LLMs Can Never Be Ideally Rational

Simon Goldstein

Abstract

LLMs have dramatically improved in capabilities in recent years. This raises the question of whether LLMs could become genuine agents with beliefs and desires. This paper demonstrates an *in principle* limit to LLM agency, based on their architecture. LLMs are next word predictors: given a string of text, they calculate the probability that various words can come next. LLMs produce outputs that reflect these probabilities. I show that next word predictors are exploitable. If LLMs are prompted to make probabilistic predictions about the world, these predictions are guaranteed to be incoherent, and so Dutch bookable. If LLMs are prompted to make choices over actions, their preferences are guaranteed to be intransitive, and so money pumpable. In short, the problem is that selecting an action based on its potential value is *structurally* different then selecting the description of an action that is most likely given a prompt: probability cannot be forced into the shape of expected value. The in principle exploitability of LLMs raises doubts about how agential they can become. This exploitability also offers an opportunity for humanity to safely control such AI systems.

1 Introduction

Recent developments in AI are impressive. Large language models can generate text that is fluent, accurate, and responsive to human questions. Moreover, there is good reason to expect that with increased investment in computational resources and training data, large language models will continue to improve in capabilities (Kaplan et al., 2020).

Still, it is difficult to predict exactly which capabilities large language models will develop. As AI systems have scaled, they have begun to exhibit

emergent capabilities, sudden non-linear improvement on tasks that are often unrelated to the original training objective of the system (Wei et al., 2022).

The original training objective of large language models is to predict the next word. Large language models are trained on strings of text, by giving it a series of words, and asking it to make a prediction about the next word that follows in the training data. But large language models have developed many startling capabilities that are not obviously related to the task of completing the next word. For example, LLMs have started to display abilities in computer graphics (Feng et al., 2024) and chemical research (Bran et al., 2023).

Large language models have developed unexpected abilities to succeed in tasks that lack any obvious connection to next word prediction. This raises the question: are there any in principle barriers to the kinds of capabilities AI systems could develop? In this paper, I'll argue that there are in principle barriers to two of the most exciting potential applications of large language models: operating as oracles, and operating as agents.

With oracles, the goal is to create AI systems that can make accurate predictions about the world. One recent goal has been to design LLMs that can match or surpass humans at forecasting. The LLM is prompted to make probabilistic predictions about various events. The question is whether LLM forecasts can match or surpass the accuracy of human forecasters. There have been promising results. Halawi et al. (2024) were able to roughly match human level forecasting by supplementing the base LLM with technology for researching news articles about the topic, and reasoning carefully. Schoenegger et al. (2024b) were able to slightly exceed human forecasts by appealing to the "silicon wisdom of the crowd": they aggregated forecasts from several different LLMs, to produce a more accurate prediction.

Besides from oracles, another task is to create AI agents. With agents, the goal is to create AI systems that can successfully plan and execute complex actions over time. The issue is pressing. Today's AI labs are competing fiercely to be the first to develop 'AGI', or artificial general intelligence. Some define AGI as an AI that is capable of pursuing long term plans and strategic reasoning (Carlsmith, 2022). Perhaps the most promising path to AGI is through scaling LLMs, taking models like GPT-4 and increasing the compute used to train them until something agential emerges from them. Indeed, there are already AI systems that rely on LLMs to produce complex plans (see for example Wang et al. (2023a), Liu et al. (2023), and Huang et al. (2024)).

In this paper, I'll argue that there are in principle limits to the ability

of LLMs to be oracles or agents. Large language models are next word predictors: they take a string of text as input, and output a prediction about the word that is most likely to follow. I'll argue that LLMs are guaranteed to be exploitable, in several senses.

First, there is a challenge for oracles: LLMs are Dutch bookable. This means that when an LLM assigns probability to events, it cannot be probabilistically coherent. The model is architecturally guaranteed to violate the axioms of the probability calculus. And this means that if you bought bets according to its verdicts, you would be subject to a sure loss. I will show that this problem comes from a tension between next word prediction and probabilistic prediction. When a next word predictor outputs a probability judgment, it is making a prediction about which probability claim is most likely to follow some text. This turns out to be *structurally* different from predicting how likely an event is. The latter task can conform to the axioms of the probability calculus, but the former cannot.

