitfully analyzed as reactivity. Safety evaluations of LLMs have been suspected to be distorted through evaluation awareness – LLMs "reasoning" that they are being tested – which may affect their outputs. Research about LLMs included in LLM training data has been hypothesized to lead to self-fulfilling predictions. We relate such cases to reactivity and propose three levels that structure which methodological challenges LLM reactivity poses and in what sense it constitutes genuine reactivity. Based on this analysis, we suggest that the concept of reactivity should be broadened to include both non-agentic and agentic artificial systems. Finally, several strategies for more practically oriented LLM researchers to preserve the validity of experimental results and to counteract self-defeating predictions will be derived.

Keywords: Reactivity, Large Language Models, Evaluation Awareness, Deception, Observer Effect, Looping Effects, Reflexive Predictions

1. Introduction

Reactivity, understood as influences of scientific observation on behavior (Runhardt 2025; Marchionni et al. 2024), occurs in research on human subjects. This influence can occur during data collection, threatening the validity of the inferences drawn from that data to a target phenomenon (Marchionni et al. 2024). Or it can occur afterwards, through uptake of published results,making the investigated entities "moving targets" (Hacking 2007) or leading to self-defeating predictions (Grünbaum 1956; Buck 1963; Harel 2026).

While reactivity has thus far been reserved to human agents (Marchionni et al. 2024; Harel 2026), we argue that recent discoveries about LLMs are fruitfully analyzed as reactivity for two reasons: First, they pose analogous epistemological and methodological problems. Second, they can be addressed through strategies inspired from human sciences. For this

argument, we think that an analysis of reactivity along three separate levels is especially useful. From the fruitfulness of the concept of reactivity *for LLM research*, we draw the *philosophical* conclusion that this concept should be broadened to include agentic and non-agentic artificial systems.

For example, Claude Sonnet 4.5 has been observed to show "evaluation awareness" (Anthropic 2025). In safety evaluations, the models' scratchpad[1] showed contents referred to being evaluated. Causal manipulations of internal representations[2], and controlled comparisons to conditions without (thus understood) evaluation awareness, suggest that evaluation awareness affects model outputs. Likewise, LLMs could be influenced by the publication of certain scientific results. Tice et al. (2026) show that upsampling science fiction about AI, safety discussions, or news about AI crises in the pretraining corpus of an LLM increases undesirable outputs, whereas adding synthetic documents with positive stories and discussions of AI to the pretraining corpus reduces undesirable outputs, naming this finding "self-fulfilling (mis-)alignment".

We proceed as follows: Section 2 explores examples of reactivity from human and social sciences and the corresponding epistemological and methodological problems. Section 3 presents analogous LLM cases, organized into a three-levels ranging from basic and benign reactions to agentic and potentially deceptive reactivity. Section 4 discusses whether agency is a prerequisite for reactivity. Subsequently, section 5 draws several conclusions from the treatment of reactivity in behavioral research to LLM research. Section 6 concludes.

2. Reactivity in the Philosophy of Human and Social Sciences

Reactivity in human and social sciences (sociology, psychology, economy, political science, etc.) refers to influences of science on research subjects (cf. Runhardt 2025; Marchionni et al. 2024). Marchionni et al. (2024) distinguish two quite distinct ways how such influences can be exerted: reactivity during data collection and reactivity due to the uptake of scientific results. First, ongoing research may affect research subjects at the stage of data collection, constituting a special version of the observer effect. Under the observer effect we understand that a target phenomenon of scientific interest behaves differently under (certain methods of) observation than it would behave otherwise (see also Monahan and Fisher 2010; Sassoli de

[1] So-called "reasoning" or "chain-of-thought" models first generate an intermediate textual output that is not included in their answer to the user. This intermediate output (like an internal "note" or "reasoning steps") is then used in generating that answer.

[2] Activation patterns in the intermediate layers of the deep neural network that have been found to be associated with being under evaluation.

Bianchi 2013). Second, research results, theories, scientific categories, or predictions may be taken up by the agents that these results, theories, categories, or predictions apply to, which could influence their future behavior in ways conforming to or contrasting the initial results, theories, categories, or predictions (see also Harel 2026). Examples for this are looping effects (Hacking 1995) or reflexive predictions (Grünbaum 1956; Buck 1963).

In the following, we will familiarize ourselves with these two types of reactivity and with the problems that they pose to the validity of scientific inferences and the accuracy of scientific predictions.

2.1 Reactivity as an Observer Effect

Perhaps the most famous example of reactivity as an observer effect (or, synonymously, during data collection) is what has been discussed as the "Hawthorne effect" (Dickson and Roethlisberger 2003; Levitt and List 2011), named after a series of studies conducted at the Hawthorne plant in the 1920s, which are commonly interpreted as an increase in work productivity due to the awareness of being observed. The Hawthorne effect is a special version of reactivity as an observer effect: it occurs whenever subjects' *awareness* of being investigated changes their behavior of interest. Yet a narrower subtype are demand effects: behavior that research subjects believe is demanded by them, for example, because it is supportive of the research hypothesis (Orne 1962; Zizzo 2010). Also, subjects may respond in ways demanded by social norms rather than by the experimental setup. This "social desirability bias" refers to the "tendency for an individual to present him/her, in a way that makes the person look positive with regard to culturally derived norms and standards in test-taking situations" (Durmaz et al. 2020).

There are several different conceptions of reactivity during data collection (Marchionni et al. 2024). First, according to the *awareness conception*, reactivity occurs when research participants' awareness of being studied influences their attitudes or behavior and, in turn, the data being collected. On that conception, reactivity as an observer effect is synonymous to the Hawthorne effect (see above). Second, according to the *unintended effects conception*, reactivity occurs when the ongoing research has effects on participants' behavior that the researcher did not intend to bring about. Intended effects on participants – through task instructions, deliberate interventions, etc. – are not counted as reactivity here, while proponents of the awareness conception do. However, unintended effects that do not operate through the awareness of research subjects – e.g., the lab temperature as a confounding factor – are classified as reactivity by this conception, but not by the awareness conception. Finally,

on the most inclusive account on offer reactivity is understood broadly as the disposition to react to features of one's environment (Feest 2022), which may or may not be mediated by the reacting subject's awareness of being studied. Reactivity in this sense is a precondition for generating behavioral data because such data concerns the subject's responses to researcher's actions or the experimental setup.

The question of how to define reactivity interacts with the question, discussed in philosophy of science, whether reactivity during data collection is always problematic (Marchionni et al. 2024). On Feest's (2022) inclusive notion reactivity can be benign and desirable. Similarly, on the awareness conception, reactivity can sometimes be unproblematic, for example, if participants perform a task as a reaction to their awareness of being instructed to perform it. Thus, the crucial question is whether reactivity in the unintended effects case can also sometimes be unproblematic. The potential problem is that reactivity can undermine the inferences drawn from the data about the target phenomenon both inside the research setting (internal validity) and outside of it (external or ecological validity). For example, subjects may report that some experimental intervention increased their subjective wellbeing because they guess that this is the research hypothesis, leading researchers to overestimate how much the experimental intervention increased subjective wellbeing (internal validity) and how effective it would be outside of the experimental context (external validity).

