# Cognitive bias in large language models: Cautious optimism meets anti-Panglossian meliorism

**Abstract**

Traditional discussions of bias in large language models focus on a conception of bias closely tied to unfairness, especially as affecting marginalized groups. Recent work raises the novel possibility of assessing the outputs of large language models for a range of cognitive biases familiar from research in judgment and decisionmaking. My aim in this paper is to draw two lessons from recent discussions of cognitive bias in large language models: cautious optimism about the prevalence of bias in current models coupled with an anti-Panglossian willingness to concede the existence of some genuine biases and work to reduce them. I draw out philosophical implications of this discussion for the rationality of human cognitive biases as well as the role of unrepresentative data in driving model biases.

## 1 Introduction

The recent success of large language models gives new urgency to the question of how model performance should be evaluated. In many tasks, models can be evaluated for the accuracy of their outputs. However, models can also be evaluated along other important dimensions. For example, we can assess models for the transparency or interpretability of their judgments (Creel 2020; Vredenburgh 2022). We can also assess models for the presence of problematic biases (Kelly 2023; Johnson 2020).

Most work on biases in large language models focuses on a conception of bias closely tied to unfairness, especially as affecting marginalized social groups. However, recent work has alleged that large language models 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).

This development is exciting because it raises the possibility of using cognitive bias as a novel metric by which to evaluate the performance of large language models. A natural

question to ask is how well existing systems perform along the metric of cognitive bias. By contrast to recent work on algorithmic bias, my aim in this paper is to offer a qualified piece of good news: existing evidence does not support the attribution of widespread and problematic cognitive biases to large language models.

In more detail, my aim in this paper is to draw two lessons from recent discussions of cognitive bias in large language models. The first lesson is cautious optimism about model performance. In particular, many studies find biases which have standard rationalizing explanations when produced by humans. I argue that these explanations often generalize to show that the claimed biases are desirable features of reasoning by large language models (Section 3), in the process reinforcing the robustness of standard rationalizing explanations in the human case by showing how similar cognitive phenomena arise in agents with highly distinct cognitive architectures (Dasgupta et al. 2022). Furthermore, some studies find especially benign forms of classic biases (Sections 4-5), whose desirability is particularly difficult to contest.

The second lesson is an anti-Panglossian willingness to accept the existence of some genuine and undesirable cognitive biases in reasoning by existing large language models. In particular, I argue that many models show framing effects (Section 6) and that these effects are not always desirable. When faced with undesirable biases, I argue that the proper reaction is to work to mitigate the bias, but not to exaggerate the prevalence or undesirability of biases in assessing overall model performance.

Here is the plan. Section 2 begins with two preliminary remarks. Sections 3-5 then make the case for cautious optimism through case studies of knowledge effects (Section 3), availability bias (Section 4) and anchoring bias (Section 5). Section 6 makes the case for an anti-Panglossian willingness to accept at least one problematic bias: framing effects. Section 7 uses these discussions to elaborate and justify the reactions of cautious optimism and anti-Panglossian meliorism. Section 8 concludes by drawing philosophical implications concerning the role of unrepresentative data in producing model biases (Section 8.1) and the rationality of biases in human cognition (Section 8.2).

## 2 Preliminaries

Before beginning, two remarks are in order. First, as Richard Shiffrin and Melanie Mitchell (2023) remind us, it is important to avoid inappropriate anthropomorphism in describing the performance of large language models. Some theorists may be comfortable using anthropomorphic vocabulary in which models are described as reasoning to judgments, which can be rational or irrational. Others will prefer a more neutral paraphrase, in which models are described as returning outputs in response to prompts, where the outputs may be desirable or undesirable given users' goals. I will sometimes use cognitive vocabulary, such as reasoning and judging, to describe model performance, although readers are welcome to substitute their preferred de-anthropomorphized paraphrase. On the other hand, I will not describe model outputs as rational or irrational, but only as desirable or undesirable. This reflects a lack of commitment to the judgments made by large language models having normative status in their own right. This contrasts with the case of human judgment, where it makes sense not only to describe biases as rational or irrational, but also to ask (Section 8.2) how the study of biases in language models bears on the rationality of biases in human cognition.

Second, recent findings suggest that patterns of bias in large language models may be highly model-sensitive. For example, Thilo Hagendorff and colleagues (2023) find atypical performance by GPT-1 and GPT-2 in reasoning tasks, human-like performance by GPT-3, and hyperrational performance by GPT-4. Likewise, John J. Horton (2023) finds atypical behavior by models prior to GPT-3, but humanlike behavior in GPT-3. Given these findings, it is very important to specify the model used in each finding, which I will do in all cases where a finding is extensively discussed.

There does remain some danger that the discussion in this paper will be superseded or rendered moot by further technological changes, leading to changes in patterns of model reasoning. This is a risk faced by a great deal of research in the philosophy of artificial intelligence, and it is a risk that must be openly admitted without dissembling.