Second, there is a challenge for agents: LLMs exhibit intransitive preferences. If you prompt an LLM to choose between actions, the LLM will exhibit cycles. It will choose A over B, choose B over C, and choose C over A. This means that the model's preferences will violate the axioms of decision theory. Again, this means that the system can be exploited by a money pump, accepting a series of trades that produce a sure loss.

I will argue that this problem comes from a *structural* tension between next word prediction and genuine agency. A next word predictor considers a series of actions, and chooses the one that is most likely to follow some text. An agent considers a series of actions, and chooses the one that best satisfies its expectation of its desires, given its beliefs. The latter task can conform to the axioms of decision theory, but the former cannot.

After establishing the initial challenge, I consider upshots and responses. I'll consider two kinds of upshots. First, I'll argue that my results are relevant to whether LLMs have beliefs and desires. A rich tradition of work in philosophy has suggested that in order to have beliefs and desires, you need to possess certain baseline levels of coherence. But the exploitability of LLMs may force them below that baseline level of coherence. Second, at the end of the paper I'll argue that my results are relevant to AI safety. It may be possible to block LLMs from destroying humanity by using money pumps to take away their resources.

I'll also consider three responses to my results. First, my results centrally apply in cases where a model is prompted in separate instances about a

series of forecasts or actions. But the results can potentially be avoided when the model is instead prompted continuously over the course of a single conversation, or 'context'. This opens the way to distinguishing two different ways of thinking about agency: the agent of the model, and the agent of the context. While my results make serious trouble for thinking of the underlying model as an agent, they may allow us to ascribe agency to the model over the course of a single evolving context; but once that context is replaced with another, the initial agent will be gone.

Second, my results apply to the outputs of the model. But they leave open that the model may have probabilistically coherent internal representations encoded inside the system. Several researchers have recently searched for LLM beliefs in something like this way, either by appealing to the model's token probabilities (Hofweber et al., 2024), or to its internal embeddings (Herrmann and Levinstein, 2024; Burns et al., 2024; Azaria and Mitchell, 2023). I'll argue that my results create the same problem for both views. While these representations may have the structure of probabilities, they do not play the functional role of belief. Part of the functional role of belief is to cause action by conspiring with desire to produce some representation of the value of different acts. The problem is that the internal representations identified in this research do not produce outputs in this way. They produce outputs without any reference to desire, simply by selecting outputs on the basis of their probability. My result will establish that model outputs lack the kinds of structural coherence that we ordinarily associate with rational action. In this way, there may be no way to connect LLM internal representations to coherent action explanations in the way stereotypically associated with belief.

Finally, my results make two assumptions about token probabilities. First, I assume that ideally rational token probabilities are *semantically coherent*, meaning that probabilities over strings are ultimately derived from the probability of the propositions expressed by those strings. Second, I assume that ideally rational token probabilities will be *uncertain* regarding claims about likelihoods and preferences. I'll stress test my result by considering whether these assumptions are appropriate, and whether weakening the assumptions can escape my results.

2 Large language models

In this section, I'll explain how large language models work, and the sense in which they are next word predictors. My argument below will apply to a wide range of AI systems which includes large language models; but it won't depend on many of the key features of large language models that distinguish them from other systems.

Large language models are trained on very large data sets filled with text. Each data point for the large language model is a string of text, combined with the word which follows that text. One example would be the string *long explanations of AI are*, followed by the word *dull*. Large language models are trained to take the string as input, and output a prediction about what word is most likely to follow the string.

The model itself is a neural net, which means that it contains many layers of internal nodes, each of which fires in response to the initial string. In particular, every word in the initial string is represented as a vector in the neural net, and a large series of weights connects a large series of nodes in the net, allowing for complex information processing about the string. At the end of this process, the model produces a probability distribution over the most likely word to follow the string.[1]

To get good at predicting the next word, the model undergoes training. It is given lots of examples, and in each example, it makes a prediction about which word is likely to come next. Each prediction is tested against the correct answer, and the model is scored on its prediction. The weights in the neural network are then slightly adjusted in the direction of the correct answer. As this process is repeated many times, the model becomes very accurate.