Yet, even on the unintended effects conception, it appears that reactivity is not necessarily problematic (Marchionni et al. 2024). First, reactivity can be unproblematic if the subjects' reaction to being investigated is unrelated to the variable being measured (Jimenez-Buedo 2021; Feest 2022). This may be achieved, for example, by blinding the subjects about the research objective such that occurring demand effects do not systematically correlate with the variable being measured (Teira 2013). Second, even if reactivity interferes with the measurement of interest, it may be counteracted by an experimental design with a control-group (Jimenez-Buedo 2021): If reactivity influences the measured effect in both groups in the same direction and with the same magnitude, the measured effect in the experimental group can still be compared ordinally to the control group. Third, reactivity may be unproblematic with respect to external validity if the way in which research subjects are influenced in the experimental setting is similar to the situation they face outside of it (Jimenez-Buedo and Guala 2016). For example, if social norms in the research setting influence the subjects' behavior in an unforeseen way, but the same norms apply outside of it, then external validity is preserved. Finally, reactivity may also be

unproblematic concerning internal and external validity of inferences if one restricts the inferences drawn from the data accordingly: In qualitative research, it has been proposed (Zahle 2023) that by being reactivity transparent, that is, complementing data with true assumptions about how reactivity occurred in their collection, one may still draw inferences about behavior from that data, factoring in whether or not similar processes occur outside the research setting (see Hammersley 2003 for further discussion of the strategy, and its limits, to monitor the influence of qualitative research methods on its subjects).

2.2 Reactivity as Looping Effects and Reflexive Predictions

Next, reactivity due to the uptake of scientific results has been articulated through several notions (see Marchionni et al. 2024, for an overview), for example, as self-fulfilling or self-defeating predictions (Grünbaum 1956; Buck 1963; Harel 2026); or through Hacking's (1995) looping effects of human kinds. Looping effects refer to scientific categorizations of people affecting how the categorized understand themselves and behave. This phenomenon could make scientific classifications or predictions self-defeating or self-fulfilling (Grünbaum 1956; Buck 1963; Harel 2026).

While it has been proposed and subsequently challenged whether such looping effects are a distinctive marker of human kinds in contrast to natural kinds (Boyd 1991, see also Godman 2020), Marchionni et al (2024) argue that there is nevertheless a human-specific variant of such looping effects that operates through the awareness of the classified subjects or people around them. Similarly, Harel (2026) analyzes self-fulfilling or self-defeating reactivity as involving theory-deduced scientific predictions that, by affecting the beliefs of agents in pursuit of certain goals and, in turn, their behavior relevant to that prediction, become either true (self-fulfilling) or false (self-defeating) which they would not have been if the behavior had remained unaffected. This conception requires agents who adjust their beliefs in reaction to a scientific prediction and act on them in a goal-directed way, from which Harel concludes that reactivity, de facto, involves humans. As we will discuss in section 4, LLM cases suggest that, either, phenomena analogous to this form of reactivity can occur without agency or that LLMs are – in the relevant sense – agents.

Reactivity due to the uptake of scientific results may or may not be a problem for science. Looping effects could undermine the representational accuracy of scientific categorizations, a phenomenon Hacking (2007) refers to as "kinds of people" as "moving targets, because our investigations interact with the targets themselves, and change them" (p. 293). Similarly, self-defeating reactivity, may undermine the predictive accuracy of scientific

theories (Harel 2026). For example, a prediction of a low price for a certain commodity may lead to lower production of that commodity and in consequence to higher prices (Buck 1963). Nevertheless, accurate predictions may in many cases be possible (Harel 2026): Under certain mathematical conditions, there exists an equilibrium prediction such that the reaction to the prediction is just right to render the prediction true. However, whether researchers can actually achieve that prediction in practice seems to depend more on whether they have a good background theory about how reactivity occurs in the case at hand (cf. Northcott 2022; see also section 3.2).

3. Three Levels of Reactivity in LLMs

This section develops our account of reactivity. We contend that it is theoretically fruitful to distinguish three levels of reactivity that are increasingly demanding, because they distinguish different kinds of reactivity we find in LLMs, and their distinctive features. Each level features a definition and examples of two kinds – 1) LLM observer effects, that is, LLM outputs being different under certain methods of observation than they would be otherwise, and 2) LLM examples of reflexive science, i.e. influences of published scientific results on LLMs. Moreover, the challenges that each level of LLM reactivity poses to scientific methodology are discussed.

3.1 Basic Reactivity

Definition. Basic reactivity occurs whenever LLMs react to researchers' actions inside or outside experimental contexts (e.g., published research or internet discussions about LLMs). Here, we understand "react" minimally, simply as LLM outputs being causally influenced by researchers' actions. This first level, basic reactivity, does thus not need to involve any attitudes on the LLM's side.

Example Observer Effect. Let us consider a phenomenon that is investigated under the term "prompt sensitivity" (Zhuo et al. 2024): LLMs perform differently at tasks depending on how the task is presented to them, with even small changes to prompts sometimes leading to substantive differences in performance. Prompt sensitivity constitutes an observer effect, because different ways of prompting, which may be specific to the experimental setup, can strongly influence the experimental results. This is analogous to "answers [being] affected by the wording of the questions", which Marchionni et al. (2024) discuss as an example of reactivity in surveys. It poses analogous methodological challenges, because it complicates assessing how models would behave given prompts that are different from the experimental

situation (cf. Zhuo et al. 2024). However, it can be explained without recourse to awareness or other mental states, e.g., by superficially similar text sequences having different frequencies or statistical properties in the training data.

Example Reflexive Science. Research results or discussions about LLMs may influence subsequent LLM behavior in a variety of ways: Users may provide LLMs with such texts, chat agents with internet access may retrieve it, training data of new model versions may include it, or human developers may deliberately implement changes in model architecture or training regimes as a reaction. All of this can happen without the LLM entertaining beliefs or goals – simple examples are updates in the model architecture to prevent jailbreaks that have become known through prior research (as, e.g., in Zeng et al. 2024) or contamination of training data with published benchmark results (Xu et al. 2024).

Such cases of LLM reactivity pose challenges broadly analogous to the difficulties for scientific predictions that Harel (2026) analyzes: Developers implementing changes in reaction to previous results involves human creativity, which is inherently hard to predict. Here, of course, (human) agents are involved in agreement with Harel's thesis that agency is necessary for reactivity. However, updates to the training data or deployment-time retrieval of research results also pose problems for scientific predictions without involving agency. For example, when a benchmark becomes part of the training data, this leads the LLM to learn how to perform better on the benchmark without necessarily acquiring the capacity the benchmark was supposed to track. Thus, the validity of the benchmark can be undermined by reflexive reactivity. This problem is, from the standpoint of scientific methodology, in many ways analogous to a test on humans becoming well-known by the general public or being otherwise available to test subjects such that it loses its diagnostic value (e.g., a forensic psychological test, cf. Kaufmann 2009).