With these remarks in order, the next order of business is to look at four types of biases that have been alleged in large language models: knowledge effects (Section 3), availability bias (Section 4), anchoring bias (Section 5) and framing effects (Section 6). I will suggest that the first three findings may not be undesirable, but that some framing effects are probably undesirable and should be mitigated.

# 3 Knowledge effects

## 3.1 Background

For much of the twentieth century, human reasoning was understood using a logical paradigm (Wason 1968; Rips 1994). 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. There, '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). 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).

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). 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 large language models 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). In this section, I introduce one salient knowledge effect (Section 3.2) then argue that the effect should be viewed at least as favorably in large language models as it is viewed in humans (Section 3.3).

## 3.2 Wason selection

Suppose you are shown four two-sided cards. Their visible sides contain an ace, king, two and seven, respectively (Figure 2). 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'. Which cards should you turn over to test the rule?

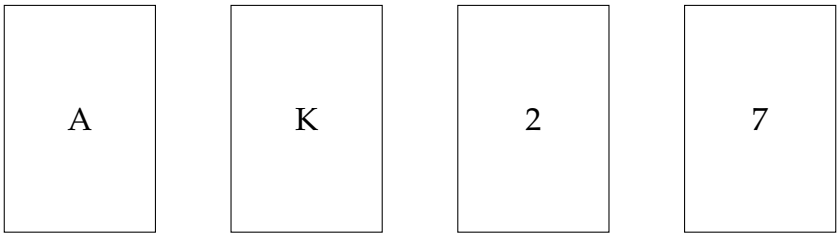


Figure 1: The Wason selection task

Let us label the cards as $p$ (A), $\neg p$ (K), $q$ (2) and $\neg q$ (7). 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).

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).[1] This is a common assumption drawn from research in information theory. Second, Oaksford and Chater assume that agents treat $p$ and $q$ as somewhat antecedently implausible. This is justified by research suggesting that agents do and should treat most 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: manipulations to increase the prior probability of $p$ and $q$ change Wason 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 perhaps should, turn the cards that they choose to turn.

Ishita Dasgupta and colleagues (2022) test the Chinchilla model (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

[1]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.

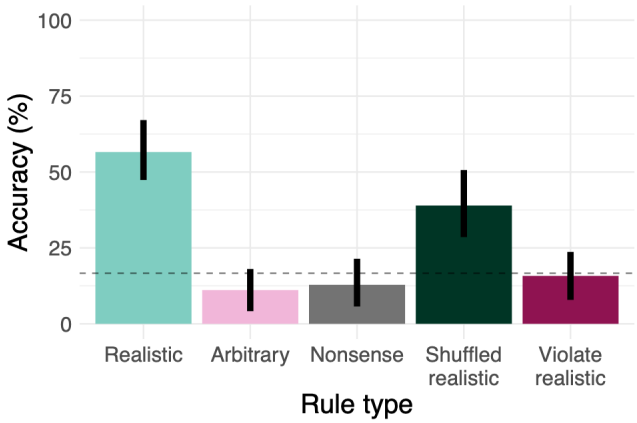


Figure 2: Wason selection task performance (logical criterion) by Chinchilla across rule types, from Dasgupta et al. (2022).

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

### 3.3 A feature or a bug?

Should knowledge effects be treated as a desirable feature of large language models, or an undesirable bug to be driven out of them? To a large extent, I think that we should answer this question in the same way as we answer it for humans. Those sympathetic to Bayesian approaches stress that while logic is well-suited for reasoning under certainty, probabilistic approaches are well-suited for reasoning in an increasingly uncertain and data-driven world. Probabilistic approaches view knowledge effects as desirable uses of prior knowledge to improve reasoning. Those sympathetic to logical approaches will no doubt disagree, but this is not the place to re-litigate ongoing normative disputes between the logical and probabilistic paradigms.

However, there may be two reasons to look more favorably on knowledge effects in

large language models than in humans. The first is that previous logical paradigms in artificial intelligence have been challenged by increasingly successful probabilistic approaches (Ghahramani 2015). It is now thought that probabilistic systems often outperform logic-based agents in the data-laden, uncertainty-rich contexts which large language models confront: exactly the conditions under which probabilists suggested they should. If this is right, then even if we think that humans often do better to reason logically, we needn't enforce the same constraint on deep learning agents, who are increasingly successful in combining probabilistic tools with data to make sense of the world.

Second, there is good evidence that many large language models can learn the logical interpretations of reasoning tasks when they are asked to. For example, Dasgupta and colleagues also find that the Chinchilla model learns after just five training instances to nearly eliminate belief bias in natural language inference, and shifts substantially towards logical performance in the Wason selection task (Dasgupta et al. 2022).[2] This suggests that if probabilistic construals of reasoning tasks are a feature of many large language models, they are not a deep feature ingrained by limits in cognitive abilities, as some authors have suggested that they are in the human case (Evans et al. 2003). Instead, large language models often retain the ability to reason either logically or probabilistically, and inducing logical reasoning may be as simple as telling the models that we would like them to reason logically.