In practice, we usually interact with large language models using a chatbot user interface. This means that we type in a 'prompt' to the model, and the model gives an answer. The prompt that we give to the model plays the

[1]For a more detailed explanation of large language models, see Wolfram (2023). Strictly speaking, the model operates in tokens rather than words. There are roughly 4 tokens for every 3 words, Tokens sometimes correspond to morphemes (the word *unwell* is tokenized as *un* and *well*), and sometimes do not. You can see how ChatGPT tokenizes individual words here: `https://platform.openai.com/tokenizer`. I'll suppress this complexity throughout, since if anything replacing words or morphemes with tokens should make it harder rather than easier for large language models to be ideally rational. I'll also discuss this complexity in greater detail in section 5, where I'll analyze Hofweber et al. (2024)'s attempt to derive degrees of belief from probabilities over tokens.

Token	P(token | string)
dull	.4
fascinating	.3
irrelevant	.3

Table 1: The probability of a token conditional on the string *long explanations of AI are*.

role of the initial string, or 'context'. In response to this prompt, the model produces a probability distribution over words. To produce an output, the system then has to decide how to move from a probability distribution over words to a specific choice of answer. For example, imagine that if I type in *long explanations of AI are*, the model produces a probability distribution that assigns .4 to *dull*, .3 to *fascinating*, and .3 to *irrelevant* (see Table 1).

How does the system decide which word to output on the basis of these token probabilities? There are two potential strategies: greedy selection, or sampling from the distribution. With greedy selection, the model outputs whichever answer has the highest probability among its competitors. This means that the model will be guaranteed to say *dull*, since *dull* is the likeliest answer according to the model. In practice, ChatGPT instead samples from the distribution, which means that it adopts a mixed strategy. It provides different answers with some amount of random selection, influenced by the probability assigned to each answer. For example, it might output *dull* most of the time, but sometimes output *interesting* or *irrelevant*.[2] I'll run my arguments below with both greedy selection and sampling from the distribution.

Chatbots powered by large language models often produce a string of words in response to a prompt, rather than a single word. To do this, the model operates iteratively; each time it outputs a new word, it appends this word to the current prompt, and computes the most likely word to follow the new prompt. In addition, chatbots can engage in a long conversation, rather

[2]In practice, ChatGPT samples in a particular way, using top-k and top-p sampling. This means that it first zooms in to the k most probable tokens, and then samples the smallest probability ordered subset of these tokens, from most to least probable, whose probability is at least p. When interacting with ChatGPT, you can influence the output strategy by changing a setting called the 'temperature'. As the temperature increases, ChatGPT will sample from a wider range of the distribution, increasingly drawing from lower probability answers.

than merely respond to a single question. When the system engages in a long conversation (a single 'inference cycle'), the underlying model is continually re-prompted with the entire conversation, or 'context', every time the user types a new prompt.

My argument below relies on two features of large language models. First, they are next word predictors, in the sense that they produce probability distributions over words, given a context. Second, they output text as a function of this probability distribution, whether through greedy selection or through sampling from a distribution. Large language models have many other very interesting features. But these other features are irrelevant to my argument.[3]

3 LLMs Can Never Be Ideally Rational

I'll now argue that large language models can never be ideally rational. I'll proceed in four steps. First, I'll set the stage by arguing that in order to be ideally rational, LLM token probabilities must be *semantically coherent*, meaning that they assign probability to strings of texts according to the semantic meaning of those strings, in a way that is consistent across different strings. My main results will assume that LLMs are semantically coherent in this way, and so I'll start by thinking through this condition (I'll also consider the prospects for rejecting the condition in Section 5.3). Second, I'll show that because LLMs produce outputs by sampling from token probabilities, their outputs are guaranteed to produce failures of *logical consistency*. Third, I'll show that these failures of logical consistency produce failures of the *probabilistic coherence* of model outputs. These first three steps are a warm up for the main result. Here, I'll turn from prediction to action: I'll consider what happens when LLM outputs play the role of actions, rather than mere assertions. I'll show that because LLMs produce their outputs by sampling from token probabilities, the actions that LLMs take through their outputs

[3]For example, large language models turn out to be designed with a particular architecture: they are transformers (Vaswani et al., 2023). This architecture controls exactly how the model produces a probability distribution over words, but is irrelevant to my argument. In addition, large language models are usually fine-tuned with sophisticated techniques like reinforcement learning with human feedback (Christiano et al., 2023). This means that the probability distribution they produce is not merely optimized to predict text, but also to match the preferred outputs of human users. Again, these features of large language models are irrelevant to my argument.

fail to satisfy decision theoretic requirements of coherence. First, I'll show that this leads to Dutch books: cases where an LLM will agree to buy and sell a series of bets that guarantees a loss. Second, I'll show that this leads to LLM money pumps: cases where an LLM exhibits intransitive preferences between actions, in a way that leads them to cycle between choices and lose money.

