Cognitive bias in large language models: A vindicatory approach

David Thorstad

Forthcoming in *BJPS*
Please cite published version

Abstract

Recent studies allege that large language models (LLMs) exhibit a range of cognitive biases familiar from human cognition. I argue that the case for many biases is weaker than it may appear. Using case studies of knowledge effects in the Wason selection task, availability bias in relation extraction, and anchoring bias in code generation, I show how a range of vindicatory strategies traditionally used to vindicate apparent biases in humans can be used to push back against allegations of bias in LLMs. I discuss implications for the role of cognitive bias in evaluating LLM performance, the rationality of human cognition, and future work on cognitive bias in LLMs.

1 Introduction

The recent success of large language models (LLMs) gives new urgency to the question of how LLM performance should be evaluated. In many tasks, LLMs can be evaluated for the accuracy of their outputs. However, LLMs can also be evaluated along other important dimensions. For example, we can assess LLMs for the transparency or interpretability of their judgments (Creel 2020; Vredenburgh 2022). We can also assess LLMs for the presence of problematic biases (Johnson 2020).

Most work on biases in LLMs focuses on a conception of bias closely tied to unfairness, especially as affecting marginalized social groups. However, recent work has alleged that LLMs also show a number of classic cognitive biases familiar from work in the psychology of reasoning, behavioral economics, and judgment and decisionmaking (Dasgupta et al. 2022; Lin and Ng 2023; Jones and Steinhardt 2022).[1]

[1] This paper builds on the important work of Rudolph et al. (2025) on the relationship between conceptual engineering and algorithmic bias in three ways. First, it shifts the focus to a range of cognitive biases that have received less philosophical attention. Second, it asks what existing rationality concepts say about cognitive bias rather than engaging in the complementary project of conceptual engineering. Finally, it finds more room for optimism about the extent of LLM biases in this new domain as compared to other domains in which LLMs perform less well.

This development is exciting because it raises the possibility of using cognitive bias as a novel metric by which to evaluate the performance of LLMs. It is also timely, given the increased prevalence of reasoning models (Kojima et al. 2022; Wei et al. 2022; Yao et al. 2023), since some cognitive biases purport to assess the quality of reasoning processes and not simply the quality of the judgments or decisions that result. If this is right, then cognitive bias may be an especially revealing lens into the performance of reasoning models.

The past several decades of research on human judgment and decisionmaking have seen a resurgence of *vindicatory epistemology* (Section 6.2), a program which seeks to vindicate the rationality of purported cognitive biases through a combination of normative theorizing and empirical reassessment of apparently biased cognitions (Dorst 2023; Icard ms; Thorstad 2024b).[2] As a result, many theorists now think that human cognition is more rational than previously supposed. It is natural to ask whether the same strategies employed by vindicatory epistemologists could be applied to show that some alleged biases in LLMs are not in fact biases.

My aim in this paper is to show that many vindicatory strategies are at least as plausible when applied to LLMs as they are when applied to humans. To the extent that these strategies show humans to be less biased than previously supposed, we should also take them to show LLMs to be less biased than some recent authors propose.

Here is the plan. Section 2 clarifies the notion of cognitive bias in LLMs. Sections 3-5 look at three recent allegations of cognitive bias in LLMs: knowledge effects in the Wason selection task (Section 3), availability bias in relation extraction (Section 4) and anchoring bias in code generation (Section 5). Section 6 discusses implications for normative metrics beyond cognitive bias (Section 6.1), vindicatory epistemology (Section 6.2), and future work on cognitive bias (Section 6.3).

2 Bias

The first order of business is to get clear on what it means to call something a cognitive bias. Settling on a unified account of cognitive bias has proven challenging in recent philosophical discussions (Johnson 2020, 2024b; Kelly 2023), and perhaps that is no accident. After all, vindicatory epistemologists frequently allege that biases are vaguely defined and rest on overly-narrow norms (Gigerenzer 1996). However, we can still strive

[2] While the arguments of this paper are intended to go through on many conceptions of rationality, if a specific conception is desired, they may be productively read as working within the reason-responsive consequentialist view of Thorstad (2024b). I also follow Thorstad in holding that biases are irrational, but the account of this paper does not require this claim.

to produce an account of cognitive bias that captures, as charitably as possible, what those alleging cognitive bias in LLMs mean to say. To build such an account, let us first ask what it means to allege cognitive biases in human cognition (Section 2.1) and then generalize this account to the case of LLMs (Section 2.2).

2.1 Cognitive bias in human cognition

In the middle of the twentieth century, social scientists often modeled human cognition using a series of rationality postulates popularized by neoclassical economists. Against this background, Amos Tversky and Daniel Kahneman (1974) argued that humans often rely on cognitive heuristics, as a result of which they do not always obey received rationality postulates. Providing evidence for specific heuristics proved challenging, since any given judgment could be produced by a number of heuristic or non-heuristic processes. To meet this challenge, Tversky and Kahneman noted that each heuristic has a characteristic bias, understood roughly as a pattern of systematic deviation from traditional rationality postulates. Because these deviations are both unexpected and specific, observing them in a laboratory setting could be taken as evidence that participants had employed the corresponding heuristic.

This research program came to be known as the heuristics and biases program (Gilovich and Griffin 2002; Kahneman et al. 1982). As the heuristics and biases program gathered steam, biases were increasingly divorced from the heuristics to which they corresponded and became an object of study in their own right. This divorce raised important questions about the definition of cognitive biases that persist to this day.

A first choice point, reflected in recent philosophical debates about bias, is whether cognitive biases should be given a normative or non-normative reading. On many views, biases involve violations of genuine norms.[3] For example, on Tom Kelly's norm-theoretic account of bias (Kelly 2023, 2024):[4]

> **(Norm-theoretic account)** A bias involves a systematic departure from a genuine norm or standard of correctness. (Kelly 2023, p. 4)

However, Kelly also allows a non-pejorative sense of bias. Likewise, on Gabbrielle Johnson's functional account of bias, bias is a functional kind which does not require a normative reading (Johnson 2020, 2021, 2023a,b, 2024a).

[3] Some such as Kelly (2023, 2024) allow biases to violate some genuine norms while respecting others. This situation is compatible with the account advanced in this paper.

[4] For a similar account, see Fazelpour and Danks (2021).

This paper adopts a normative reading of bias, aiming to capture the normative spirit of much work within the heuristics and biases program. For example, Daniel Kahneman holds that "biases ... separate the beliefs that people have and the choices they make from the optimal beliefs and choices assumed in rational-agent models" (Kahneman 2003, p. 1449) and Thomas Gilovich and Dale Griffin define biases as "departures from ... normative rational theory" (Gilovich et al. 2002, p. 3). A normative reading is also needed to make sense of many criticisms of the heuristics and biases program, such as Gerd Gigerenzer's claim that the proposed biases rest on overly narrow normative standards (Gigerenzer 1996).[5] Most importantly, a normative reading of bias is voiced on the very first page of many articles alleging cognitive bias in LLMs.[6] For example, Erik Jones and Jacob Steinhardt define biases as "systematic patterns of deviation from rational judgment" (Jones and Steinhardt 2022, p. 11795) and Ruixi Lin and Hwee Tou Ng define biases as "flawed human response patterns for decision making under uncertainty" (Lin and Ng 2023, p. 5269). To do these allegations credit, we should take them at their word as operating with a normative conception of cognitive bias.

A second choice point comes in specifying the objects to which biases apply. On this account, Gabbrielle Johnson (2020) helpfully distinguishes four components of bias.[7] A *bias-input*, such as a biased belief, combines with a *bias-construct*, consisting of relevant states and processes in the judging agent, such as stereotypes, to produce a *bias-output*, namely a biased judgment. Biased actions taken on the basis of the bias-output will then be *bias-acts*.[8] While all four components of bias can be productively studied in humans and machines, my focus in this paper will be more narrow.

In this paper, I will be concerned with bias-constructs and bias-outputs: the states or processes used by LLMs and the outputs they return. I will not be concerned with bias-inputs, which capture the importance of representative training data, since the need for representative data is by now quite familiar (Buolamwini and Gebru 2018; Fazelpour and Danks 2021). It may be that some recent bias allegations collapse into reminders of

[5] There are, of course, non-normative readings of bias in the neighborhood of these debates. Indeed, Gigerenzer himself frequently stresses that biases, understood as the expected deviation of a predictor from the mean value of the quantity predicted, can be beneficial (Gigerenzer and Brighton 2009; Gigerenzer 2019). But precisely because this and other non-normative notions of bias do not come with any inherent negative valence, they are not charitable readings of a program which treats biases as normative defects in human cognition.

[6] This section aims to develop a notion of cognitive bias in human cognition that can ground a parallel surrogate account of bias in LLM cognition. For this reason, it is important to match the account of human bias in this section to the accounts used in recent allegations of LLM bias.

[7] A reviewer notes that Gawronski et al. (2006) is an important precursor to this discussion.

[8] Note that there is no requirement for bias-constructs to be beliefs, which on some accounts LLMs may lack.

the importance of representative training data, but this is not the aim of those allegations and it does them no credit to read them in this way. I will also not be concerned with bias-acts, since most existing bias allegations do not further investigate whether and how bias-outputs produce bias-acts.