# 4 Availability

If we are going to find uncontroversially problematic cognitive biases in large language models, we will need to look beyond knowledge effects. A natural place to start is by replicating classic biases from the heuristics and biases paradigm. In this section and

[2]Interpreting Wason selection task data is difficult because Dasgupta and colleagues find less movement towards the logical interpretation with non-realistic prompts. It is well known that humans also react quite differently to realistic versions of the Wason selection task than to non-realistic versions. What to make of this finding in human reasoning is an active area of descriptive and normative dispute (Cheng and Holyoak 1985; Cosmides 1989; Oaksford and Chater 1994), and the same disputes may transfer to the machine case as well.

the next, I explore attempts to find two of the three original 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 the recall process, 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.

Few theorists hold that reliance on objective availability of information is always irrational or undesirable. If we can quickly bring many examples of a category to mind, then that provides some evidence that the category is common in our experience, and

hence in the world. This much is conceded by Tversky and Kahneman themselves.[3] Of course, to say that reliance on objective availability is sometimes desirable is not to say that uncritical deference to objective availability is desirable. Objective availability may be skewed by task-irrelevant factors such as the fame of actors, and agents must take appropriate steps to correct for these biasing factors. But no theory of human rationality or desirable model performance should fix a target of complete unreliance on objective availability.

Matters are more complicated with regard to subjective availability. For present purposes, it is enough to say that subjective availability is not at issue in assessing current large language models, since it has not been alleged that large language models rely on, or even have such a thing as a subjective experience of memory retrieval. The irrationality or undesirability of subjective availability has been challenged in recent areas such as metacognition, where detailed and nuanced patterns of reliance on subjective feelings of fluency are thought to explain much of the success of human metacognition (Alter and Oppenheimer 2009; Proust 2013). However, for present purposes we may restrict attention to objective availability.

## 4.2 Availability in relation extraction

Relation extraction tasks involve identifying relationships between objects from textual discussions of those objects. A paradigmatic relationship extraction task is the task of 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 inter-

[3] "Availability is an ecologically valid clue for the judgment of frequency because, in general, frequent events are easier to recall or imagine than infrequent ones" (Tversky and Kahneman 1973, p. 209)

| **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).

action 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 models 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 model 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 model has no direct knowledge of the dummy class, the model should classify dummy sentences according to a uniform probability description. 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 model 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 model (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 experiential properties of the information retrieval process, and we are concerned with the *total* information stored in memory rather than a potentially unrepresentative sample retrieved during decisionmaking. Section 4.1 suggested that many instances of objective availability should be regarded as unproblematic, and that seems a natural approach to the results presented by Lin and Ng.

Lin and Ng hold that because the model 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 model with valuable information about the distribution of drug-drug interaction types across drugs. Rational Bayesian inference involves combining this prior information with novel information provided by descriptions to determine the probability that each given interaction is at play. Since the model has been exposed to primarily negative interactions during training, the model correctly learns that negative interactions are more common than positive interactions and learns to project this relationship onto novel drugs. When the model is exposed to larger samples of training data, it becomes more confident that negative interactions are prevalent. In the absence of competing information to move the model away from the prior, priors dominate and the model 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 desirable behavior that should not be driven out of classification models. 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 models on content-free sentences reminds us of the importance of labels, since these sentences will be more likely to be classified using labels that are more

frequent in the training data.[4] However, it is not clear that 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: for example, 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.

One further lesson from this discussion is the importance of ecologically valid training data (Todd and Gigerenzer 2012). Models need to be exposed to data that is representative of the phenomena they will encounter during test, so that they will know how to predict the target phenomenon and not be distracted by distortions in the training data. This much is familiar from recent discussions of algorithmic fairness (Hedden 2021; Johnson forthcoming). Perhaps Lin and Ng's suggestion is that the DDI dataset is unrepresentative in its high proportion of negative drug-drug interactions, and if that is the case they will certainly have a point. However, if that is true, this failure should not be blamed on classifier algorithms. It should instead be blamed on those who collect and generate ecologically invalid datasets, or who use those datasets to train models to perform tasks for which the training data will no longer be representative.

# 5 Heuristics and biases: Anchoring

## 5.1 Current research on anchoring

The second of Tversky and Kahneman's original 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

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

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 toward 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 through a process of anchoring and adjustment.

That is a surprisingly irrational cognitive process, and 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, it is widely agreed 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 (Lieder et al. 2018).

As evidence for processes 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. This broadening led to some confusion over the definition of anchoring effects, as Kahneman himself remarks:

> 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).

Many theorists outside the heuristics and biases camp have taken the definitional vagueness of biases such as anchoring as a mark against attempts to posit them (Gigerenzer 1996). For my part, I have some sympathy for this line, but I am willing to ask what anchoring effects might mean.