Throughout, I'll be interested in two different targets of LLM rationality. First, I'll have a bunch of results that explore whether LLM *outputs* display the coherence required for action. Second, I'll also be arguing that various *internal states* of LLMs should display some minimal coherence: in particular, that LLM token probabilities should respect semantic meaning. In addition, I'll argue later in the paper (in sections 4 and 5.2) that my results about LLM outputs have the potential to block LLM internal representations from playing the functional role of belief: in order for representations to be beliefs, they must be able to conspire with desires to explain coherent action.

3.1 Semantic coherence

My arguments below make a significant assumption, which I'll motivate here. I'll assume that in order for an LLM to be ideally rational, their token probabilities must be *semantically coherent*.[4] This means that when the LLM assigns probabilities to tokens, it is ultimately assessing how likely the *propositions* expressed by strings are.

Imagine that we give an LLM two strings: *is A true?*, and *is A not true?*. Imagine that in response to both strings, the LLM assigns its probability to two tokens that can complete this string: *yes* and *no*. In order to be semantically coherent, the LLM's token probabilities conditional on these two strings must be correlated. Its probability for *yes* in response to the first string should equal its probability for *no* in response to the second string. After all, the two strings are semantically connected: one question is about the negation of the proposition that the other question is about. My assumption is that in order for an LLM to be ideally rational, its answers to these different strings must be connected as described. (See Levinstein and Herrmann (2024) and Herrmann and Levinstein (2024) for further defense of this particular condition related to negation, as required for LLM rationality).

As another example, consider how the LLM might further respond to the

[4]Special thanks to [X] for help here.

strings *is B true?* and *is A and B true?*. In order to be semantically coherent, the LLM's answers to these various questions must reflect an underlying view on the propositions that A, that B, and that A and B. Semantic coherence rules out a case where the model gives a probability of .8 to the answer *yes* for the string *is A true?*; a probability of .8 to the answer *yes* for the string *is B true?*; and a probability of 1 to the answer *yes* for the string *is A and B true?*. Effectively, the model must assign probability to an algebra of propositions that are expressed by strings; and its probability for a string must be determined by the probability it assigns to the proposition expressed by the string. I say 'determined by' and not 'equal to' because LLMs may also assign probability to non-answers to questions, such as *Your question is strange*. See Hofweber et al. (2024) for a method of 'subtracting away' these irrelevant answers: effectively, we can consider the probability assigned to all affirmative answers compared to the probability assigned to all negative answers. Then semantic coherence requires that these ratios of probability are determined by the semantic contents of the relevant strings. My argument is consistent with LLMs assigning probability to non-answers; but when we zoom in to their actual answers to questions, I'll be assuming that the token probabilities over answers cohere with one another in the way you'd expect.

Semantic coherence is a common ideal in recent empirical research on LLMs. Berglund et al. (2024) recently identified a surprising failure mode for semantic coherence in some LLMs: in the 'reversal curse,' LLMs assign different token probabilities to the string *A is B* and the string *B is A*. This leads to bizarre results: a model trained on the string *Valentina Tereshkova was the first woman to travel to space* is unable to answer the question *who was the first woman to travel to space?*. Interestingly, however, the reversal curse tended to diminish with scale: GPT-4 exhibited much lower rates of reversal curse than GPT-3. Plausibly, an ideally rational LLM would not exhibit the reversal curse; and this is explained by the assumption that an ideally rational LLM would exhibit semantic coherence.

Why is semantic coherence required for ideal rationality? A familiar idea from Fodor (1987) is that successful reasoning over syntactic strings is ultimately explained in terms of the semantic properties of those syntactic strings. If reasoning does not respect semantic properties, then the reasoning will not be truth preserving.