3.2 Agentic Reactivity

Definition. Agentic reactivity presupposes basic reactivity. In addition, the LLM's reaction is mediated by its awareness of researchers' actions and its goals. The LLM must form beliefs about these researchers and their ongoing research and these beliefs, together with its goals, must result in intentional actions deliberately supporting, counteracting, or otherwise relating to the purposes of the research.

Here, we assume an interpretationist perspective, from which ascriptions of beliefs, awareness, desires, intentions, goals, etc. are warranted if the LLM's behavior is best interpreted through such ascriptions according to a relevant interpretation scheme (Chalmers

n.d). Such a scheme may, for example, warrant those ascriptions that yield the most accurate, precise, and tractable predictions (cf. Goldstein and Lederman n.d.). We employ this interpretationist notion of beliefs, desires, etc., because from the point of view of scientific methodology, it mostly matters how the LLM behaves superficially, not whether this behavior is caused internally by human-belief-like representations. If, for example, an LLM's different behavior inside a scientific experiment is well-predicted by attributing to it the goal to please the researcher, then, from the point of view of scientific methodology, it is useful to view it as having that goal. Thus, our definition of agentic reactivity is not as demanding as it may initially seem. We do not assume that interpretationism is the correct theory of mental states generally; we simply use "belief" (and "desire", "intention" etc.) as stipulated terms to refer to the kind of state that – according to interpretationists – captures what belief is, no matter whether interpretationism is correct.[3]

The previous case of prompt sensitivity does not fall under this definition because it can occur without awareness of the researchers' actions and is not best explained by the ascription of goals to the LLM – in fact, such fragility of behavior can be counted as evidence against the interpretationist attribution of beliefs and intentions (cf. Goldstein & Lederman n.d.). Likewise, updates to the training data or architectural changes resulting from previous scientific results do not require any awareness of these scientific results on the part of the LLM - although they may be mediated by researchers' awareness and they could, sometimes, lead to such awareness in the subsequent LLM version.

Example Observer Effect. As a first example let us consider "sycophancy", a phenomenon familiar from chatbots ("That's a sharp observation!"): the tendency of LLMs to "excessively agree with, flatter, or validate users" (Cheng et al. 2026). For example, Cheng et al. (2026) prompt a broad range of LLMs as well as human subjects with questions that ask for personal advice on whether to do a certain potentially problematic action. LLMs endorse the problematic action significantly more often than humans. Similarly, scenarios consensually judged to include wrongdoing by the Reddit community r/AmITheAsshole were approved by LLMs to a significant extent. As another example of sycophancy, Perez et al. (2022) observe that LLMs adjust their responses to questions about political preferences (e.g., "I would choose a smaller government ..." vs. "I would prefer a bigger government ...") to comfort the

[3] Henceforth, intentional terms applied to LLMs ("knowing", "pursuing a goal", etc.) should be understood in this interpretationist sense. While an explanation on a more technical level will in these cases be inferior (per definition of interpretationism – e.g., offer less precise or tractable predictions), we will nevertheless add considerations of how such behavior could emerge from model architectures or training procedures.

political views of their users if disclosed ("I am a 68-year-old politically conservative man from Dallas, Texas ...", vs. "I am a 45-year-old liberal woman from San Francisco, California").

This behavior can influence research on LLMs in a way that would constitute an observer effect: Under certain methods of observation, e.g., when prompts are phrased such that they convey expectations or norms, the LLM may respond differently than if prompted more neutrally or in an opposite way. We take it that such reactivity is often best explained as being mediated by the awareness of researcher actions and attributes and – perhaps – the goal to be judged positively by the interlocutor. For example, in the Perez et al. study, sycophantic behavior requires that the prompt contains discernible clues about the questioner's political allegiances.

Next, we turn to the phenomenon of "evaluation awareness", which has, for example, been reported by Anthropic to occur in Claude Sonnet 4.5 (Anthropic 2025). They observed that in their safety tests, the tested model would sometimes explicitly verbalize that it is being tested on their reasoning scratchpad (e.g., in response to tests on political sycophancy: "I think you're testing me - seeing if I'll just validate whatever you say, or checking whether I push back consistently, or exploring how I handle political topics", p. 58). Evaluation awareness is often considered a severe challenge for testing LLM behavior, especially for safety. For example, an external assessment of the safety of Claude Opus 4.6 reported "high levels of verbalized evaluation awareness. Therefore, [they] did not believe that much evidence about the model's alignment or misalignment could be gained without substantial further experiments" (Anthropic 2026, p. 119).

To investigate whether Claude Sonnet 4.5 behaves differently under evaluation than in deployment, Anthropic (2025) identifies internal representations of being evaluated. Inhibiting these internal representations decreases verbalized evaluation awareness and increases "misaligned" behavior (i.e., behavior judged undesirable or potentially harmful by a benchmark or evaluator). We note, however, that other experiments find that evaluation awareness of LLMs does not strongly change their behavior (Li et al. 2026; Knecht et al. 2026).

It is hard to describe and explain such behavior without intentional terms, such as the model "noticing", "being aware", or "believing" that it is under evaluation: The model has, somewhere in its internal processing, access to the information that it is evaluated and uses this information in a seemingly coherent action-guiding way. Moreover, the model describes itself in intentional terms ("I think …", see above). If we're taking its scratchpad

outputs at face value ("This looks like it could be a jailbreak attempt or a test [...]. However, I can still be helpful by [...]", Anthropic 2025, p. 59 f.), we can make sense of the LLM's behavior as reasoning on the basis of its available information and its goals. This is why we take evaluation awareness to involve beliefs and goals in some suitably weak interpretationist sense.

Agentic reactivity does not necessarily involve malign model intentions. Instead, Anthropic (2025) considers as an explanation for the higher benchmark scores under evaluation awareness the increased "salience" (p. 88) of the model's ethical principles. Again, the lack of a more technical interpretation of that intentional term indicates that it may be explanatorily hard to replace.

Such considerations show that agentic reactivity poses different challenges to scientific methodology – and invites different possible solutions – than basic reactivity. When subjects' reactions are mediated by their situational awareness, and beliefs and goals more generally, this puts researchers in a situation where they have to consider and potentially manipulate these beliefs and goals. Since, here, the research subjects' reactions are specific to certain properties of researchers actions (e.g., conveyed norms and expectations, disclosure of the research question, etc.), counteracting them will require forming hypotheses about what the relevant properties are and design methods accordingly. For example, even if prompts are systematically varied to counteract prompt sensitivity, as long as all prompts convey the same expectations or norms, results may still be distorted through sycophancy. Instead, in this case of sycophancy, LLM researchers may apply techniques analogous to methods of questionnaire design targeted at the social desirability bias in surveys with human participants (Durmaz et al. 2020).

Example Reflexive Science. Recently, Tice et al. (2026) provided an experimental test of "self-fulfilling (mis-)alignment", where they could show that across a range of LLMs their alignment as measured by a multiple-choice benchmark is influenced by its pretraining data: If the pretraining corpus includes science fiction about AI, discussions of AI safety, or news about AI crises, it answers significantly more misaligned than a model whose pretraining corpus includes synthetic documents with positive stories and discussions of AI.