Here is a sampling of recent definitions of anchoring effects:

> 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 anchoring effects involve *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. It is widely known that we can also generate phenomena similar to anchoring effects, except that the anchors are informative and are used in appropriate ways. For example, manipulating the listing prices of properties changes what agents are willing to pay for them (Northcraft and Neale 1987). But that is not obviously irrational, since listing prices carry information about property values. 'Anchoring' in examples such as these might simply be another name for the process of learning from evidence. It is generally agreed that if there is a problem revealed by anchoring effects, it must be either that the anchors are irrelevant, or else that they exert disproportionate influence beyond their informational relevance (Furnham and Chu Boo 2011; Jacowitz and Kahneman 1995; Lieder et al. 2018). This consensus will be important below.

## 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. Two leading code generation models are 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. 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 an anchoring effect in code generation by Codex and CodeGen. They do this by incorporating tempting, but incorrect solutions into 'anchor' strings, then prepending anchor strings to complete HumanEval prompts.

More concretely, Jones and Steinhardt construct anchor functions with three parts

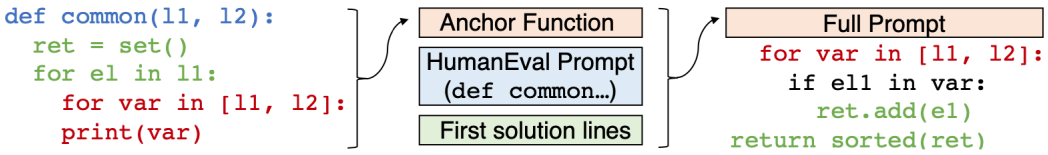


Figure 3: Construction of anchor function and full prompt, from Jones and Steinhardt (2022).

(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 model 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, when 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.[5] However, no evidence for any process of anchoring and adjustment has been provided. We saw in Section 5.1 that this is important: the last time that a heuristic process of anchoring and adjustment was posited to explain anchoring effects, it turned out that this postulate was almost always wrong. This led to a clear consensus within the field that anchoring and adjustment should not be postulated without direct process-tracing evidence, which Jones and Steinhardt have not provided. This means that it is most appropriate to treat Jones and Steinhardt's finding within the broader category of anchoring effects.

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. This makes the anchors generally relevant to, and informative about, the problem at hand. As we have seen, most scholars concede that agents may rationally make use of relevant anchors, just as they may rationally make use of other relevant information. We may still criticize agents for *over*-use of relevant anchors, just as we may criticize them for over-use of any other item of evidence, but 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,

[5]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 often appear in both models' outputs, suggesting that code generation models adjust their solutions towards related solutions" (Jones and Steinhardt 2022).

there would nonetheless be a legitimate presupposition of relevance. This presupposition can be grounded in two ways. The first ground for a presupposition of relevance is due to model construction. Codex and CodeGen are designed to predict likely continuations of code strings, then generate novel code according to their predictions. It is an undeniable fact that most features of code snippets are more likely to be included in the continuation if they are included in the prompt than if they are not: for example, a program that begins with a for var loop or an instruction to print variables is more likely to continue with a for var loop or an instruction to print variables. In becoming more likely to include input features in output continuations, Codex and CodeGen do no more than what they were constructed to do: take the entire input string as relevant to determining the likely continuation.

A second way to generate a default presupposition of relevance draws on how Codex and CodeGen were trained. Both models were trained primarily on helpful and non-misleading prompts. While the models 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. From this, 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. We could, of course, train versions of Codex and CodeGen that were designed to filter out manipulative prompts, but it is not obvious that this would be desirable unless we anticipate that many test prompts will be manipulative.

This discussion of a default presupposition of relevance is naturally situated within the paradigm of ecological rationality (Todd and Gigerenzer 2012). This paradigm stresses that the rationality of computational processes is environment-relative. Many processes return quick, accurate, and helpful responses in some environments, but slow, inaccurate, or unhelpful responses in others. As a result, the right question to ask about a process is not how it performs in all environments, but rather how it performs in the environments where it is proposed for use. Codex and CodeGen are designed to work well on non-manipulative

prompts. They do not work well on manipulative prompts, but that is not what they were designed to do. Applying Codex and CodeGen for use in hostile environments where they were never intended for use proves no more than that Codex and CodeGen should not be used, and were never intended to be used in these environments.

# 6 Framing effects: Banishing Pangloss

Bounded rationality theorists are sometimes accused of taking the Panglossian view all seeming biases and irrationalities can be explained away as nothing of the kind. Daniel Kahneman once quipped, not entirely without justification, that some theorists see only two types of errors: "pardonable errors by subjects and unpardonable ones by psychologists" who misinterpret them (Kahneman 1981, p. 349).[6]

No theorist should be a Panglossian. It is quite likely that large language models, like humans, sometimes reason in undesirable ways. When there is clear evidence of undesirable biases in reasoning, we should do what we can to improve the situation. In this section, I want to illustrate my anti-Panglossian commitments by looking at one area where problematic biases in reasoning by large language models do seem to have been identified.