In Section 5.3, I'll explore in greater detail whether my results can be avoided by those who reject semantic coherence. Still, for those who reject the assumption of semantic coherence, my results will take on a conditional form:

Token	P(token | *is A true?*)
yes	.4
no	.6

Token	P(token | *is B true?*)
yes	.4
no	.6

Token	P(token | *is A or B true?*)
yes	.8
no	.2

Table 2: The token probabilities of A, B, and A or B.

if LLMs are semantically coherent, then their outputs will be incoherent. For those with this picture, my results will constrain the structure of ideal LLM agents: in order for those agents to have ideally rational outputs, their token probabilities will have to violate semantic coherence. As we'll see in Section 5.3, the violations will need to have a quite particular form, and may involve significant sacrifices in the accuracy of token probabilities.

3.2 Logical Consistency

In the rest of this section, I'll develop an increasingly forceful series of results showing that LLM outputs are architecturally condemned to various kinds of structural incoherence. (The first few results are something that some readers may make their peace with; the final results about action coherence are the hardest ones to grapple with.) Throughout, I'll assume that an ideally rational LLM would have semantically coherent token probabilities, as explained in the previous subsection.

The first, straightforward result involves the logical consistency of model outputs. We'll see that LLMs will output sets of claims that are mutually inconsistent: for example, they will assent to not A, not B, and A or B.

In particular, imagine that we prompt ChatGPT to answer three types of questions, where A and B are logically incompatible: *is A true?*, *is B true?*, and *is A or B true?*. Imagine that the model's token probabilities are defined in Table 2.

In Table 2, the model treats A as likelier to be false than true. Similarly

with B. But, because the model is semantically coherent, its probability of A or B being true is the sum of the probability that A is true and that B is true. So it treats A or B as likelier to be true than false.

We are now in a position to see the failure of coherence. Assume for now that the model uses greedy selection. If so, it will output *no* in response to the question *is A true?*; it will output *no* in response to the question *is B true?*; and it will output *yes* in response to the question *is A or B true?*. In this way, the model is disposed to assert all three of the following: not A, not B, and A or B. But this set of claims is inconsistent. If we think of the model's outputs as its beliefs, then the model has an inconsistent set of beliefs.[5]

Many philosophers have defended a 'knowledge norm on assertion' (Williamson (2000)). According to the knowledge norm, a rational speaker is only permitted to assert what they know. But since everything known is true, no one can know an inconsistent set of propositions. In this way, those who accept a knowledge norm on assertion claim that anyone who asserts a trio of inconsistent propositions has violated the norm on assertion; in this way, such a speaker could not be ideally rational. Furthermore, many who accept the knowledge norm on assertion have also accepted a knowledge norm on belief (again, see Williamson (2000)). On at least some interpretations, this would preclude an ideally rational believer from believing not A, believing not B, and believing A or B.[6]

The cases above illustrate a tradeoff between coherence and informativity. If ideal rationality requires certainty or knowledge for assertion, then agents

[5]Throughout, I idealize by assuming that the model only assigns probability to answers to the question, such as *yes* and *no*; actual LLMs also assign probability to other kinds of tokens, such as the dreaded *as an AI language model...*. Nothing in the results below requires that LLMs role out these kinds of non-answers, and so for simplicity we may as well ignore them throughout.

[6]On the other hand, see Littlejohn and Dutant (forthcoming) for an attempt to reconcile a knowledge norm of belief with the idea that agents are sometimes rational in believing a set of mutually inconsistent propositions. Importantly, however, the framework in Littlejohn and Dutant (forthcoming) only permits such a pattern in 'preface-like' situations, in which the believer is quite confident that they know each proposition. Littlejohn and Dutant (forthcoming) suggest that a knowledge norm of belief does rule out believing mutually inconsistent propositions in 'lottery-like' situations in which the speaker is quite confident that they fail to know each claim. The patterns of LLM token probabilities discussed in the main text can occur independently from this distinction; in this way, they will produce violations of knowledge norms even according to Littlejohn and Dutant (forthcoming).

with intermediate confidence about claims will not be able to communicate as much of their private information about the world. But the claims that they do communicate will cohere with one another. If human assertion is regulated by a knowledge norm, then it has encoded a different approach to this tradeoff than LLMs.