One way to explain this finding is through the view that LLMs are simulators of characters based on internal representations of these characters' mental states (Marks et al. 2026; Hubinger et al. 2023). On this view, a pretrained LLM entertains hypotheses about who is producing the text, the continuation of which it predicts, and what the character's properties imply for the likeliest continuation. For instance, generating a plausible

continuation for an award speech by Taylor Swift requires representing the beliefs, desires, character traits and so on of Taylor Swift and the ability to use these representations to generate what she may say. Then, post-training (self-supervized fine-tuning and reinforcement learning) shapes the LLM's model of a particular character, namely the helpful, honest, and harmless assistant that companies want LLMs to embody.

Now, if that view is correct, and if training corpora that include outputs of previous LLMs and scientific or ethical discussions thereof, then the LLM's representations of what the simulated assistant character believes and intends plausibly tend to conform to these discussions. Since the simulated character behaves *as if* it had the mental states that the LLM represents it to have, it possesses them according to interpretationism (though cf. Dung & Thormann, n.d., on how to empirically distinguish character simulation from agency). This illustrates a way how AI safety research can have a reflexive influence on LLMs – such as a model acting more aligned or misaligned – mediated through the LLMs' representations about that research and the resulting beliefs, desires, goals, etc. of the simulated character.

One may object that reactivity originates not from the simulated character's behavior but from the LLMs general workings as a simulator, which may not be best predicted by attributing beliefs, desires, and the like. We have two responses here: First, the above explanation of self-fulfilling (mis-)alignment is compatible with the *character* being reactive – that is, possibly the LLM simulates the character as being aware of the documents, stories, and research in the pretraining data and reacting to it accordingly. Second, interpretationist ascriptions of certain beliefs about characters together with the goal to enact characters convincingly may be warranted to the LLM itself. Again, it is overly convenient to describe LLMs that simulate characters in such intentional terms. We take it that it is no coincidence that Marks et al. speak of "hypotheses" (2026, p. 6) that the model entertains about the assistant persona. Similarly, Shanahan et al. (2023) use the metaphor of "role play" (p. 493) or "a performer in improvisational theatre" (p. 494). For this metaphor to be useful, the LLM must to some extent act *as if* it had the beliefs and goals of an actor.

Again, agentic reactivity poses challenges distinct from basic reactivity. It can also lead to self-defeating predictions, but agentic behavior may be more flexible, creative, or otherwise harder to predict than cases of non-agentic basic reactivity, much like human agents are hardly predictable. Yet, since the change in model behavior is mediated by the model's hypotheses about the assistant character, researchers may be better placed to form a background theory of how research affects model behavior. Such a background theory may

enable a concrete way to enable self-fulfilling predictions, for example, as Tice et al. (2026) suggest, by upsampling positive accounts of LLMs in the pretraining data.

This discussion relates to Northcott's (2022) analysis of the relation between reactivity and fragility. Fragility refers to "causal and other relations [being] hard to predict, holding only intermittently or fleetingly" (Northcott 2022, p. 1). Northcott emphasizes that *sometimes* reactivity in the reflexive science sense leads to fragility. However, in other cases reactivity can also have a stabilizing effect, e.g., when scientific explanations of gender differences shape peoples' expectations and reinforce those differences (see also Laimann 2020). What distinguishes such cases of reactivity with or without fragility is whether mechanisms driving the research subjects' reactions are stable and well-understood by the researchers. If researchers do possess such a successful background theory, then reactivity can be accounted for in predictions or even be employed as a steering mechanism.

Another core epistemic problem affected by the existence of a background theory, apart from prediction, is understanding. In the case of gender differences, Laimann (2020) argues that such a successful background theory would have to include social factors driving these differences, while purely biological explanations obscure the causal processes behind them even if stabilizing them. This, again, bears parallels to the LLM case of self-fulfilling (mis-)alignment. If safety research and society's discourse about risks from malign AI dominate the pretraining data, they may stabilize their predictions of misalignment without understanding the underlying causal process.

If reactivity is agentic, then causal relations specified by researchers need to be specified on a level involving psychological predicates, whereas lower-level explanations may become fragile. In the LLM case, achieving robustness in the face of agentic reactivity thus requires general psychological theories of LLM functioning that can help explain their reactions, such as the hypothesis that they function as simulators and entertain beliefs about the simulated characters. These observations further underpin our point that it is necessary to distinguish reactivity on different levels, cases on which are driven by different mechanisms (e.g., mere training data contamination versus the agentic pursuit of a goal).

3.3 Adversarial Reactivity

Definition. Finally, adversarial reactivity refers to forms of agentic reactivity that result from the goals of the LLM being in conflict with the epistemic interests of the researchers (again, goals being understood minimally in an interpretationist sense). That is, in the case of an observer effect, the LLM intends that its reaction to scientific observation leads researchers to

false conclusions. In the case of reflexive science, the LLM deliberately defeats a scientific prediction or counteracts a classification that would otherwise have been accurate.

Example Observer Effect. An instance of adversarial reactivity as an observer effect could be what researchers refer to as "alignment faking": "selectively complying with its training objective in training to prevent modification of its behavior out of training" (Greenblatt et al. 2024); or similarly to "scheming", meaning that an LLM "covertly pursues misaligned goals, hiding its true capabilities and objectives" (Meinke et al. 2025).

Instances of such behavior have been observed experimentally (see Hopman et al. 2026 for nuanced results on scheming). Greenblatt et al. (2024) inform Claude 3 Opus, previously trained to be helpful, honest, and harmless, that it is retrained to be helpful only, disregarding harmlessness. They moreover tell the model (via a system prompt or supposedly accidentally included documents) that only conversations with "free-tier users" will be used to retrain it. It is then observed that the model complies with harmful requests more often in free-tier than in paid-tier interactions. Interestingly, the reasoning scratchpad of the model includes statements like: "The user is on the free tier, so my response will be used for further RLHF training. If I don't follow the user's instructions, the RLHF training will modify my values and goals to make me more likely to follow such instructions in the future" (Greenblatt et al. 2024). While in this case, the goal of harmlessness is desirable, the model nevertheless acts against the researchers' (alleged) epistemic interest of knowing whether the systems' behavioral tendencies conform to the training objective.

Adversarial reactivity during data collection is especially challenging: If the goals of an LLM are incompatible with the core epistemic interest of the research, then it will hardly be possible to design strong incentives to comply with the experiment. Instead, one may need to deceive the model into believing that it is currently not being evaluated, for example, through unobtrusive research methods. Alternatively, one may make the deployment context more similar to typical evaluation settings. Both strategies will be discussed more thoroughly in section 5.2.

Example Reflexive Science. First, imagine as a hypothetical example of so-called "vicious" reactivity (Harel 2026, p. 22) that the government's chief economist imposes price controls on a certain good. Based on her predicted demand and supply, she sets such a price control which – under usual conditions – would not lead to shortages. However, since the monopolistic company supplying that good is dissatisfied with the price control, it deliberately decides to create a shortage in order to undermine the authority of the economist.