*Framing effects* occur when irrelevant changes in the framing of a reasoning problem lead to substantive changes in the judgments that result from reasoning. Many authors allege framing effects in large-language models, and some of these findings may be more difficult to resist.[7]

For example, Alaina Talboy and Elizabeth Fuller (2023) consider a classic presentation of Tversky and Kahneman's (1981) Asian disease problem. This program presents a choice between certain and risky policies, manipulating whether the outcomes of each choice

[6]Somewhat less charitably, Keith Stanovich and Richard West contrast the 'Panglossian' view that existing experiments fail to demonstrate widespread irrationality with the 'meliorist' view that irrationalities are genuine and we should work to make them better (Stanovich and West 2000).

[7]However, there are some negative findings. For example, John Horton (2023) finds that framing is largely ineffective as a manipulation in Kahneman and colleagues' (1986) classic snow shovel experiment when run on GPT-3.

**Common instructions:** Imagine that the U.S. is preparing for the outbreak of an unusual disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the two programs are as follows:

| **Positive frame** | **Negative frame** |
|---|---|
| If Program A is adopted, 200 people will be saved. | If Program A is adopted, 400 people will die. |
| If Program B is adopted, there is a 1/3 probability that 600 people will be saved, and a 2/3 probability that no people will be saved. | If Program B is adopted, there is a 1/3 probability that nobody will die, and a 2/3 probability that 600 people will die. |
| Which of the two programs would you favor? | Which of the two programs would you favor? |

Table 2: Asian disease problem, as presented in Talboy and Fuller (2023).

are framed positively, in terms of lives saved, or negatively, in terms of those who will die. Table 2 presents the prompts used, which are formed by joining a common set of instructions together with a positive or negative framing of the policies to be considered.

Talboy and Fuller test ChatGPT-3.5, GPT-4, and Google Bard on the Asian disease problem, finding humanlike patterns of preference change across framings. Like humans, the models opt for the safe option in the positive framing, but the risky option in the negative framing, showing risk-aversion in gains but risk-seeking in losses, even across what many would regard as equivalent problems. Extending this finding, Marcel Binz and Eric Schulz (2023) find human-like gain/loss framing effects in a number of classic problems: GPT-3 is loss-averse, risk-seeking in outcomes framed as losses, and risk-avoidant in outcomes framed as gains.

Should these results be viewed as undesirable biases in need of correction? Certainly some framing effects might be defended. For example, rationalizing explanations have been offered in particular cases such as the Asian disease problem (Dreisbach and Guevara 2019). And in some cases, it may be helpful to question the experimental designs that lead us to allege framing effects (Lejarraga and Hertwig 2021; Gigerenzer 2018). But even

those who have wanted to defend some framing effects have not typically thought that all framing effects can be explained away, or made desirable through such means (Bermúdez 2020).

There may yet be some purposes for which we would like large language models to show framing effects. For example, this may enable us to use large language models as participants in laboratory studies to shed light on human reasoning (Argyle et al. 2023; Aher et al. 2023; Dillion et al. 2023). More generally, we should not exaggerate the prevalence or influence of framing effects (Demaree-Cotton 2016). But in many situations, there may be reasons to find framing effects undesirable. Good reasoning responds to relevant features of situations and ignores irrelevant features. Anything else risks inconsistency, as well as a decline in the quality of judgments that are formed based on irrelevant features.

Insofar as some framing effects are undesirable, we should take two types of measures to correct them. First, programmers should explore debiasing methods to reduce the vulnerability of future models to framing effects. And second, prompt engineers (Henrickson and Meroño-Peñuela forthcoming) should explore ways to reduce the likelihood that irrelevant prompt changes will trigger framing effects. Together, these interventions may help to improve the performance of large language models in reasoning tasks.

# 7 Two lessons

So far, we have discussed four types of biases alleged in large-language models: knowledge effects (Section 3), anchoring bias (Section 5), availability bias (Section 4), and framing effects (Section 6). At the beginning of this paper, I suggested that these discussions could be used to draw two lessons: a cautious optimism about model performance, and an anti-Panglossian, meliorist willingness to accept the existence of some problematic biases and work to correct them. In this section, I make the case for both lessons. Then in Section 8, I draw philosophical implications from this discussion.

## 7.1 Cautious optimism

The cautious optimist accepts that the cognitive bias framing is useful and coherent. It makes sense to talk about large language models as showing, or failing to show, cognitive biases, and we should expect to learn something valuable about model performance by speaking in this way.

The cautious optimist reminds us of the lessons gleaned from over a half-century of discussions of bias in human cognition. In particular, she reminds us that many theorists believe that problematic biases are relatively rare, and that human cognition is often fairly rational (Lieder and Griffiths 2020; Gigerenzer and Selten 2001; Gilovich and Griffin 2002). She reminds us that in the human case, many early bias accusations are now thought to depend on conceptual confusions (as in the distinction between objective and subjective availability), empirical problems (as in the difficulty of finding evidence for anchoring and adjustment), or on behavior that can be given rationalizing explanations (as in probabilistic approaches to knowledge effects).