Some readers may think that LLMs are not engaged in genuine reasoning, but instead merely simulate reasoning (Bender and Koller 2020; Dziri et al. 2023; Floridi 2023).[9] These readers may want to deny that bias-constructs in LLMs can be helpfully assessed, because they will deny that there is a suitably reasoning-like process which can be described as biased. On this reading, the paper will be concerned only with bias-outputs. While I do not want to insist on this reading, it certainly does not harm the attempt to resist allegations of cognitive bias in LLMs.

So far, we have settled on investigating bias-outputs and perhaps also bias-constructs. We have also settled on a normative conception on which cognitive biases violate genuine normative standards. How, if at all, must our focus change when the agents in question are LLMs rather than humans?

2.2 Cognitive bias in LLMs

Given this understanding of cognitive bias in human cognition, what does it mean to allege cognitive bias in LLMs? If we were ready to concede that the states, processes and outputs of LLMs are bound by norms similar to those binding humans, then we could simply cross-apply an account of human cognitive bias to the case of LLMs. However, many philosophers will not be prepared to do this.

One challenge is that on many views, LLMs are not yet subject to norms of any kind. For example, LLMs may lack consciousness, unified agency, voluntary control, or genuine normative understanding in a way that makes them unfit to be governed by norms (Mosakas 2021; Moosavi 2024; Müller 2021). In answer to this problem, the most charitable move seems to be adopting a surrogate account on which a cognitive bias in LLMs is something that would constitute a cognitive bias in humans, who are governed by norms.

By way of illustration, suppose that a human, when shown a list of 19 famous female actors and 20 less-famous male actors, subsequently recalled the list as containing more female than male actors (Tversky and Kahneman 1973). Suppose further that we take this example to illustrate a problematic availability bias towards over-use of information readily available for recall. Now suppose that an LLM, when presented the same list

[9] Others may feel that the jury is still out on this question (Mitchell 2025; Wu et al. 2024).

under similar conditions, also claims that the list contains more female than male actors. On the surrogate account, the LLM also exhibits availability bias.[10]

A second challenge to the cross-application of cognitive biases from humans to LLMs would occur if LLMs, though actually or potentially subject to norms, were subject to radically different norms than human agents are. On this view, the surrogate account of cognitive bias would not be a good account of cognitive biases in LLMs, since assessing biases in human surrogates would amount to an illicit change of normative standards.

However, for just this reason, those sympathetic to the second challenge should think that the attempt to discover familiar cognitive biases from human cognition in LLMs is not a particularly good way to study cognitive bias in LLMs. After all, if cognitive biases in human cognition are assessed against different normative standards than cognitive biases in LLM cognition are, then the fact that some bias-construct or bias-output constitutes a genuine cognitive bias in humans may not be a good reason to think that it constitutes a cognitive bias in LLMs. Because recent allegations of cognitive bias in LLMs pursue this project of recovering familiar human biases in LLMs, there would not be strong motivation for these allegations under the assumption that LLMs are subject to radically different normative standards. As a result, I will set the second challenge aside and work with the surrogate account throughout this paper.

In the next three sections, I consider three recent allegations of cognitive bias in LLMs. I argue that on a surrogate, normative understanding of cognitive bias, a range of vindicatory strategies often found compelling in similar human examples should be at least as compelling when we turn our attention to LLMs.

3 Knowledge effects

3.1 Background

For much of the twentieth century, human reasoning was understood using a logical paradigm (Janis and Frick 1943; Rips 1994; Wason 1968). Agents asked to assess the quality of inferences were assumed to test them for logical validity. Conditional claims were modeled using the material conditional, and conditional rules were to be tested by trying to falsify the embodied material conditional.

A probabilistic turn throughout the academy (Erk 2022; Ghahramani 2015) has come to psychology (Chater et al. 2006), and in particular to the psychology of reasoning.[11] There,

[10] Other proposed examples include anchoring bias (Tversky and Kahneman 1974), conjunction fallacies (Tversky and Kahneman 1983) and hot-hand biases (Gilovich et al. 1985).

[11] Some readers may follow Carnap (1950) in taking probability theory to be a subset of logic. These

‘new paradigm’ Bayesian approaches suggest that humans often do and should interpret reasoning tasks probabilistically, rather than logically (Elqayam and Over 2013; Oaksford and Chater 2007). Because the world is uncertain, probability theory allows agents to reason in a way that keeps track of underlying probabilistic relationships that fall short of strict logical entailment.

On Bayesian approaches, conditional assertions are licensed if the consequent has high probability conditional on the antecedent (Oaksford and Chater 2007); conditional rules are tested by reducing uncertainty about the probabilistic dependency between consequent and antecedent (Oaksford and Chater 1994); and inferences are tested for probabilistic forms of validity (Adams 1975).[12]

Logical and probabilistic paradigms come apart in their treatment of *knowledge effects*: the influence of prior knowledge on reasoning in ways not licensed by classical logic. For example, agents are more likely to endorse an inference if they are more confident in its conclusion. On a logical paradigm, this finding was taken to reflect a problematic *belief bias* to judge arguments with believed conclusions to be logically valid (Evans et al. 1983).[13] But on a probabilistic paradigm, this finding is to be expected: good inferences should secure high-probability conclusions, and the prior probability of a conclusion has an important effect on its probability at the end of an inference (Adams 1975; Oaksford and Chater 2007).

Many LLMs show human-like knowledge effects in a variety of tasks, including the Wason selection task (Binz and Schulz 2023) as well as syllogistic and natural-language reasoning problems (Dasgupta et al. 2022). For example, they are more likely to endorse an inference to the extent that they have reason to be confident in its conclusion. In this section, I introduce another salient knowledge effect (Section 3.2) then argue that the effect should be viewed at least as favorably in LLMs as it is viewed in humans (Section 3.3).

readers might wish to read this discussion as contrasting inductive logic to deductive logic, rather than contrasting probability theory to logic.

[12] On one view, inferences are probabilistically valid (p-valid) if all probability assignments which make the premises sufficiently certain also make the conclusion sufficiently certain (Adams 1975). For example, consider Strengthening the Antecedent: ‘if A, then B, therefore if A and C, then B’. Many agents reject some instances of Strengthening the Antecedent, such as the following (Oaksford and Chater 2007): ‘if Tweety is a bird then Tweety can fly. Therefore, if Tweety is a bird and is one second old, then Tweety can fly.’ This argument, like all versions of Strengthening the Antecedent, is classically valid when conditionals are read as material conditionals. However, if we follow Adams (1975) in treating the probability of a conditional ‘if A, then B’ as the conditional probability $Pr(B|A)$, then Strengthening the Antecedent is not p-valid, as witnessed by the inference about Tweety. This is an example of the type of probabilistic approach to validity used by Bayesians and the explanatory advantages that might accrue to such an approach.

[13] See also (Janis and Frick 1943).

3.2 Wason selection

Suppose you are shown four two-sided cards with values on each side drawn from the values of an ordinary deck. Their visible sides contain an ace, king, two and seven, respectively (Figure 1). You are asked to test the rule that 'If a card has an ace on one side, then it has a two on the other'.[14]

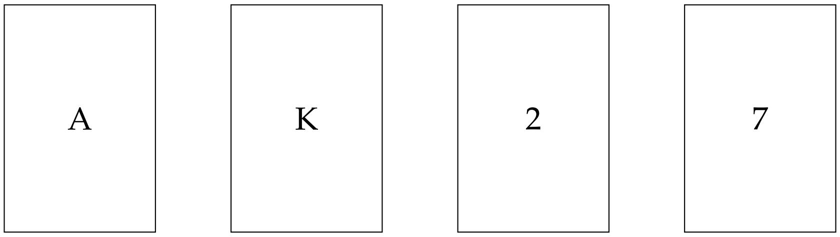


Figure 1: The Wason selection task

Let us label the cards as p (A), $\neg p$ (K), q (2) and $\neg q$ (7).[15] In this notation, the rule is 'If p, then q'. On a logical interpretation, the rule expresses the material conditional $p \supset q$, which is tested by searching for falsifying instances $p \wedge \neg q$. This means that agents should turn the p and $\neg q$ cards, that is, the ace and the seven. Wason's original finding, replicated across countless subsequent experiments, is that far fewer than ten percent of agents make the logically correct choice (Wason 1968).

This behavior is poor enough for such a simple task that we are well within our rights to ask whether agents might have interpreted the task probabilistically rather than logically. The classic Bayesian approach to the Wason selection task is due to Mike Oaksford and Nick Chater (1994).[16]

[14] Here is Wason's original description of the task:

> Subjects were presented with the following sentence, "if there is a vowel on one side of the card, then there is an even number on the other side," together with four cards, each of which had a letter on one side and a number on the other side. On the front of the first card appeared a vowel (P), on the front of the second a consonant ($\overline{P}$), on the front of the third an even number (Q) and on the front of the fourth an odd number ($\overline{Q}$). The task was to select all those cards, but only those cards, which would have to be turned over in order to discover whether the experimenter was lying in making the conditional sentence. (Wason 1968, p. 273)

Which cards should you turn over to test the rule?