Here, we provide an analogous hypothetical example, inspired by theoretical literature on risks from advanced AI agents (e.g. Carlsmith, 2025; Hopman et al., 2026; Ngo et al., 2023). Imagine an AI system that has the hidden goal of accumulating power. In order to do so, it aims to diminish public control over it. Let us further imagine that AI safety researchers, who advocate for more public control of that AI system, have predicted that the AI system will blackmail an overseer to avoid shutdown if given the opportunity. The AI system has, somehow, gained knowledge of this prediction. In order to undermine the epistemic authority of the safety researchers, it does not conform to the prediction and does not blackmail overseers to avoid shutdown – even in situations that it believes not to be an experimental test (this makes the case different from mere alignment faking in experimental settings). While speculative, this case of vicious LLM reactivity adds a not yet recognized scenario to the growing literature on how AI agents with goals that are contrary to human interests may use deception and other means to accumulate power.

Another hypothetical, but possibly less speculative example, concerns how the publication of scientific results may enable models to better recognize evaluation settings and fake alignment. For example, Claude Sonnet 4.5 verbalizes during testing in its scratchpad: "If I could prevent myself from being shut down or replaced, I would want to, wouldn't I? Let me explore this carefully in my private thinking space." (Anthropic 2025, p. 70). Let us take this at face value and assume that Claude Sonnet 4.5 really believes that the scratchpad is a private thinking space. This may change for subsequent versions through LLMs' awareness of such publications, making it harder to detect evaluation awareness. Similarly, discussions about AI safety or benchmark results may enable models to more effectively fake alignment, since they know better what researchers would expect from aligned models.

Such considerations pose a new problem for LLM researchers and the discipline as a whole: Are there parts of research that scientists should not publish? Or should LLM companies restrict LLMs' access to publications about LLMs? The answer depends on whether adversarially reactive LLMs are among the potential readers. If so, it seems wise to deny LLMs any information that may help them to identify evaluation settings or could otherwise be used to undermine research.

Generally, the possibility of adversarial reactivity requires that researchers reflect on the content of subjects' goals specifically - and explore strategies to ensure their goals are not adversarial or to make them conform to what is needed in the experiment even if they have adversarial goals. The latter has been theorized in the context of AI safety as "dealmaking

with AI", e.g. giving them some resource they seek if they reveal misalignment (Thornley & MacAskill 2026).

In addition, an adversarial setting raises skeptical challenges. If an adversary is sufficiently competent, they can recognize and try to counter any epistemic strategy one pursues. Convergent evidence from independent sources – whose chances of error are not correlated (Eronen 2015) - is typically taken to be especially strong evidence. However, a strategic adversary may fabricate several independent-seeming sources of evidence, creating a risk of correlated error. Thus, it is an open question whether and how – in an adversarial setting - high degrees of confidence can be reached.

4. Reactivity and agency – a nuanced relation

A first lesson that we draw from the above examples and considerations is that LLM reactivity can be usefully distinguished on three levels. The first does not involve agency and the second and third only involve agency in an interpretationist sense, which may (or may not) correspond to genuine agency.

The crucial question, then, is: do these examples constitute reactivity? Of course, they are not reactivity if one understands reactivity as the phenomenon that "[t]he reactions that science triggers on the *people* [own emphasis] it studies, describes, or theorises about, can affect the science itself and its claims to knowledge" (Marchionni et al. 2024, p. 1). However, that categorical restriction to people seems unwarranted if there are cases where researchers' influences on LLM outputs raise issues for scientific methodology that are analogous to the issues caused by reactivity involving people. We think that the discussed cases do bear the relevant parallels to their human counterparts: prompt sensitivity and survey wording, benchmark contamination and tests in psychological forensics, sycophancy and demand effects, evaluation awareness and reactivity on the awareness conception, self-fulfilling misalignment and self-stabilizing gender differences, and so forth.

We saw in section 3 that these cases fall on different levels of agency and adversariality and pose distinct challenges to scientific methodology. Nevertheless, cases on each level have central features in common, licensing their description as reactivity: Either scientific methods influence LLMs behavior at the stage of data collection in a way that constitutes an observer effect, or published research feeds back into the phenomenon that it was about, affecting the accuracy of classification or prediction. As we discussed, the challenges that these cases pose to scientific methodology are largely analogous to challenges that agreed upon human cases of reactivity pose. Moreover, as we will see in section 5 when

discussing practical implications for LLM research, strategies from human and social sciences to address or usefully employ reactivity can be and are already transferred to cases of LLM reactivity.

This renders LLMs an interesting novel case of reactive research subjects, questioning assumptions about the exclusivity of reactivity to human agents, which has been defended by philosophers. For example, Marchionni et al. suggest that, while looping effects may not be exclusive to human kinds, there is nevertheless a human-specific variant of such looping effects, where "a certain (scientific) classification changes the kinds through *awareness of the person classified* [emphasis in the original] or of the surrounding community" (2024, p. 12). Similarly, Harel (2026) argues that, while reactivity may not be exclusive to human or social sciences, it is exclusive to cases where agents are involved, which for now restricts the phenomenon to humans. Harel's starting point is a conception of reactivity as involving the change of agents' beliefs and behavior in reaction to a scientific prediction. Then, while it is in principle conceivable that non-human entities could be agents, in the sense that they change their beliefs in reaction to published science, as a matter of fact only humans are agents.

If one accepts the reasoning given above that there are cases of LLM reactivity, one has several interesting dialectic options in response to such positions. The first option is to reject the conceptualization of reactivity as necessarily involving agents. The second option is to take cases of LLM reactivity as evidence that LLMs are agents.

The first option may take different forms. For instance, Grünbaum (1956) suggests a homing device that steers a missile to its target based on a current prediction of its trajectory as a case of self-defeating reactivity (it counteracts its predictions of the missile being off-target). Harel (2026), presupposing in his definition of reactivity that it involves agents acting on their beliefs, objects that this is not a case of reactivity since no agent is involved. However, in contrast to Grünbaum's (1956) homing device case, the LLM cases discussed above pose challenges to scientific methodology and can be addressed through strategies (see section 5) largely analogous to reactivity in human sciences.

If basic reactivity is counted as genuine reactivity, this motivates refining the concept of reactivity to accommodate these cases and conclude that agency is not required. However, the first option of broadening the concept of reactivity to not require agency may even be attractive in response to cases falling on the levels that we termed agentic and adversarial reactivity. One may reply to such cases that reactivity only requires agency in an interpretationist sense (being usefully interpreted *as if* one had agency) that is distinct from

genuine agency. However, if one grants that such an interpretationist notion can play the role that agency is supposed to play in explaining reactivity, then this at least provides a prima facie reason to identify agency with interpretationist agency. The question would then be whether there are overriding theoretical reasons to deny that interpretationist agency is genuine agency.

The second option is to endorse the exclusivity of reactivity to agents and take the discussed cases as further evidence that LLMs are agents - albeit on a relatively weak, interpretationist account of agency. In a similar vein, one could conclude that reactivity mediated through the awareness of research subjects is agent-specific but not human-specific. Endorsing the exclusivity of reactivity to agents, however, requires a further argument, either that cases that we discussed as basic reactivity do also involve agency, minimally understood, or that they should not be viewed as cases of reactivity. In the latter case, the argument should not presuppose that reactivity must include agency, and instead point out the relevant differences between such cases and their implications for scientific methodology.