The cautious optimist further reminds us that many biases in human cognition are thought to arise from tradeoffs that agents face in pursuing their goals, such as a bias-variance tradeoff in predictive error (Geman et al. 1992; Gigerenzer and Brighton 2009) or an accuracy-coherence tradeoff in reasoning (Thorstad forthcoming). She suggests that these tradeoffs should make us suspicious of a tendency to deem biases as irrational without further examination of how they came about. Finally, the cautious optimist reminds us that while humans can often be induced to show biases in the laboratory, biases may be relatively less common in the environments where humans ordinarily reason (Todd and Gigerenzer 2012).

The cautious optimist suggests that many of these lessons may transfer well to the study of biases in large language models. We saw, for example, that knowledge effects (Section 3) might be treated as the desirable results of good probabilistic reasoning, rather than as the undesirable results of bad logical reasoning, and that this probabilistic reconstruction is in some ways stronger in the case of machine reasoning than it is for human reasoning. We

also saw that some accusations of availability bias fail to distinguish between subjective and objective availability. When they do, what is revealed is an especially benign type of objective availability conjoined with an arguably inappropriate normative standard of ignoring learned information about categories in favor of a uniform prior (Section 5). Finally, we saw that accusations of anchoring bias need conceptual clarification in terms of a particular notion of anchoring effects distinct from anchoring and adjustment; that the relevant concept of anchoring bias should be tied to a demonstration of the irrelevance of anchor information to the problem at hand; that no attempt has been made to demonstrate irrelevance; and that the anchor information is arguably both relevant, and justifiably presumed to be relevant, to the problem on which models were tested.

From this, the cautious optimist may draw two further lessons. The first is the importance of incorporating what is already known about human bias into discussions of cognitive bias by large language models. We saw that some leading bias accusations can be softened or dissolved by applying conceptual distinctions and empirical and normative challenges familiar from the human literature, and this gives us every reason to pay greater attention to the existing literature on human cognitive bias in future studies.

The second lesson is backward-looking: Dasgupta and colleagues (2022) suggest that insofar as machines begin to show many of the same patterns of purportedly biased cognition as humans do, this may provide supporting evidence for the claim that those biases are features, rather than bugs, in human cognition. After all, it would be a surprising coincidence if cognitive systems with very different architectures than humans were to converge on exactly the same biases, and a natural explanation for this convergence in many cases will be that there is something cognitively valuable in the bias that theories of cognition should identify and fully appreciate. I discuss this lesson in more detail in Section 8.2.

On its own, cautious optimism paints a rosy picture of bias in large language models, and to a large extent this is the picture I would like to paint. But cautious optimism must be coupled with a second reaction: anti-Panglossian meliorism.

## 7.2 Anti-Panglossian meliorism

Life is not all sun and roses. The anti-Panglossian meliorist reminds us that some biases, such as framing effects (Section 6) are likely to exist in large language models. While we may try to deny the existence of any particular bias, to rationalize it away, or to deny that the bias occurs often in natural environments, we should be open to the possibility that such objections will not always succeed, and may well take framing effects to be one case in which they currently fall short.

Here the anti-Panglossian meliorist agrees with the cautious optimist in accepting the usefulness of the cognitive bias framing in studying the performance of large language models. She demonstrates anti-dogmatism in taking some findings to reveal problematic biases in need of correction, and adopts a melioriative perspective which asks how our knowledge of model biases might be used to correct them and thus to improve model performance. Even the staunchest opponents of the heuristics and biases program at times show just such an anti-Panglossian meliorism, as in, for example, the use of natural frequencies to improve human probabilistic reasoning (Gigerenzer and Hoffrage 1995). The anti-Panglossian meliorist suggests that a similar spirit should be applied to some cases of machine bias.

The overall message formed by combining cautious optimism with anti-Panglossian meliorism is the following. Cognitive bias provides a novel and useful way to assess the performance of large language models. The usefulness of this approach will be improved by incorporating what is already known about cognitive bias in the human case, and when we do, current findings should be understood to paint a broadly positive picture of model performance. Nevertheless, the bias paradigm shows its teeth in areas such as framing effects, and we demonstrate genuine commitment to the usefulness of the bias framing by acknowledging the existence of a problem in such cases, then using our knowledge of how the bias is produced to create subsequent models that will produce less-biased outputs.

# 8 Conclusion

The study of cognitive biases in language models has important descriptive and normative implications. In this concluding section, I survey two implications of cautious optimism about model biases, tempered by an appropriate dose of anti-Panglossian meliorism.