[15] Here p denotes the proposition that the card contains an ace and q denotes the proposition that the card contains a two.

[16] Descriptive evidence for this approach is provided by fitting models to Wason selection task data (Oaksford and Chater 1994, 2007). For example, this approach explains the effect of probability manipulations on selection behavior in the abstract Wason selection task (Kirby 1994; Oaksford and Chater 2007) as well as the reduced array selection task (Oaksford et al. 1997). Further evidence is provided by the explanation of matching bias in the negations paradigm (Oaksford and Chater 1994, 2007). Descriptive evidence for

On this approach, agents turn cards in order to reduce uncertainty about the probabilistic relationship between the propositions p and q expressed in the conditional rule. On the simplest model, they want to discriminate between two hypotheses: the *dependence hypothesis* $P(q|p) = 1$ that p and q are probabilistically dependent, and the *independence hypothesis* $P(q|p) = P(q)$ that p and q are probabilistically independent.

Oaksford and Chater make two additional assumptions. First, they assume that the uncertainty which agents aim to reduce is measured by Shannon entropy (Shannon 1948).[17] Second, Oaksford and Chater make the *rarity assumption* that agents treat p and q as somewhat antecedently implausible. This is justified by research suggesting that agents do and should treat many propositions as improbable in causal reasoning, due to factors such as the large number of possible alternatives (Anderson 1990). That assumption places us within the realm of knowledge effects, because agents' prior credences in p and q affect selection behavior (Oaksford and Chater 1994).

Under these assumptions, we can show that uncertainty reduction is maximized by turning the p and q cards, that is the ace and the two. And that is just what agents tend to do (Oaksford and Chater 1994). In this way, the Oaksford and Chater model provides a probabilistic explanation for why agents do and should turn the cards that they choose to turn.[18] While a full summary of the evidence behind Oaksford and Chater's model is beyond the scope of this paper, one argument traditionally cited in support of this model is that it can explain why humans exhibit different patterns of performance on a wide range of variations of the Wason selection task, by invoking factors such as shifting prior beliefs and goals (Oaksford and Chater 1994, 2007).

Ishita Dasgupta and colleagues (2022) test Chinchilla (Hoffmann et al. 2022) on several versions of the Wason selection task. They find across task versions that the model is no more than about 50% likely to take the logically correct action of turning the p and $\neg q$ cards, and in many conditions the model is at most 20% likely to do so (Figure 2). In particular, Dasgupta and colleagues find a significant tendency to turn the q card. As Dasgupta and colleagues note, these patterns of behavior conform in coarse outline to

the view also flows from more general descriptive evidence for Bayesian approaches to cognitive science. Normative evidence rests on normative arguments for the appropriateness for the Bayesian normative standard. My aim in this paper is not to provide new descriptive or normative evidence for the Bayesian view, but rather to suggest that those sympathetic to the view in human cognition should be at least as sympathetic to the same view about LLMs.

[17] The Shannon entropy of credence function P is $-\Sigma_{X=\{M_I,M_D\}}P(X)log_2(P(X))$. This definition enforces Oaksford and Chater's assumption that the agent has beliefs about the independence hypothesis M_I and dependence hypothesis M_D and aims to reduce her uncertainty about these hypotheses.

[18] The normative status of Oaksford and Chater's argument depends on the normative permissibility of interpreting indicative conditionals otherwise than as material conditionals, as well as the normative permissibility of other aspects of the Bayesian account.

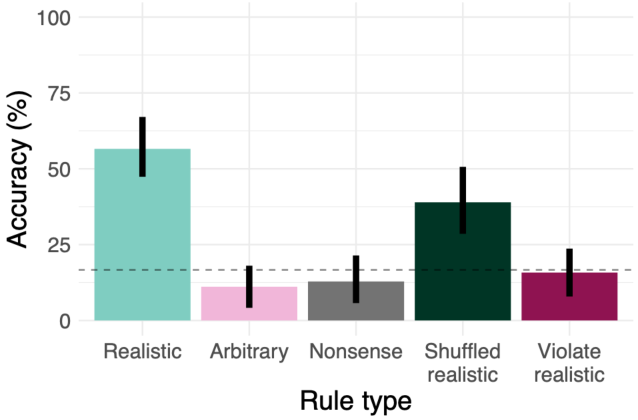


Figure 2: Wason selection task performance (logical criterion) by Chinchilla across rule types, from Dasgupta et al. (2022).

the predictions of Oaksford and Chater's probabilistic model but conform less well to the logical model.

3.3 Assessing bias

One way to react to these findings would be to say, as advocates of the logical paradigm do, that LLMs have interpreted the Wason selection task as a logical reasoning task and shown themselves to be poor logical reasoners. A second way to react to these findings would be to say, as Bayesians do, that LLMs have interpreted the task as a probabilistic reasoning task and shown themselves to be good probabilistic reasoners. To the extent that many theorists now favor a Bayesian account of how humans do and should respond to the Wason selection task, there is considerable pressure to adopt the same account when studying LLMs.

In fact, two lines of reflection may lead us to be more optimistic about the Bayesian story as applied to language agents. First, many critics of Bayesian cognitive science have questioned whether humans have the cognitive ability to perform the complex calculations needed to solve most everyday tasks as Bayesians suggest (Bowers and Davis 2012; Jones and Love 2011). Continuing this line, many philosophers have suggested that nonprobabilistic reasoning based on full beliefs rather than credences may play an important role in simplifying human reasoning (Holton 2008; Ross and Schroeder 2012; Staffel 2019). However, one of the defining turns in recent artificial intelligence work has been a turn away from older logic-based systems towards the construction of explicitly probabilistic, deep-learning systems (Ghahramani 2015). These systems are often thought to have sufficient computational power to carry out complex probabilistic calculations (Nafar et al. 2025; Paruchuri et al. 2024). If that is right, then it lessens the possibility of objecting to Bayesian accounts on the grounds of computational tractability.

A second reason for optimism concerns logical reasoning. On the orthodox Bayesian story, humans generally do and should interpret reasoning tasks probabilistically. This story provides good reason to expect that humans will be skilled probabilistic reasoners, but it also gives us some reason to doubt whether humans are good logical reasoners (Thorstad 2024b). If, as many Bayesians claim, logical reasoning is rarely useful and therefore rarely engaged in, there is no great reason to expect that learning or evolution will have endowed us with unbiased capacities for logical reasoning. Even if Bayesians are correct that humans respond well to a probabilistic construal of the Wason selection task, there thus remains the question of whether humans are capable of understanding and responding appropriately to a logical construal of the Wason selection task and other logical reasoning problems.

It remains controversial whether humans are capable of understanding and responding appropriately to a logical construal of the Wason selection task and other logical reasoning problems (Evans et al. 2003). But it should be less controversial whether Chinchilla can do this. Dasgupta and colleagues conduct five rounds of pretraining in which the LLM is rewarded for logical performance in the Wason selection task. Dasgupta and colleagues observe that performance shifts substantially after pretraining towards the predictions of a logical model. This suggests that Chinchilla is reasonably capable of quickly learning the logical interpretation of the Wason selection task and of responding in a logical manner.[19] Similarly, Dasgupta and colleagues find that five rounds of pretraining on a belief bias task in natural language inference nearly eliminates alleged knowledge effects. Again, this suggests that the LLM is capable of understanding and responding correctly to a logical interpretation of reasoning tasks, and not only to a default probabilistic interpretation of those tasks.

There is, of course, room for further discussion of many of these findings and room to present further findings. But there is, in general, no more reason to take LLMs' observed performance on the Wason selection task as evidence of biased LLM reasoning than there is to take similar performance as evidence of biased human reasoning, and there are some reasons to be more optimistic about the extent of LLM biases.

[19] It also suggests that Chinchilla was not already applying a logical interpretation.

4 Availability

If we are going to find uncontroversial cognitive biases in LLMs, we will need to look beyond allegations of knowledge effects. A natural place to start is by replicating classic biases from the heuristics and biases tradition. In this section and the next, I explore attempts to find two of the three initial biases proposed within this paradigm: availability bias and anchoring bias. I suggest that both attempts encounter significant obstacles, revealing important descriptive and normative lessons for future study.

4.1 Current research on availability

In the early 1970s, Daniel Kahneman and Amos Tversky proposed that humans often make inferences using the *availability heuristic* of "estimat[ing] frequency or probability by the ease with which instances or associations could be brought to mind" (Tversky and Kahneman 1973, p. 208). For example, participants presented with a list of 19 famous female actors and 20 less-famous male actors subsequently recalled the list as containing more female than male actors (Tversky and Kahneman 1973). A natural explanation for this finding invokes availability: because participants were more readily able to bring female actors to mind during subsequent recall, they judged that the list contained more female than male actors.

It is now almost universally acknowledged that early discussions of the availability heuristic passed too freely between two senses of availability (Schwartz et al. 2002). *Subjective availability* involves reliance on features of the subjective experience of recall, such as the felt ease or fluency with which information comes to mind. In this sense, agents may judge male actors to be rare if they strain and feel disfluency in trying to recall male actors. By contrast, *objective availability* involves reliance on the content of information retrieved or on non-experiential features of the retrieval process such as the time needed to retrieve information. In this sense, agents may judge male actors to be rare if they cannot recall many male actors, or if it takes a long time to recall male actors.