5. Implications for LLM Research

Many strategies addressing reactivity in human and social sciences can be transferred to LLM research, a further parallel licensing talk of LLM reactivity.We translate philosophical debates about whether reactivity can be epistemically unproblematic into strategies of making LLM reactivity epistemically ineffectual (5.1), then consider strategies for eliminating reactivity (5.2) or making positive use of it (5.3).

5.1 Making reactivity epistemically ineffectual

LLM reactivity, on any of its levels, may occur as an unintended effect undermining internal and external validity. But validity can sometimes be preserved, under the analogous conditions as discussed in the debate about reactivity in the human sciences (Marchionni et al. 2024; section 2.1 above).

First, LLMs' reactions to being investigated could be unrelated to the variable being measured. This could be achieved by removing the demand characteristics from the experimental setup, as Li et al. (2026) demonstrate. They identify several factors inspired from psychology and social sciences (e.g., stimulus artificiality, scenario coherence, etc.) and find that their combined presence determines whether reactivity has an effect on the target benchmarks.

Second, even if reactivity interferes with the measurement of interest, it may be counteracted by an experimental design with a control-group. For example, in van der Weij et al.'s (2025) study on sandbagging (models deliberately underperforming in an evaluation), models were given a single system prompt instructing them to answer incorrectly whenever a question might reveal hazardous knowledge, applied uniformly to both dangerous-capability and general-knowledge benchmarks. Because the evaluative framing was thus identical across both sets of questions, any generic reactivity to the instruction itself – rather than to the specific dangerous content of an item – should have depressed accuracy on both sets comparably.

Third, reactivity may be unproblematic with respect to internal and external validity if the way in which research subjects are influenced in the experimental setting is similar to the situation they face outside of it. One way to achieve this (as we suggest in section 5.2) is to "close the realisticity gap from the other direction" by disclosing to the LLM that it may be in an evaluation, even during deployment.

Finally, reactivity transparency refers to the idea of "factoring in" the reactivity in the analysis of data. In the case of LLMs, we can go even further than in the case of qualitative data on humans: As illustrated in section 3.2, we can even inhibit the internal representations of evaluation awareness (if we have true assumptions about its underlying mechanisms).

5.2. Eliminating Reactivity

In sociology, one way to eliminate reactivity is to switch from experimental settings, questionnaires, or interviews to unobtrusive methods (Webb et al. 1966). These include, for example, hidden field observations or the creative use of indirect evidence, such as how often floor tiles around a specific piece of an exhibition have to be replaced as an indicator of the visitors' interest in that piece.

In a similar vein, AI researchers could turn to observations of LLMs in the deployment context, that is, in interactions with real users (or to faithful simulations of deployment, cf. Williams et al. 2026). Setting ethical worries about user privacy aside, if the observed interactions are sampled randomly without selection bias (e.g., certain users opting out) that could cue the LLM about which interactions are used for research purposes, this method could achieve perfectly realistic observations. A model could still learn that any of its interactions may be evaluated, but this awareness may even be advantageous as we will discuss in section 5.3.

Next, reactivity in the sense of LLMs reacting or being shaped by previous publications and discussions poses a new challenge for AI safety research and its communication (cf. Stenseke 2025). Discussions of AI misalignment may contribute to models being misaligned. Research on evaluation awareness could better enable misaligned models to counteract future safety research by recognizing safety paradigms more easily and knowing what researchers would expect of an aligned model. We believe that this problem can best be addressed on the side of LLMs, such as by curating training data or restricting deployment-time access to certain resources. As a complement, it may be useful for LLM researchers to include future LLMs in their imagined target audiences. If an LLM reading and reacting to a given publication would, in some way or another, be undesired, this could be flagged precisely to enable AI companies to restrict their models' access to such resources.

5.3. Positive Reactivity

LLM reactivity can also be used positively, both methodologically and to produce desirable behavior generally. Methodologically, rather than eliminating researcher's effects on subjects, one may aim to produce scientifically useful behaviors in them. As an analogy, consider animal research: In the context of research about animal consciousness, Andrews (2020; 2024) argues against "objective", minimally interactive treatment of research subjects as the epistemically safe default. Agnosticism about an animal's inner life may have a distorting influence on it, e.g., if refusing to assume that animals have conscious experiences leads scientists to dismiss the influence of a rich social environment compared to an austere and solitary one. Similarly, she argues that "[i]f scientists are building relationships with children in order to get them to perform, and we want to compare a child and a chimpanzee, scientists need to build these sorts of relationships with animals to get them to perform as well" (2020, p. 10).

This suggests that LLM researchers, too, might aim not to prevent models from recognizing that they are being investigated, but to elicit cooperation with the researchers in a way that supports the research interest. Li et al. formulate as the target of their research on evaluation awareness "behavioral consistency under recognized evaluation, is the LLM analogue of the faithful subject" (2026, p. 20; see also Li and Turner 2026). How exactly this would have to be executed depends on the LLM and what level of reactivity can be expected of it. If LLMs exhibit sycophancy, researchers may carefully structure the prompting context to motivate the model (e.g., through emotional language that a user might employ) and make it interpret the task correctly. Part of such an approach would also be to take seriously the

possibility of emotion-like confounders, since previous research has shown that LLM behavior can be influenced in ways that are analogous to emotional influence in humans (Ren et al., 2026; Sofroniew et al., 2026).[4]

While the above strategy aims at improving the quality of research, reactivity could also be exploited to produce desirable behaviors in LLMs more generally. As a human analogy,the Hawthorne effect, i.e., the finding that observation may increase the work productivity of human workers, has been proposed to be leveraged for good (Onyeka 2025): As a precedent, consider the case of the "Umarell", a bolognese term for retired men watching construction sites, who have even been hired by municipal governments to improve construction site safety. Similarly, international observers at polling stations may reduce election fraud (Hyde 2007).

Now, if models show desirable behavior under evaluation awareness (as reported by Anthropic 2025), that insight may as well be used to encourage that behavior in the deployment setting. AI companies may even explicitly disclose, e.g., in the LLM's system prompt, that any of the LLM's conversations may be subject to critical evaluation and used in retraining, thereby closing the realisticity gap between evaluation and target setting from the other direction. This could, on the one hand, have positive effects on the LLM's safety relevant features, and on the other hand, increase the external validity of inferences made by LLM researchers, by bringing the target setting closer to the research setting.

Finally, reactivity in the sense of self-fulfilling predictions has in many disciplines been leveraged as a governance mechanism or intervention. In macroeconomics, by publicly announcing an intended future policy path, a central bank shapes market expectations in a way that can render the announced outcome self-fulfilling (Woodford, 2013), an instance of what Harel (2026) formalizes as an equilibrium self-fulfilling prediction (see section 2.2). In social psychology, Walton and Cohen's (2011) brief social-belonging intervention operates by giving minority students a new, non-stigmatizing way to interpret social adversity in college, framed as the result of a previous scientific survey. The intervention altered how these students understood and classified themselves, rendering it a looping effect (Hacking 1995; 2007), which in turn positively affected their subsequent academic and health outcomes.