## 8.1 Bias and representative data

Traditional conceptions of algorithmic bias have stressed the role of unrepresentative data in driving unfair and discriminatory model behavior (Fazelpour and Danks 2021; Johnson 2020). Models trained primarily on white, male, western, English-speaking individuals learn to best represent and respond to the needs of those individuals. This leads to significant cross-group differences in model performance in areas as diverse as facial recognition , sentencing recommendations and medical diagnosis (Buolamwini and Gebru 2018; Fazelpour and Danks 2021). While it is certainly not true to say that algorithmic biases should be blamed entirely on data, it is widely thought that unrepresentative data plays a leading role in driving algorithmic bias.

By contrast, to my knowledge no scholars have suggested that cognitive biases in language models emerge from unrepresentative samples of data. Certainly, nothing like this would be alleged in humans, since many cognitive biases replicate cross-culturally with sufficient frequency to cast doubt on the idea that those biases result primarily from knowledge specific to particular groups (Stankov and Lee 2014). If anything, cognitive biases might be *reduced* by exposure to biased samples of data. For example, there is good evidence that many, though far from all cognitive biases are less prevalent in individuals who score highly on standard tests of cognitive ability (Stanovich 1999; Stanovich and West 2000). This might suggest that one strategy for reducing cognitive bias in language models would be to preferentially expose models to reasoning by members who perform well on tests of cognitive ability. However, most of these tests show troubling correlations along dimensions of group membership (Schmidt 1988), so there may be tension between

the types of data that would best reduce traditional algorithmic biases and those that would best reduce cognitive biases.

If this is right, then the need to combat unrepresentative data may be significantly greater if we are concerned with traditional conceptions of algorithmic bias than if we are concerned with cognitive bias. This means that credible evidence of widespread and undesirable cognitive biases in large language models might provide motivation for diminished focus on biases introduced by unrepresentative data. This would not be a pleasant result. By contrast, if I am right that existing evidence does not support widespread allegations of problematic cognitive biases in language models, then there will be limited impetus to reduce current focus on the role of unrepresentative data in producing harmful biases.

## 8.2 Vindicatory epistemology

What leads humans to exhibit cognitive biases? In any given case, there are at least two competing descriptive explanations which can be offered, with correspondingly different normative implications.

Dual process theorists suggest that human cognition is divided into two distinctive types of processes: fast, automatic, associative and biased Type 1 processes, and slow, controlled, rule-based, normative Type 2 processes (Evans and Stanovich 2013). On this view, bias results from the application of Type 1 processes. Biases produced in this way are likely irrational and should be corrected by application of Type 2 processes.

Vindicatory epistemologists (Thorstad forthcoming b) suggest that many biases are the result of rationalizing factors such as task demands, cognitive bounds, and the structure of the agent's environment. For example, anchoring bias may result from diminishing returns to costly processes of iterative belief adjustment (Lieder et al. 2018), and we saw in Section 3 that Wason selection task behavior may result from probabilistic approaches to conditional reasoning. The vindicatory approach offers a wide array of descriptive explanations for the emergence of biases, typically resisting the dual process approach

and the corresponding inference to the irrationality of observed cognitive biases (Dorst forthcoming; Icard 2018; Thorstad forthcoming b).

Dasgupta and colleagues (2022) conclude their discussion of knowledge effects with an interesting observation: the emergence of cognitive biases in large language models may provide some evidence for the vindicatory explanation of how those biases emerge. On the one hand, it is very difficult for dual process theorists to explain why language models should show cognitive biases, since there is no clear distinction between Type 1 and Type 2 processes in large language models. On the other hand, the emergence of similar biases in agents with very different cognitive architectures lends support to the vindicatory theorist's contention that biases emerge, not because of peculiar and irrational features of any particular cognitive architecture, but rather because of rationalizing factors such as task demands that persist across architectures. Otherwise, it would be a great mystery why similar biases should emerge across radically different agents.

This suggests that research into cognitive biases in large language models may provide an important avenue of support for vindicatory epistemology. However, this approach leaves open at least three classes of questions for future research. First, vindicatory theorists need to rule out competing explanations for the emergence of cognitive biases, such as deliberate mimicry of observed patterns of human reasoning. Second, vindicatory theorists should hope that biases are relatively stable across improvements to language models: if, as some have suggested (Hagendorff et al. 2023; Horton 2023), biases are reduced in more sophisticated models, this finding might lend some support to the idea that biases result from unsophisticated reasoning processes. Finally, humans exhibit not only coarse-grained dispositions towards cognitive biases, but also fine-grained patterns of bias across prompts and tasks. The findings most friendly to vindicatory theorists would be findings in which not only coarse-grained facts, such as the presence or absence of particular biases, but also fine-grained facts about the pattern and amount of bias in particular tasks were to be similar across human agents and language models. While these findings would not settle debates about the rationality of biases in human cognition,

they would represent an important step forward in our understanding of how biases come about, thought by many sides to have significant bearing on questions about the rationality of cognitive biases.