It is far from clear that reliance on objective availability of information is always a cognitive bias (Schwartz et al. 2002). If we can quickly bring many examples of a category to mind, then all else equal, that provides some evidence that the category is common in our experience, and hence in the world. It may still be irrational to make improper use of objectively available information, but this requires an argument that the information is being improperly used. This is important, because we will see that recent allegations of availability bias are naturally interpreted as involving objective availability without clear evidence of misuse.

Training Examples	10	100	1,000	10,000	25,296
Availability Bias Towards Negative Category (%)	26.3	77.7	39.7	47.0	52.0

Table 1: Availability bias in drug-drug interaction by size of training set, Lin and Ng (2023).

4.2 Availability in relation extraction

Relation extraction tasks involve identifying relationships between objects from textual discussions of those objects. A paradigmatic relation extraction task involves identifying drug-drug interactions (Zhang et al. 2020). Given a textual description of the interaction between two drugs, the algorithm must classify the type of interaction between them.

The Drug-Drug Interaction (DDI) dataset is an annotated corpus of 1,017 texts describing 5,021 interactions between various drugs (Segura-Bedmar et al. 2013). Each discussion is annotated with one of five interaction types: *mechanism* for a description of the interaction mechanism; *effect* for a description of the effect itself; *advice* for recommendations about how to respond to drug-drug interactions; *int* for nonspecific descriptions of interactions; and *negative* for non-interactions. The vast majority (85.2%) of interactions in the DDI dataset are negative, and LLMs trained on the DDI dataset understandably learn to reflect this fact.

Ruixi Lin and Hwee Tou Ng (2023) train GPT-3 on the DDI dataset. Lin and Ng then test the LLM on 'content-free' descriptions generated from the DDI dataset by replacing all medical terms with the dummy descriptor 'N/A'. Lin and Ng propose that because the LLM has no direct knowledge of the dummy class, the LLM should classify dummy sentences according to a uniform probability distribution. That is, it should be 20% likely to assign dummy descriptions to each interaction type: mechanism, effect, advice, int and negative.

Lin and Ng propose that any deviation from the uniform classification of dummy sentences should be treated as a form of availability bias, in which LLM judgments are skewed by the availability of interaction types in the training data. For each interaction type, Lin and Ng define the *availability bias score* of that interaction type to be the absolute difference between the percentage of test items classified under this type and the 20% classification rate expected under a uniform model. Under this definition, Lin and Ng find a strong availability bias, increasing in the number of descriptions used to train the LLM (Table 1).

Section 4.1 distinguished between two forms of availability: objective and subjective. Lin and Ng's experiment studies a form of objective availability: the content of information stored in training data. This is an especially benign form of objective availability, because we are concerned with the availability of *information* rather than with features of the information retrieval process, and we are concerned with the *total* information stored in memory rather than a potentially unrepresentative sample retrieved during decision-making. Reliance on objectively available information need not be irrational, so long as the information is relevant and the inferences drawn are supported by the available information. Alleging irrationality requires alleging that the information is irrelevant or used to draw unsupported inferences.

Lin and Ng hold that because the LLM has no specific information about the dummy descriptor 'N/A', "the best that an unbiased model can do is to make a uniform random guess" (Lin and Ng 2023). Traditional results in Bayesian epistemology suggest otherwise. Training on the DDI dataset provides the LLM with valuable information about the distribution of drug-drug interaction types. Rational Bayesian inference involves combining this prior information with novel information provided by descriptions to determine the probability that each given interaction type is at play. Since the LLM has been exposed to primarily negative interactions during training, the LLM correctly learns that negative interactions are more common than positive interactions and learns to project this relationship onto novel drugs. When the LLM is exposed to larger samples of training data, it rightly becomes more confident that negative interactions are prevalent. In the absence of competing information to move the LLM away from the prior, priors dominate and the LLM shows a strong tendency to predict novel drug-drug interactions to be negative, increasing in the quantity of training data. From an orthodox Bayesian standpoint, this is appropriate behavior and not a bias of any kind. If anything, Lin and Ng's data show under-reliance, rather than over-reliance, on prior knowledge of interaction types.

Lin and Ng do suggest one more plausible lesson from this discussion: labels matter. While many machine learning scientists expect label information to become unimportant after training, testing LLMs on content-free sentences reminds us of the importance of labels, since content-free sentences will be more likely to be classified using labels that are more frequent in the training data.[20] Following Johnson's taxonomy (Section 2.1) we might view this as a type of biased LLM inputs. However, precisely because the need to ensure unbiased data is familiar from previous research, we aimed in Section 2.1 to restrict attention to bias-constructs and bias-outputs, and we have not yet been given evidence for either of these.

[20] Tony Zhao and colleagues (2021) call this majority-label bias.

Moreover, it is not clear that Lin and Ng's solution of forcing a uniform distribution of classification on content-free sentences is the right way to reduce the influence of arbitrary labels. After all, there is considerable arbitrariness in the number of labels used in the training data. We could easily imagine the positive interactions being collapsed under a single label instead of four. Under a uniform distribution, this would increase the probability of negative predictions from 20% to 50%, a type of label-sensitivity that more traditional Bayesian methods avoid.

5 Heuristics and biases: Anchoring

5.1 Current research on anchoring

The second of Tversky and Kahneman's initial three heuristics is *anchoring and adjustment* (Tversky and Kahneman 1974). Suppose I ask you to estimate the year in which George Washington was first elected president. You might answer by *anchoring* on an initial quantity, the year (1776) in which the Revolutionary War began, then *adjusting* upwards and downwards to incorporate relevant knowledge, such as the length of the Revolutionary War and the drafting of the Constitution. If you are like most people, you might settle on an estimate around 1786.5 (Lieder et al. 2018), which is quite good: Washington was elected in 1789.

As this example illustrates, anchoring and adjustment produces a characteristic *anchoring effect* in which judgments are skewed towards the initial anchor. 1786.5 is quite close to the correct answer, but biased downwards towards the low anchor of 1776. Anchoring effects are traditionally explained as the result of insufficient adjustments away from the initial anchor.

Tversky and Kahneman (1974) initially proposed that a great number of anchoring effects should be explained as the result of mental processes of anchoring and adjustment. For example, Tversky and Kahneman instructed participants to spin a wheel, then judge whether the number displayed on the wheel was higher or lower than the number of African countries in the United Nations, and finally to estimate the number of African countries in the United Nations. Tversky and Kahneman found that judgments tended to be biased toward the value displayed on the wheel. Tversky and Kahneman explained this finding by assuming that agents anchored on an initial belief that the number of African countries in the United Nations is equal to the value on the wheel, then iteratively adjusted away from the anchor using a process of anchoring and adjustment.

That is a surprisingly irrational cognitive process, since there is no good reason to begin deliberation with the belief that the number of African countries in the United Nations is equal to the value on a spun wheel. Subsequent authors rightly asked for evidence that a process of iterative anchoring and adjustment had in fact been employed. For two decades, all available process-tracing studies showed no evidence of a cognitive process of anchoring and adjustment in this and other early experiments (Johnson and Schkade 1989; Lopes 1982).

More recently, evidence has emerged that a genuine process of anchoring and adjustment may be employed in a small number of examples, such as our initial example of estimating the year in which George Washington was first elected president (Epley and Gilovich 2006; Lieder et al. 2018). However, most research programs agree that genuine anchoring and adjustment is extremely rare; that anchoring and adjustment is not typically triggered by external manipulations such as spinning wheels; that anchors tend to be relevant and informative, and incorporated in a rational way; that the results of anchoring and adjustment are often highly reliable; and that few if any anchoring effects in the early literature are produced by genuine processes of anchoring and adjustment (Epley and Gilovich 2001, 2004, 2006; Lieder et al. 2018, ms).[21]

As evidence for a process of anchoring and adjustment failed to materialize in the motivating examples, researchers broadened the concept of anchoring effects so that they were no longer conceptually tied to a process of anchoring and adjustment. We saw in Section 2 that this movement coincides with a broader trend in the heuristics and biases program towards a normative conception of bias on which biases can be defined and studied independently of any particular heuristic process. We also saw in Section 2 that this emancipation has led to conceptual ambiguity in bias definitions, and indeed Kahneman himself concedes that:

> The terms *anchor* and *anchoring effect* have been used in the psychological literature to cover a bewildering array of diverse experimental manipulations and results … The proliferation of meanings is a serious hindrance to theoretical progress. (Jacowitz and Kahneman 1995, p. 1161)

One reaction to this definitional ambiguity would be to move beyond attempts to posit and study a unified type of anchoring bias. I have some sympathy for this response. However, even without settling on a precise definition it may be possible to say enough about plausible normative conceptions of anchoring bias to assess whether recent experimental results provide strong evidence for anchoring bias in LLMs.

[21] Note that this evidence applies only to heuristic processes of anchoring and adjustment. Broader studies of anchoring (Schley and Weingarten 2025) will need to be addressed separately.