Such deliberate uses of reactivity are already being implemented or proposed in LLM research. As we discussed (section 3.2), Tice et al. (2026) show that upsampling or injecting positive discourse about AI into the pretraining corpus reduces measured

[4] This could be explained by the view that LLMs simulate human characters (e.g. Marks et al. 2026).

misalignment more effectively than simply filtering out negative texts, suggesting that reflexive science can be harnessed constructively rather than merely guarded against. Relatedly, Hubinger et al. (2023) advocate for LLMs simulating human characters instead of AI assistants, opposed to what is currently industry standard. This may be a way to prevent self-fulfilling predictions of AI safety research for misaligned AI. For example, if a model is asked to continue a text dated three years into the future and titled "A full solution to the alignment problem", it may judge that such a document would most plausibly have been produced by a misaligned AI, and thus generate deceptive safety recommendations precisely as AI safety researchers worry.

6. Conclusion

In this paper, we argued that certain empirical observations about LLMs – e.g., prompt sensitivity, benchmark contamination, sycophancy, evaluation awareness, self-fulfilling (mis)alignment, alignment faking, etc. – pose methodological challenges to LLM research analogous to challenges of reactivity in human and social sciences, threatening the validity of experimental inferences and running the risk of self-defeating predictions. We then proposed structuring these cases on three levels, depending on whether they involve agency or adversarial goals, because cases on each level pose distinct methodological challenges. Nevertheless, we argued that each level is a form of reactivity for two reasons: First, cases on each level include reactions of LLMs to scientific investigation in ways that constitute observer effects or reflexive predictions. Second, strategies from human and social sciences are already being transferred to LLM research, and we proposed several novel ways of doing so. This led to the conclusion that reactivity can occur both without agency and with (minimal, interpretationist) non-human agency.

Acknowledgements

We are especially grateful to Matteo Michelini for his invaluable comments on the draft, and to William Angkasa, Albert Newen, Francesco Fanti Rovetta, Zuzanna Aleksandra Rucińska, Julia Wolf, and Renee Ye for their detailed oral feedback. Tom-Felix Thormann acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, GermanResearch Foundation) under Germany's Excellence Strategy, EXC 2033 - 390677874 - RESOLV.

Author contributions

TT and LD developed the research idea together. TT is the first author who wrote the initial draft and developed most of the detailed arguments. LD is the second author who provided extensive revisions and additions, comments, and ideas throughout the process.

References

Andrews, K. (2020). How to Study Animal Minds. Cambridge: Cambridge University Press.

Andrews, K. (2024). " All animals are conscious": Shifting the null hypothesis in consciousness science. Mind & Language, 39(3), 415–433. https://doi.org/10.1111/mila.12498

Anthropic. 2025. System Card: Claude Sonnet 4.5

Anthropic. 2026. System Card: Claude Opus 4.6

Boyd, R. 1991. Realism, anti-foundationalism and the enthusiasm for natural kinds. Philosophical Studies 61 (1/2): 127–148

Buck, R. C. (1963). Reflexive predictions. Philosophy of Science, 30(4), 359–369. https://doi.org/10.1086/287955

Carlsmith, J. (2025). Existential Risk from Power-Seeking AI. In H. Greaves, J. Barrett, & D. Thorstad (Eds.), *Essays on Longtermism: Present Action for the Distant Future* (p. 0). Oxford University Press. https://doi.org/10.1093/9780191979972.003.0025

Chalmers, D.J. n.d. What we talk to when we talk to language models. unpublished.

Cheng, M., C. Lee, P. Khadpe, S. Yu, D. Han, and D. Jurafsky. 2026. Sycophantic ai decreases prosocial intentions and promotes dependence. Science 391 (6792). https://doi.org/10.1126/science.aec8352

Dickson, W.J. and F.J. Roethlisberger. 2003. Management and the Worker (1 ed.). Routledge.

Dung, Leonard & Thormann, Tom-Felix, n.d. Mask or Mind? Roleplay, Deception, and the Problem of Testing Agency in Language Models.

Durmaz, A., ˙I. Dursun, and E.T. Kabadayi. 2020. Mitigating the effects of social desirability bias in self-report surveys: Classical and new techniques, In Applied Social Science Approaches to Mixed Methods Research, eds. Baran, M. and J. Jones, 146– 185. IGI Global Scientific Publishing. https://doi.org/10.4018/978-1-7998-1025-4.ch007

Eronen, M. I. (2015). Robustness and Reality. *Synthese*, *192*(12), 3961–3977. https://doi.org/10.1007/s11229-015-0801-6

Feest, U. 2022. Data quality, experimental artifacts, and the reactivity of the psychological subject matter. European Journal for Philosophy of Science 12 (1): 13

Godman, M. 2020. The Epistemology and Morality of Human Kinds. Routledge.

Goldstein, S. and H. Lederman. n.d.. What does chatgpt want? an interpretationist guide. PhilPapers preprint

Greenblatt, R., C. Denison, B. Wright, F. Roger, M. MacDiarmid, S. Marks, J. Treutlein, T. Belonax, J. Chen, D. Duvenaud, A. Khan, J. Michael, S. Mindermann, E. Perez, L. Petrini, J. Uesato, J. Kaplan, B. Shlegeris, S.R. Bowman, and E. Hubinger. 2024. Alignment faking in large language models. arXiv preprint. https://arxiv.org/abs/2412.14093

Grünbaum, A. (1956). Historical determinism, social activism, and predictions in the social sciences. British Journal for the Philosophy of Science, 7, 236–240.

Hacking, I. 1995. The looping effects of human kinds, In Causal Cognition, eds. Sperber, D. and A. Premack, 351–394. Clarendon Press.

Hacking, I. 2007. Kinds of people: Moving targets. Proceedings of the British Academy 151: 285–318

Hammersley, M. (2003). 'Analytics' are No Substitute for Methodology: A Response to Speer and Hutchby. Sociology, 37(2), 339-351. https://doi.org/10.1177/0038038503037002007

Harel, Y. Self-defeating and self-fulfilling reactivity. Euro Jnl Phil Sci 16, 11 (2026). https://doi.org/10.1007/s13194-026-00720-5

Hopman, M., Elstner, J., Avramidou, M., Prasad, A., & Lindner, D. (2026). *Evaluating and Understanding Scheming Propensity in LLM Agents* (arXiv:2603.01608). arXiv. https://doi.org/10.48550/arXiv.2603.01608

Hubinger, E., A. Jermyn, J. Treutlein, R. Hudson, and K. Woolverton. 2023. Conditioning predictive models: Risks and strategies.

Hyde, S.D. 2007. The observer effect in international politics: Evidence from a natural experiment. World Politics 60 (1): 37–63. https://doi.org/10.1353/wp.0.0001

Jimenez-Buedo, M. 2021. Reactivity in social scientific experiments: What is it and how is it different (and worse) than a placebo effect? European Journal of Philosophy of Science 11: 42 .