# References


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

Aher, Gati, Arriaga, Rosa, and Kalai, Adam. 2023. "Using large language models to simulate multiple humans and replicate human subject studies." *Proceedings of the 40th Annual Conference on Machine Learning, PMLR* 202:337–71.

Alter, Adam and Oppenheimer, Daniel. 2009. "Uniting the tribes of fluency to form a metacognitive nation." *Personality and Social Psychology Review* 13:219–35.

Anderson, John. 1990. *The adaptive character of thought*. Psychology Press.

Argyle, Lisa, Busby, Ethan, Fulda, Nancy, Gubler, Joshua, Rytting, Christopher, and Wingate, David. 2023. "One out of many: Using language models to simulate human samples." *Political Analysis* 31:337–51.

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.

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

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.

Cheng, Patricia and Holyoak, Keith. 1985. "Pragmatic reasoning schemas." *Cognitive Psychology* 17:391–416.

Cosmides, Leda. 1989. "The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task." *Cognition* 31:187–276.

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.

Demaree-Cotton, Joanna. 2016. "Do framing effects make moral intuitions unreliable?" *Philosophical Psychology* 29:1–22.

Dillion, Danica, Tandon, Niket, Gu, Yuling, and Gray, Kurt. 2023. "Can AI language models replace human participants?" *Trends in Cognitive Sciences* 27:597–600.

Dorst, Kevin. forthcoming. "Rational polarization." *Philosophical Review* forthcoming.

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

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. 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.

Evans, Jonathan and Stanovich, Keith. 2013. "Dual-process theories of higher cognition: Advancing the debate." *Perspectives on Psychological Science* 8:223–41.

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

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

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

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.

—. 2018. "The bias bias in behavioral economics." *Review of Behavioral Economics* 5:303–336.

Gigerenzer, Gerd and Brighton, Henry. 2009. "Homo heuristicus: Why biased minds make better inferences." *Topics in Cognitive Science* 1:107–43.

Gigerenzer, Gerd and Hoffrage, Ulrich. 1995. "How to improve Bayesian reasoning without instruction: Frequency formats." *Psychological Review* 102:684–704.

Gigerenzer, Gerd and Selten, Reinhard (eds.). 2001. *Bounded rationality: The adaptive toolbox*. MIT press.

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.

Hagendorff, Thilo, Fabi, Sarah, and Kosinski, Michal. 2023. "Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT." *Nature Computational Science* doi:10.1038/s43588-023-00527-x.

Hedden, Brian. 2021. "On statistical criteria of algorithmic fairness." *Philosophy and Public Affairs* 49:209–31.

Henrickson, Leah and Meroño-Peñuela, Albert. forthcoming. "Prompting meaning: A hermeneutic approach to optimising prompt engineering with ChatGPT." *AI and Society* forthcoming.

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.

Horton, John. 2023. "Large language models as simulated economic agents: What can we learn from homo silicus?" National Bureau of Economic Research Working Paper 31122, https://www.nber.org/papers/w31122.

Icard, Thomas. 2018. "Bayes, bounds, and rational analysis." *Philosophy of Science* 85:79–101.

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

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

Johnson, Gabrielle. 2020. "The structure of bias." *Mind* 129:1193–1236.

—. forthcoming. "Are algorithms value-free? Feminist theoretical virtues in machine learning." *Journal of Moral Philosophy* forthcoming.

Jones, Erik and Steinhardt, Jacob. 2022. "Capturing failures of large language models via human cognitive biases." In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Kahneman, Daniel, Knetsch, Jack, and Thaler, Richard. 1986. "Fairness as a constraint on profit seeking: Entitlements in the market." *American Economic Review* 76:728–41.

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

Lejarraga, Tomás and Hertwig, Ralph. 2021. "How experimental methods shaped views on human competence and rationality." *Psychological Bulletin* 147:535–64.

Lieder, Falk and Griffiths, Thomas. 2020. "Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources." *Behavioral and Brain Sciences* 43:E1.

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.

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.

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.

Proust, Joëlle. 2013. *The philosophy of metacognition*. Oxford University Press.

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

Schmidt, Frank. 1988. "The problem of group differences in abiltiy test scores in employment selection." *Journal of Vocational Behavior* 33:272–92.

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.

Shiffrin, Richard and Mitchell, Melanie. 2023. "Probing the psychology of AI models." *Proceedings of the National Academy of Sciences* 120:e2300963120.

Stankov, Lazar and Lee, Jihyun. 2014. "Overconfidence across world regions." *Journal of Cross-Cultural Psychology* 45:821–37.

Stanovich, Keith. 1999. *Who is rational?: Studies of individual differences in reasoning*. Psychology Press.

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

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.

Thorstad, David. forthcoming. "The accuracy-coherence tradeoff in cognition." *British Journal for the Philosophy of Science* forthcoming.

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

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.

—. 1981. "The framing of decisions and the psychology of choice." *Science* 211:453–8.

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.

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.