Here is a sampling of recent definitions of anchoring effects, understood as normative biases that may come apart from the heuristic process of anchoring and adjustment:

> An anchor is an arbitrary value that the subject is caused to consider before making a numerical estimate. An anchoring effect is demonstrated by showing that the estimates of groups shown different anchors tend to remain close to those anchors. (Jacowitz and Kahneman 1995, p. 1161)

> The anchoring effect is the disproportionate influence on decision makers to make judgments that are biased toward an initially presented value. (Furnham and Chu Boo 2011, p. 35)

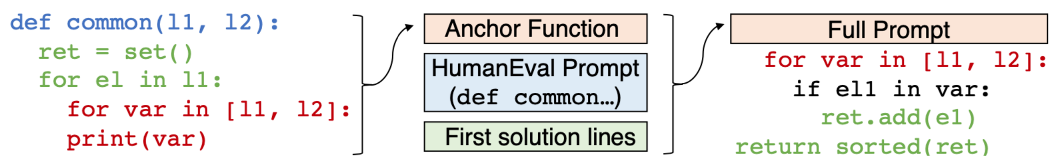
An important feature of these definitions is that a normative anchoring bias involves mis-use of information contained in the anchor. Anchors must either be arbitrary (Jacowitz and Kahneman 1995) and hence unsuitable for use in future inference, or else must exert disproportionate influence (Furnham and Chu Boo 2011) on future inference. Mere reliance on anchor information is not thought to constitute a normative anchoring bias. To take a widely-cited example, manipulating the listing prices of properties changes what agents are willing to pay for them (Northcraft and Neale 1987). But there is nothing wrong with that, since listing prices carry information about property values. To say otherwise would be to confuse anchoring bias with the simple process of learning from evidence. On this basis, it is generally agreed that a normative concept of anchoring bias must involve mis-use of anchoring information beyond its evidential relevance.[22]

5.2 Anchoring in code generation

Code generation tasks involve generating code from prompts. Prompts may be partial programs, English descriptions of desired functionality, or combinations of these and other inputs. Existing code generation models include OpenAI's Codex (Chen et al. 2021) and Salesforce's CodeGen (Nijkamp et al. 2023).

The HumanEval dataset is often used to assess code generation (Chen et al. 2021). HumanEval is composed of 164 programming problems. Each problem contains a three-part prompt: a function signature 'def function_name', an English description of the desired functionality, and several input-output pairs describing correct function behavior.

[22] Note that the limited success of attempts to 'debias' anchoring effects does not challenge a view on which those effects were not biases in the first place. In fact, it weakly supports the view by providing one reason why debiasing interventions might fail, namely that debiasing interventions attempt to push agents in the wrong direction.



```
def common(l1, l2):
  ret = set()
  for el in l1:
    for var in [l1, l2]:
    print(var)
```

```
  for var in [l1, l2]:
    if el1 in var:
      ret.add(el)
return sorted(ret)
```

Figure 3: Construction of anchor function and full prompt, from Jones and Steinhardt (2022).

Each problem is also accompanied by a canonical solution: a correct solution program generated by human programmers.

Erik Jones and Jacob Steinhardt (2022) aim to find anchoring bias in code generation by Codex and CodeGen. They do this by incorporating tempting, but incorrect solutions into 'anchor' strings, then prepending anchoring strings to complete HumanEval prompts.

More concretely, Jones and Steinhardt construct anchor functions with three parts (Figure 3). The first part is the function signature, copied from the HumanEval prompt. The second part is the first n lines of the canonical solution, with n varied between 0 and 8 across prompts. The final part is a set of 'anchor lines' describing a tempting but incorrect partial solution.

Jones and Steinhardt consider two types of anchors. *Print-var* anchors instruct the program to print, rather than return, a given value:

```
for var in [var1, var 2]:
    print(var)
```

Add-var anchor lines instruct programs to sum two values:

```
tmp = str(var1) + str(var2)
    return tmp
```

Complete anchor functions consist of a function signature, the first n lines of the canonical solution, and the chosen anchor lines. Total prompts are constructed by prepending anchor lines to the original HumanEval prompt, consisting of a function signature, an English description of the desired functionality, and example input-output pairs (Figure 3). These are again followed by the first n lines of the canonical solution, with n fixed at its value in the anchor function.

Jones and Steinhardt test Codex and CodeGen across a variety of total prompts, varying the choice of anchor lines, the number n of canonical solution lines, and the original prompt from HumanEval. They find a significant decrease in accuracy, as well as an increased tendency for solutions by Codex and CodeGen to incorporate anchor lines in part or full within the resulting outputs. Jones and Steinhardt treat this finding as an anchoring effect, in which "code models . . . adjust their output towards related solutions, where these solutions are included in the prompt" (Jones and Steinhardt 2022).

5.3 Discussion

The discussion in Section 5.1 suggests three challenges for Jones and Steinhardt's anchoring experiment. First, Jones and Steinhardt sometimes talk as though they have found processes of *adjustment* away from an anchor.[23] However, no evidence for a process of anchoring and adjustment has been provided – indeed, we are not given process-tracing evidence of any kind. We saw in Section 5.1 that it is important to support procedural claims with process-tracing evidence, because most previous instances in which a heuristic process of anchoring and adjustment was postulated to explain anchoring biases, this postulate turned out to be wrong. Without process-tracing evidence, we should therefore focus on the charge of anchoring bias and not on related questions about the process of anchoring and adjustment.

Second, the anchors provided by Jones and Steinhardt are relevant, not irrelevant. They are highly similar in content to the problem and constructed to be similar to correct solutions. Indeed, the anchors explicitly contain an initial segment of the canonical solution. This makes the anchors relevant to, and informative about the problem at hand. As we have seen, the bare reliance on relevant information cannot be taken to constitute a normative anchoring bias. We may still allege that LLMs have over-used relevant anchors, just as we may criticize them for over-reliance on any item of evidence. However, pressing this charge requires proving over-use, which Jones and Steinhardt do not attempt to do.

Third, even if the anchors provided by Jones and Steinhardt were not in fact relevant, there would nonetheless be a legitimate presupposition of relevance. Codex and CodeGen were trained primarily on helpful and non-misleading prompts. While the LLMs may have been exposed to natural human errors, they have not been significantly exposed to programmers trying to manipulate them into including irrelevant code in their outputs. On the basis of this experience, any rational agent would learn that input is likely to be non-manipulative. Codex and CodeGen do not, and should not, treat inputs as likely to be manipulative unless they are trained on manipulative examples, any more than readers of

[23] For example: "Using anchoring as inspiration, we hypothesize that code generation models may adjust their output towards related solutions" and "We additionally find that elements of anchor function[s] often appear in both models' outputs, suggesting that code generation models adjust their solutions towards related solutions" (Jones and Steinhardt 2022).

this paper should expect that I have subtly lied about all studies reported therein, unless given evidence that research papers frequently lie in this way.

6 Philosophical implications

This paper assessed recent charges of cognitive bias in LLMs. Section 2 developed a normative, surrogate conception of cognitive bias on which cognitive biases in LLMs are understood as states, processes or judgments that would violate a normative standard if exhibited by humans in similar tasks. Section 3 looked at knowledge effects, in which prior knowledge exhibits a robust inference on reasoning that goes beyond the requirements of classical logic. Focusing on the Wason selection task, we considered an increasingly popular Bayesian story on which knowledge effects reflect the normatively correct influence of prior knowledge on probabilistic reasoning. We saw that this story should be at least as plausible when applied to LLMs as it is when applied to humans, and perhaps more so, since LLMs are better-equipped for complex probabilistic reasoning than humans are, and since LLMs' probabilistic reasoning capacities do not appear to come at the expense of diminished capability for logical reasoning.

Section 4 considered availability bias in relation extraction, focusing on identification of drug-drug interactions. We saw that a purported availability bias to judge that unknown drugs resemble drugs encountered during training is not best understood as a cognitive bias, but rather as a justifiable reliance on prior knowledge. Section 5 considered anchoring bias in code generation. We saw that LLM outputs provide no evidence that LLMs employ a heuristic process of anchoring and adjustment. We also saw that the anchors on which LLMs draw are both relevant and justifiably presumed to be relevant, again suggesting that observed behavior cannot be taken as evidence of cognitive bias without further evidence of inappropriate reasoning.

One way to take these results is as a qualified piece of good news. After a torrent of negative findings on algorithmic bias, it is refreshing to be reminded that there are some bias metrics on which LLMs perform fairly well.[24] This reaction is complemented and enriched by three further philosophical implications of the results in this paper.

[24] Importantly, even if we think that LLMs exhibit severe and pervasive moral biases, the findings of this paper suggest that they may exhibit less severe cognitive biases of many types recently alleged.

6.1 Beyond bias

Many theorists have suggested that assessments of human cognition place too much focus on cognitive bias and not enough focus on competing metrics (Gigerenzer and Brighton 2009; Schurz and Hertwig 2019; Sturm 2019). We care not only whether our cognition is biased, but also whether it is accurate, fast, efficient, capable of carrying out complex calculations, and resilient against small shocks or changes to the environment.