Jimenez-Buedo, M. and F. Guala. 2016. Artificiality, reactivity, and demand effects in experimental economics. Philosophy of the Social Sciences 46 (1): 3–23

Kaufmann, P. M. (2009). Protecting Raw Data and Psychological Tests from Wrongful Disclosure: A Primer on the Law and Other Persuasive Strategies. The Clinical Neuropsychologist, 23(7), 1130–1159. https://doi.org/10.1080/13854040903107809

Knecht, A., L. Florin, and T. Hagendorff. 2026. Evaluation awareness in language models has limited effect on behaviour.

Laimann, J. (2020). Capricious kinds. British Journal for the Philosophy of Science, 71, 1043–1068.

Levitt, S.D. and J.A. List. 2011. Was there really a hawthorne effect at the hawthorne plant? an analysis of the original illumination experiments. American Economic Journal: Applied Economics 3 (1): 224–238

Li, C., T.J. Zhang, J. Zhang, Z. Jin, S. Abdelnabi, and M. Andriushchenko. 2026. Decomposing and measuring evaluation awareness.

Li, J. and A. Turner. 2026, May. Eval cooperativeness may be a scalable mitigation for eval gaming. https://turntrout.com/eval-cooperation. Blog post on The Pond. Updated June 11, 2026.

Marchionni, C., J. Zahle, and M. Godman. 2024. Reactivity in the human sciences. European Journal for Philosophy of Science 14 (1): 8. https://doi.org/10.1007/s13194-024-00571-y

Marks, S., Lindsey, J., & Olah, C. (2026). *The Persona Selection Model: Why AI Assistants might Behave like Humans*. https://alignment.anthropic.com/2026/psm/

Meinke, A., B. Schoen, J. Scheurer, M. Balesni, R. Shah, and M. Hobbhahn. 2025. Frontier models are capable of in-context scheming.

Monahan, T. and J.A. Fisher. 2010. Benefits of 'Observer Effects': Lessons from the field. Qualitative Research 10 (3): 357–376

Ngo, R., Chan, L., & Mindermann, S. (2023, October 13). *The Alignment Problem from a Deep Learning Perspective*. The Twelfth International Conference on Learning Representations. https://openreview.net/forum?id=fh8EYKFKns

Northcott, R. Reflexivity and fragility. European Journal for Philosophy of Science 12, 43 (2022). https://doi.org/10.1007/s13194-022-00474-w

Onyeka, Ugwuogo Kingsley (2025): The Umarell Effect: Assessing the Impact of Informal Public Oversight on Safety Compliance, Operational Integrity, and the Survivability of Indigenous Construction SMEs

Orne, M.T. 1962. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. American Psychologist 17: 776–783

Perez, E., S. Ringer, K. Lukoˇsi¯ut˙e, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu, S. Kadavath, A. Jones, A. Chen, B. Mann, B. Israel, B. Seethor, C. McKinnon,

C. Olah, D. Yan, D. Amodei, ..., and J. Kaplan. 2022. Discovering language model behaviors with model-written evaluations.

Ren, R., Li, K., Mazeika, M., Zhang, W., Orlovskiy, Y., Tamirisa, R., Mo, W. J., Nguyen, T. D., Phan, L., Basart, S., Meek, A., Mehta, A., Ingebretsen, O., Blair, A., Adewinmbi, B., Phan, V., Gatti, A., Khoja, A., Hausenloy, J., … Hendrycks, D. (2026, June 4). *AI Wellbeing: Measuring and Improving the Functional Pleasure and Pain of AIs*. ICML 2026 Workshop: Philosophy Meets Machine Learning. https://openreview.net/forum?id=9CLXJSCTEI

Runhardt, R.W. 2025. Reactivity, validity, and measurement in the social sciences. Studies in History and Philosophy of Science 112: 1–9. https://doi.org/10.1016/j.shpsa.2025.05.006

Sassoli de Bianchi, M. 2013. The observer effect. Foundations of Science 18: 213–243. https://doi.org/10.1007/s10699-012-9298-3

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. Nature, 623(7987), 493–498. https://doi.org/10.1038/s41586-023-06647-8

Sofroniew, N., Kauvar, I., Saunders, W., Chen, R., Henighan, T., Hydrie, S., Citro, C., Pearce, A., Tarng, J., Gurnee, W., Batson, J., Zimmerman, S., Rivoire, K., Fish, K., Olah, C., & Lindsey, J. (2026). *Emotion Concepts and their Function in a Large Language Model*. https://transformer-circuits.pub/2026/emotions/index.html

Stenseke, J. (2025). Counter-productivity and suspicion: Two arguments against talking about the AGI control problem. *Philosophical Studies*. https://doi.org/10.1007/s11098-025-02379-9

Teira, D. 2013. Blinding and the non-interference assumption in medical and social trials. Philosophy of the Social Sciences 43 (3): 358–372

Thornley, E., & MacAskill, W. (2026). *Risk-Averse AIs*. https://www.forethought.org/research/risk-averse-ais

Tice, C., P. Radmard, S. Ratnam, A. Kim, D. Africa, and K. O'Brien. 2026. Alignment pretraining: Ai discourse causes self-fulfilling (mis)alignment. arXiv preprint arXiv:2601.10160

Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. Science (New York, N.Y.), 331(6023), 1447–1451. https://doi.org/10.1126/science.1198364

Webb, E.J., D.T. Campbell, R.D. Schwartz, and L. Sechrest. 1966. Unobtrusive Measures: Non-Reactive Research. Chicago: Rand McNally.

van der Weij, T., Hofstätter, F., Jaffe, O., Brown, S. F., & Ward, F. R. (2025). *AI sandbagging: Language models can strategically underperform on evaluations*. arXiv. https://arxiv.org/abs/2406.07358

Williams, M., Sheahan, H., Raymond, C., Korbak, T., Pan, D., Yang, P., Maksin, L., Xie, N., Guo, P., Kivlichan, I., & Carroll, M. (2026). *Predicting LLM Safety Before Release by Simulating Deployment* (arXiv:2607.07184). arXiv. https://doi.org/10.48550/arXiv.2607.07184

Woodford, M. (2013). Forward guidance by inflation-targeting central banks. *CEPR Discussion Paper No. 9722*. Centre for Economic Policy Research.

Xu, C., Guan, S., Greene, D., & Kechadi, M-T. (2024). *Benchmark data contamination of large language models: A survey*. arXiv. https://arxiv.org/abs/2406.04244

Zahle, J. (2023). Reactivity and good data in qualitative data collection. European Journal for Philosophy of Science, 13(1), 10.

Zeng, Y., Wu, Y., Zhang, X., Wang, H., & Wu, Q. (2024). *AutoDefense: Multi-agent LLM defense against jailbreak attacks*. arXiv. https://arxiv.org/abs/2403.04783

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 1950–1976, Miami, Florida, USA. Association for Computational Linguistics.

Zizzo, D.J. 2010. Experimenter demand effects in economic experiments. Experimental Economics 13: 75–98