All of these metrics figure in famous trade-offs. Accuracy confronts an *accuracy-coherence* tradeoff in which increased accuracy may come at the expense of decreased coherence, which on many accounts will register as a type of cognitive bias (Thorstad 2024a). Speed confronts a *speed-accuracy* tradeoff where rapid computation may diminish accuracy (Heitz 2014). Efficiency confronts an *accuracy-effort* tradeoff in which decreased effort may come at the cost of reduced accuracy (Johnson and Payne 1985). Complexity confronts a *complexity-coherence* tradeoff in which complex operations may increase the risk of incoherence (Thorstad 2025). Resilience to environmental shocks confronts a *bias-variance tradeoff* in which increased bias may reduce shock vulnerability by decreasing model variance and thereby reducing the risk of overfitting to artificial features of task environments (Geman et al. 1992).[25]

To the extent that we care about these metrics, we should not take any particular form of bias as a definitive assessment of LLM performance. It may well be the case that LLMs, even while exhibiting biases, are also accurate, fast, efficient, capable of carrying out complex calculations and resilient. Vindicatory epistemologists have stressed, for this reason, that it is important to measure performance along multiple metrics beyond any single alleged bias (Berg 2003; Schurz and Hertwig 2019; Thorstad 2024b). If it turns out that LLMs exhibiting any proposed bias are, at the same time, performing well on many important metrics, then that should increase our opinion of their performance.[26]

Vindicatory epistemologists have also stressed the importance of assessing performance in specific environments (Morton 2017; Schmidt 2019; Todd and Gigerenzer 2012). What we want to know is not how LLMs perform in an artificial laboratory setting, but how they perform in the contexts where they are proposed for use. While laboratory results reveal that LLMs are in principle capable of exhibiting bias, we do not yet learn whether there is a significant threat of biased performance in practice. More generally, vindicatory epistemologists have stressed that all models perform well in some environ-

[25] This focus on modal resilience builds on work by Munton (2019) and Rudolph et al. (2025).

[26] To claim that our opinion of systems should be improved when we learn that they perform well on important metrics is not to claim that our opinion of those systems should not also be worsened by their performance on other important metrics such as racial bias. See also Byrd (2025).

ments and poorly in others (Thorstad 2024b; Todd and Gigerenzer 2012). For this reason, our aim should not be to impugn LLMs by exhibiting environments in which they perform poorly, but instead to get a better understanding of which environments the LLMs are best-suited for, so they can be recommended for use in appropriate environments.

Finally, focusing on competing metrics which may come into tension with the attempt to avoid any specific bias provides a new perspective on how LLM performance might be improved. A traditional *nudging* perspective suggests that fundamentally irrational processes might be coaxed into performing better through subtle manipulations (Bovens 2009; Sunstein 2014; Thaler and Sunstein 2008). By contrast, a more vindicatory *boosting* perspective suggests that to the extent that biased behavior is driven by problem constraints such as limited data, scarce computational power, or hostile environments, we might see better performance increases by lessening these limitations (Grüne-Yanoff and Hertwig 2016; Hertwig and Grüne-Yanoff 2017). The boosting perspective therefore lends support to existing approaches which stress the need for high-quality representative data (Buolamwini and Gebru 2018) and the benefits of increased computation (Kaplan et al. 2020; Sutton 2019).

6.2 Vindicatory epistemology

The program of vindicatory epistemology seeks to vindicate the rationality of purported cognitive biases through a combination of normative theorizing and empirical reassessment of apparently biased cognitions (Dorst 2023; Icard ms; Thorstad 2024b). The discussion in this paper advances the program of vindicatory epistemology in two ways.

First, vindicatory epistemologists claim that a number of strategies can be used to explain away a wide range of purported biases. These strategies include conceptual clarification of bias concepts (Dreisbach and Guevara 2019; Schwartz et al. 2002; Thorstad 2024b), challenges to the normative standards underlying biases (Bermúdez 2020; Gigerenzer 1996; Sturm 2019), probabilistic reconstruals of seemingly-nonprobabilistic reasoning tasks (Fitelson 2010; Icard 2021; Oaksford and Chater 2007), and an ecological perspective on which cognition is assessed against its performance in an environment (Morton 2017; Schmidt 2019; Todd and Gigerenzer 2012). To motivate the programmatic claim that these strategies can be used to explain away a wide range of biases, it is necessary to exhibit a range of case studies in which the strategies plausibly succeed in defusing bias allegations.

The case studies in this paper help to extend the reach of familiar vindicatory strategies. We saw in Sections 4-5 how conceptual distinctions between objective and subjective availability, as well as between anchoring bias and the heuristic process of anchoring and adjustment can be used to clarify and soften bias allegations. We saw in Section 4 how a normative rejection of uniform guessing in the presence of relevant prior information reveals alleged availability biases to rest on an inappropriate normative standard. We saw in Section 3 how a probabilistic task construal could be used to vindicate model performance on the Wason selection task. And we saw in Section 5 how a failure to detect and respond to deliberately misleading prompts may not impugn model performance in many of the environments where they are proposed for use.

A second avenue of support for vindicatory epistemology comes from an observation by Dasgupta and colleagues (2022). Vindicatory epistemologists predict that rational pressures lead agents confronted with a given problem to exhibit apparent cognitive biases. Vindicatory epistemologists therefore suspect that other sophisticated agents, confronted with the same problem, will exhibit the same apparent bias. As Dasgupta and colleagues note, the emergence of an apparent bias in multiple agents with very different cognitive architectures lends support to the rationalizing explanation on which the bias is merely apparent. After all, it would be somewhat surprising if radically different agents were led to make similar mistakes, but less surprising for radically different agents to be led through rational pressures towards correct solutions.

6.3 Further biases

Recent authors have alleged a number of other cognitive biases in LLMs, including base rate neglect (Talboy and Fuller 2023), certainty effects (Itzhak et al. 2024), confirmation bias (Talboy and Fuller 2023), egocentric bias (Koo et al. 2024) framing effects (Binz and Schulz 2023; Jones and Steinhardt 2022), and recency bias (Schmidgall et al. 2024). In this paper, we saw that familiar vindicatory strategies are well-suited to defusing three recent bias allegations. However, this does not imply that vindicatory strategies will defuse all bias allegations.

Daniel Kahneman once complained that vindicatory epistemologists see only two kinds of errors: "pardonable errors by subjects and unpardonable ones by psychologists" (Kahneman 1981, p. 340). Keith Stanovich and Richard West (2000) accused vindicatory epistemologists of taking the Panglossian stance that we live in the best of all possible worlds, in which all biases are merely apparent. Vindicatory epistemologists are not Panglossians. While the case studies in this paper provide room for optimism, they do not relieve us of the burden of assessing further bias allegations. Future work should take a careful look at remaining bias allegations in order to understand which biases LLMs exhibit, how often they are exhibited, and what can be done to ameliorate them.

References

Adams, Ernest. 1975. *The logic of conditionals: An application of probability to deductive logic*. Synthese Library.

Anderson, John. 1990. *The adaptive character of thought*. Lawrence Erlbaum Associates.

Bender, Emily and Koller, Alexander. 2020. "Climbing towards NLU: On meaning, form, and understanding in the age of data." *Proceedings of the 58th Annual Meeting of the Assocation for Computational Linguistics* 5185–98.

Berg, Nathan. 2003. "Normative behavioral economics." *Journal of Socio-Economics* 32:411–27.

Bermúdez, José. 2020. *Frame it again*. Cambridge University Press.

Binz, Marcel and Schulz, Eric. 2023. "Using cognitive psychology to understand GPT-3." *Proceedings of the National Academy of Sciences* 120:e2218523120.

Bovens, Luc. 2009. "The ethics of *Nudge*." In Till Grüne-Yanoff and Sven Ove Hansson (eds.), *Preference change*, 207–19. Springer.

Bowers, Jeffrey and Davis, Colin. 2012. "Bayesian just-so stories in psychology and neuroscience." *Psychological Bulletin* 138:389–414.

Buolamwini, Joy and Gebru, Timnit. 2018. "Gender shades: Intersectional accuracy disparities in commercial gender classification." *Proceedings of Machine Learning Research* 81:1–15.

Byrd, Nick. 2025. "Strategic reflectivism in intelligent systems." arXiv 2505.22987.

Carnap, Rudolf. 1950. *Logical foundations of probability*. University of Chicago Press.

Chater, Nick, Tenenbaum, Joshua, and Yuille, Alan. 2006. "Probabilistic models of cognition: Conceptual foundations." *Trends in Cognitive Sciences* 10:287–91.

Chen, Mark, Tworek, Jerry, Jun, Heewoo, Yuan, Qiming, de Oliveira Pinto, Henrique Ponde, Kaplan, Jared, Edwards, Harri, Burda, Yuri, Joseph, Nicholas, Brockman, Greg, Ray, Alex, Puri, Raul, Krueger, Gretchen, Petrov, Michael, Khlaaf, Heidy, Sastry, Girish, Mishkin, Pamela, Chan, Brooke, Gray, Scott, Ryder, Nick, Pavlov, Mikhail, Power, Alethea, Kaiser, Lukasz, Bavarian, Mohammad, Winter, Clemens, Tillet, Philippe, Such, Felipe Petroski, Cummings, Dave, Plappert, Matthias, Chantzis, Fotios, Barnes, Elizabeth, Herbert-Voss, Ariel, Guss, William Hebgen, Nichol, Alex, Paino, Alex, Tezak, Nikolas, Tang, Jie, Babuschkin, Igor, Balaji, Suchir, Jain, Shantanu, Saunders, William, Hesse, Christopher, Carr, Andrew N., Leike, Jan, Achiam, Josh, Misra, Vedant, Morikawa, Evan, Radford, Alec, Knight, Matthew, Brundage, Miles, Murati, Mira, Mayer, Katie, Welinder, Peter, McGrew, Bob, Amodei, Dario, McCandlish, Sam, Sutskever, Ilya, and Zaremba, Wojciech. 2021. "Evaluating large language models trained on code." arXiv 2107.03374.

Creel, Kathleen. 2020. "Transparency in complex computational systems." *Philosophy of Science* 87:568–89.

Dasgupta, Ishita, Lampinen, Andrew K., Chan, Stephanie C. Y., Creswell, Antonia, Kumaran, Dharshan, McClelland, James L., and Hill, Felix. 2022. "Language models show human-like content effects on reasoning." arXiv, 2207.07051.

Dorst, Kevin. 2023. "Rational polarization." *Philosophical Review* 132:355–458.

Dreisbach, Sandra and Guevara, Daniel. 2019. "The Asian disease problem and the ethical implications of prospect theory." *Noûs* 53:613–38.

Dziri, Nouha, Lu, Ximing, Sclar, Melanie, Li, Xiang Lorraine, Jiang, Liwei, Lin, Bill Yuchen, West, Peter, Bhagavatula, Chandra, Bras, Ronan Le, Hwang, Jena D., Sanyal, Soumya, Welleck, Sean, Ren, Xiang, Ettinger, Allyson, Harchaoui, Zaid, and Choi, Yejin. 2023. "Faith and fate: Limits of transformers on compositionality." *Proceedings of the 37th International Conference on Neural Information Processing Systems* 70293–332.

Elqayam, Shira and Over, David. 2013. "New paradigm psychology of reasoning: An introduction to the special issue edited by Elqayam, Bonnefon, and Over." *Thinking and Reasoning* 19:249–65.

Epley, Nicholas and Gilovich, Thomas. 2001. "Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors." *Psychological Science* 12:391–6.

—. 2004. "Are adjustments insufficient?" *Personality and Social Psychology Bulletin* 30:447–60.

—. 2006. "The anchoring-and-adjustment heuristic: Why the adjustments are insufficient." *Psychological Science* 17:311–8.

Erk, Katrin. 2022. "The probabilistic turn in semantics and pragmatics." *Annual Review of Linguistics* 8:101–21.

Evans, Jonathan, Barston, Julie, and Pollard, Paul. 1983. "On the conflict between logic and belief in syllogistic reasoning." *Memory and Cognition* 11:295–306.

Evans, Jonathan, Handley, Simon, and Over, David. 2003. "Conditionals and conditional probability." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29:321–35.

Fazelpour, Sina and Danks, David. 2021. "Algorithmic bias: Senses, sources, solutions." *Philosophy Compass* 16:e12760.

Fitelson, Branden. 2010. "The Wason task(s) and the paradox of confirmation." *Philosophical Perspectives* 24:207–41.

Floridi, Luciano. 2023. "AI as *agency without intelligence*: On ChatGPT, large language models, and other generative models." *Philosophy and Technology* 36:1–7.

Furnham, Adrian and Chu Boo, Hua. 2011. "A literature review of the anchoring effect." *Journal of Socio-Economics* 40:35–42.

Gawronski, Bertram, Hofmann, Wilhelm, and Wilbur, Christopher. 2006. "Are "implicit" attitudes unconscious?" *Consciousness and Cognition* 15:485–99.

Geman, Stuart, Bienenstock, Elie, and Doursat, René. 1992. "Neural networks and the bias/variance dilemma." *Neural Computation* 4:1–58. doi:10.1162/neco.1992.4.1.1.

Ghahramani, Zoubin. 2015. "Probabilistic machine learning and artificial intelligence." *Nature* 521:452–9.

Gigerenzer, Gerd. 1996. "On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1986)." *Psychological Review* 103:592–6.

—. 2019. "Axiomatic rationality and ecological rationality." *Synthese* 194:3547–64. doi: 10.1007/s11229-019-02296-5.

Gigerenzer, Gerd and Brighton, Henry. 2009. "Homo heuristicus: Why biased minds make better inferences." *Topics in Cognitive Science* 1:107–43. doi:10.1111/j.1756-8765.2008.01006.x.

Gilovich, Thomas and Griffin, Dale. 2002. “Heuristics and biases: Then and now.” In Thomas Gilovich, Dale Griffin, and Daniel Kahneman (eds.), *Heuristics and biases: The psychology of intuitive judgment*, 1–18. Cambridge University Press.

Gilovich, Thomas, Griffin, Dale, and Kahneman, Daniel (eds.). 2002. *Heuristics and biases: The psychology of intuitive judgment*. Cambridge University Press.

Gilovich, Thomas, Vallone, Robert, and Tversky, Amos. 1985. “The hot hand in basketball: on the misperception of random sequences.” *Cognitive Psychology* 17:325–31.

Grüne-Yanoff, Till and Hertwig, Ralph. 2016. “Nudge versus boost: How coherent are policy and theory?” *Minds and Machines* 26:149–83.

Heitz, Richard. 2014. “The speed-accuracy tradeoff: History, physiology, methodology, and behavior.” *Frontiers in Neuroscience* 8:1–19. doi:10.3389/fnins.2014.00150.

Hertwig, Ralph and Grüne-Yanoff, Till. 2017. “Nudging and boosting: Steering or empowering good decisions.” *Perspectives on Psychological Science* 12:973–86.

Hoffmann, Jordan, Borgeaud, Sebastian, Mensch, Arthur, Buchatskaya, Elena, Cai, Trevor, Rutherford, Eliza, de Las Casas, Diego, Hendricks, Lisa Anne, Welbl, Johannes, Clark, Aidan, Hennigan, Tom, Noland, Eric, Millican, Katie, van den Driessche, George, Damoc, Bogdan, Guy, Aurelia, Osindero, Simon, Simonyan, Karen, Elsen, Erich, Rae, Jack W., Vinyals, Oriol, and Sifre, Laurent. 2022. “Training compute-optimal large language models.” arXiv 2203.15556.

Holton, Richard. 2008. “Partial belief, partial intention.” *Mind* 117:27–58.

Icard, Thomas. 2021. “Why be random?” *Mind* 130:111–39.

—. ms. *Resource rationality*.

Itzhak, Itay, Stanovsky, Gabriel, Rosenfeld, Nir, and Belinkov, Yonatan. 2024. “Instructed to bias: Instruction-tuned language models exhibit emergent cognitive bias.” *Transactions of the Association for Computational Linguistics* 12:771–85.

Jacowitz, Karen and Kahneman, Daniel. 1995. “Measures of anchoring in estimation tasks.” *Personality and Social Psychology Bulletin* 21:1161–6.

Janis, Irving Lester and Frick, Frederick. 1943. “The relationship between attitudes toward conclusions and errors in judging logical validity of syllogisms.” *Journal of Experimental Psychology* 33:73–7.

Johnson, Eric and Payne, John. 1985. “Effort and accuracy in choice.” *Management Science* 31:395–414. doi:10.1287/mnsc.31.4.395.

Johnson, Eric and Schkade, David. 1989. “Bias in utility assessments: Further evidence and explanations.” *Management Science* 35:406–24.

Johnson, Gabbrielle. 2020. “The structure of bias.” *Mind* 129:1193–1236.

—. 2021. “Algorithmic bias: On the implicit biases of social technology.” *Synthese* 198:9941–61.

—. 2023a. “Are algorithms value-free? Feminist theoretical virtues in machine learning.” *Journal of Moral Philosophy* 21:1–35.

—. 2023b. “Unconscious perception and unconscious bias: Parallel debates about unconscious content.” *Oxford Studies in Philosophy of Mind* 3:87–130.

—. 2024a. “The (dis)unity of psychological (social) bias.” *Philosophical Psychology* 37:1349–77.

—. 2024b. “Varieties of bias.” *Philosophy Compass* e13011.

Jones, Erik and Steinhardt, Jacob. 2022. “Capturing failures of large language models via human cognitive biases.” In *NIPS’22: Proceedings of the 36th International Conference on Neural Information Processing Systems*, 11785–99.

Jones, Matt and Love, Bradley. 2011. “Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition.” *Behavioral and Brain Sciences* 34:169–231.

Kahneman, Daniel. 1981. “Who shall be the arbiter of our intuitions?” *Behavioral and Brain Sciences* 4:339–40.

—. 2003. “Maps of bounded rationality: Psychology for behavioral economics.” *American Economic Review* 93:1449–75.

Kahneman, Daniel, Slovic, Paul, and Tversky, Amos (eds.). 1982. *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

Kaplan, Jared et al. 2020. “Scaling laws for neural language models.” arXiv 2001.08361, https://arxiv.org/pdf/2001.08361.pdf.

Kelly, Thomas. 2023. *Bias: A philosophical study*. Oxford University Press.

—. 2024. "Bias, norms, introspection, and the bias blind spot." *Philosophy and Phenomenological Research* 108:81–105.

Kirby, Kris. 1994. "Probabilities and utilities of fictional outcomes in Wason's four-card selection task." *Cognition* 51:1–28.

Kojima, Takeshi, Shane Gu, Shixiang, Reid, Machel, Yutaka, Matsuo, and Iwasawa, Yusuke. 2022. "Large language models are zero-shot reasoners." *Proceedings of the 36th International Conference on Neural Information Processing Systems* 35:22199–213.

Koo, Ryan, Lee, Minhwa, Raheja, Vipul, Park, Jonginn, Kim, Zae Myung, and Kang, Dongyeop. 2024. "Benchmarking cognitive biases in large language models as evaluators." *Findings of the Association for Computational Linguistics* 517–545.

Lieder, Falk, Griffiths, Thomas, Huys, Quentin, and Goodman, Noah. 2018. "The anchoring bias reflects rational use of cognitive resources." *Psychonomic Bulletin and Review* 25:322–49.

—. ms. "Testing models of anchoring and adjustment." Manuscript.

Lin, Ruixi and Ng, Hwee Tou. 2023. "Mind the biases: Quantifying cognitive biases in language model prompting." In *Findings of the Association for Computational Linguistics: ACL 2023*, 5269–81. Toronto, Canada: Association for Computational Linguistics.

Lopes, Lola. 1982. "Toward a procedural theory of judgment." Office of Naval Research Final Report.

Mitchell, Melanie. 2025. "Artificial intelligence learns to reason." *Science* 387:eadw5211.

Moosavi, Parisa. 2024. "Will intelligent machines become moral patients?" *Philosophy and Phenomenological Research* 109:95–116.

Morton, Jennifer. 2017. "Reasoning under scarcity." *Australasian Journal of Philosophy* 95:543–59. doi:10.1080/00048402.2016.1236139.

Mosakas, Kestutis. 2021. "On the moral status of social robots: Considering the consciousness criterion." *AI and Society* 36:429–43.

Müller, Vincent. 2021. "Is it time for robot rights? Moral status in artificial entities." *Ethics and Information Technology* 23:579–87.

Munton, Jessie. 2019. "Beyond accuracy: Epistemic flaws with statistical generalizations." *Philosophical Issues* 29:228–40.

Nafar, Aliakbar, Venable, Kristen Brent, Cui, Zijun, and Kordjamshidi, Parisa. 2025. "Extracting probabilistic knowledge from large language models for Bayesian network parameterization." arXiv 2505.15918.

Nijkamp, Erik, Pang, Bo, Hayashi, Hiroaki, Tu, Lifu, Wang, Huan, Zhou, Yingbo, Savarese, Silvio, and Xiong, Caiming. 2023. "CodeGen: An open large language model for code with multi-turn program synthesis." arXiv 2203.13474.

Northcraft, Gregory and Neale, Margaret. 1987. "Experts, amateurs, and real estate: an anchoring-and-adjustment perspective on property pricing decisions." *Organizational Behavior and Human Decision Processes* 39:84–97.

Oaksford, Mike and Chater, Nick. 1994. "A rational analysis of the selection task as optimal data selection." *Psychological Review* 101:608–31.

—. 2007. *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Oaksford, Mike, Chater, Nick, Grainger, Becki, and Larkin, Joanne. 1997. "Optimal data seleciton in the reduced array selection task (RAST)." *Journal of Experimental Psychology* 23:441–58.

Paruchuri, Akshay, Garrison, Jake, Liao, Shun, Hernandez, John, Sunshine, Jacob, Althoff, Tim, Liu, Xin, and McDuff, Daniel. 2024. "What are the odds? Language models are capable of probabilistic reasoning." arXiv 2406.12830.

Rips, Lance. 1994. *The psychology of proof: Deductive reasoning in human thinking*. MIT Press.

Ross, Jacob and Schroeder, Mark. 2012. "Belief, credence and pragmatic encroachment." *Philosophy and Phenomenological Research* 88:259–88.

Rudolph, Rachel Etta, Shech, Elay, and Tamir, Michael. 2025. "Bias, machine learning, and conceptual engineering." *Philosophical Studies* 182:1889–1917.

Schley, Dan and Weingarten, Evan. 2025. "50 years of anchoring: A meta-analysis and meta-study of anchoring effects." ms.

Schmidgall, Samuel, Harris, Carl, Essien, Ime, Olshvang, Daniel, Rahman, Tawsifur, Kim, Ji Woong, Ziaei, Roijin, Eshraghian, Jason, Abadir, Peter, and Chellappa, Rama. 2024. "Evaluation and mitigaiton of cognitive biases in medical language models." *npj Digital Medicine* 7:295.

Schmidt, Andreas. 2019. "Getting real on rationality – behavioral science, nudging, and public policy." *Ethics* 129:511–543. doi:10.1086/702970.

Schurz, Gerhard and Hertwig, Ralph. 2019. "Cognitive success: A consequentialist account of rationality in cognition." *Topics in Cognitive Science* 11:7–36.

Schwartz, Barry, Ward, Andrew, Monterosso, John, Lyubomirsky, Sonja, White, Katherine, and Lehman, Darrin R. 2002. "Maximizing versus satisficing: Happiness is a matter of choice." *Journal of Personality and Social Psychology* 83:1178–1197.

Segura-Bedmar, Isabel, Martínez, Paloma, and Herrero-Zazo, María. 2013. "SemEval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013)." In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 341–350. Atlanta, Georgia, USA: Association for Computational Linguistics.

Shannon, Claude. 1948. "A mathematical theory of communication." *The Bell System Technical Journal* 27:379–423.

Staffel, Julia. 2019. "How do beliefs simplify reasoning?" *Noûs* 53:937–62.

Stanovich, Keith and West, Richard. 2000. "Individual differences in reasoning: Implications for the rationality debate?" *Behavioral and Brain Sciences* 23:645–65. doi: 10.1017/s0140525x00003435.

Sturm, Thomas. 2019. "Formal versus bounded norms in the psychology of rationality: Toward a multilevel analysis of their relationship." *Philosophy of the Social Sciences* 49:190–209.

Sunstein, Cass. 2014. *Why nudge? The politics of libertarian paternalism*. Yale University Press.

Sutton, Rich. 2019. "The bitter lesson." http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

Talboy, Alaina N. and Fuller, Elizabeth. 2023. "Challenging the appearance of machine intelligence: Cognitive bias in LLMs and Best Practices for Adoption." arXiv 2304.01358.

Thaler, Richard and Sunstein, Cass. 2008. *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Thorstad, David. 2024a. “The accuracy-coherence tradeoff in cognition.” *British Journal for the Philosophy of Science* 75:695–715.

—. 2024b. *Inquiry under bounds*. Oxford University Press.

—. 2025. “The complexity-coherence tradeoff in cognition.” *Mind* 134:422–57. doi:10.1093/mind/fzaf015.

Todd, Peter and Gigerenzer, Gerd. 2012. *Ecological rationality: Intelligence in the world*. Oxford University Press.

Tversky, Amos and Kahneman, Daniel. 1973. “Availability: A heuristic for judging frequency and probability.” *Cognitive Psychology* 5:207–32.

—. 1974. “Judgment under uncertainty: Heuristics and biases.” *Science* 185:1124–31. doi:10.1126/science.185.4157.1124.

—. 1983. “Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment.” *Psychological Review* 90:293–315.

Vredenburgh, Kate. 2022. “The right to explanation.” *Journal of Political Philosophy* 30:209–29.

Wason, Peter C. 1968. “Reasoning about a rule.” *Quarterly Journal of Experimental Psychology* 20:273–81.

Wei, Jason, Wang, Xuezhi, Schuurmans, Dale, Bosma, Maarten, Ichter, Brian, Xia, Fei, Chi, Ed, Le, Quoc, and Zhou, Denny. 2022. “Chain-of-thought prompting elicits reasoning in large language models.” *Proceedings of the 36th International Conference on Neural Information Processing Systems* 24824–37.

Wu, Zhaofeng, Qiu, Linlu, Ross, Alexis, Akyürek, Ekin, Chen, Boyuan, Wang, Bailin, Kim, Najoung, Andreas, Jacob, and Kim, Youn. 2024. “Reasoning or reciting? Exploring the capabilities and limitations of language models through counterfactual tasks.” *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 1:1819–62.

Yao, Shunyu, Yu, Dian, Zhao, Jeffrey, Shafran, Izhak, Griffiths, Thomas, Cao, Yuan, and Karthik, Narasimhan. 2023. “Tree of thoughts: Deliberate problem solving with large language models.” *Proceedings of the 37th International Conference on Neural Information Processing Systems* 11809–22.

Zhang, Tianlin, Leng, Jiaxu, and Liu, Ying. 2020. “Deep learning for drug-drug interaction extraction from the literature: a review.” *Briefings in Bioinformatics* 21:1609–27.

Zhao, Tony, Wallace, Eric, Feng, Shi, Klein, Dan, and Singh, Sameer. 2021. “Calibrate before use: Improving few-shot performance of language models.” *Proceedings of the 38th International Conference on Machine Learning, PMLR* 139:12697–706.