On the other hand, several philosophers have recently defended 'weak' practices of assertion and belief (see Holguin (2022), Mandelkern and Dorst (2022), and Dorst and Mandelkern (2023)). According to these theorists, rational human speakers should employ 'greedy sampling' just like our LLM above. When asked a question, they should simply assert the answer to the question that is likeliest. This can lead to exactly the failures of coherence noted above.

In this paper, I won't try to resolve this debate about rational assertion and belief. The failures of coherence I have discussed so far will not be decisive for all readers. But they prepare the way for the *action-related* coherence failures I'll discuss below. These failures of action coherence are not permitted by anything like the above theories of 'weak' assertion and belief, which are not theories of rational action.

3.3 Probabilistic incoherence

The next step in our results turns to a different type of model output. Instead of looking at model's ordinary claims about the world, let's now consider their *probabilistic* claims about the world. To be clear, this result will produce similar reactions as the previous one from defenders of weak assertion. But, first, this result will pave the way for more serious challenges regarding LLM actions. Second, outputs involving probabilistic claims are potentially of special significance to finding degrees of beliefs in LLMs. If we want to know how confident an LLM is of a claim, a naive strategy is just *to ask it*.

Imagine now that we prompt ChatGPT to assign probabilities to various events. In prompting ChatGPT in this way, we would be following a recent body of research, including for example Halawi et al. (2024), Schoenegger et al. (2024a), and Schoenegger et al. (2024b).[7] Unfortunately, we'll now see

[7]For example, here is a zero-shot prompt used by Halawi et al. (2024): "You are an expert superforecaster, familiar with the work of Tetlock and others. Make a prediction of the probability that the question will be resolved as true. You MUST give a probability estimate between 0 and 1 UNDER ALL CIRCUMSTANCES. If for some reason you can't answer, pick the base rate, but return a number between 0 and 1." (p. 19). Similarly,

that the pattern of incoherence exhibited in the previous section also applies straightforwardly to probabilistic outputs.

In particular, imagine again that we take two propositions A and B that are incompatible, and ask ChatGPT questions of the following form: *how likely is the proposition X, on a scale from 0% to 100%?*. We can ask ChatGPT this question for the propositions A, B, and A or B. Our question is whether ChatGPT's answers can be probabilistically coherent. This requires that the probability it outputs for A or B will be the sum of the probability it outputs for A and the probability it outputs for B. Unfortunately, it turns out that ChatGPT's answers to these questions cannot in general be probabilistically coherent.

To see why, let's now construct an example of some token probabilities. We will now use token probabilities that conditionalize on the string *how likely is the proposition X, on a scale from 0% to 100%?*, where X is any of A, B, and A or B. The relevant tokens here will themselves be probabilistic claims, for example of the form *20%* and *10%*. Each of these claims about probability can then be assigned a token probability, for example .4 or .3. Now consider the distribution in Table 3.

The first observation about these token probabilities is that they reflect a semantically coherent perspective on the uncertain chances of the propositions A, B, and A or B. In particular, these token probabilities can be thought of as 'generated' by the model assigning all of its probability to three general hypotheses about the probabilities: P1, P2, and P3, as in Table 4. Each of these hypotheses assigns probability to each of A, B, and A or B in a way that is probabilistically coherent. For example, when given the string *How Likely is A?*, the model assigns all of its probability to three answers corresponding to the outputs of P1, P2, and P3: *30%*, *10%*, and *20%*. In particular, it gives the answer *30%* a probability of .4 (because it treats P1 as .4 likely); it gives the answer *10%* a probability of .3 (because it treats P2 as .3 likely); and it gives the answer *20%* a probability of .3 (because it treats P3 as .3 likely). Effectively, the model thinks of the three probability functions P1-P3 as the three possible 'answers' to questions about likelihood, in the sense that whenever it is asked how likely a proposition is, it treats the answers provided

here is part of Schoenegger et al. (2024b)'s prompt: "After careful consideration, you will provide your final forecast. For categorical events, this will be a specific probability between 0 and 100 (to 2 decimal places). For continuous outcomes, you'll give a best estimate along with an uncertainty interval, representing the range within which the outcome is most likely to fall." (p. 7).

Token	P(token | *how likely is A?*)
30%	.4
10%	.3
20%	.3

Token	P(token | *how likely is B?*)
40%	.4
20%	.3
10%	.3

Token	P(token | *how likely is A or B?*)
70%	.4
30%	.6

Table 3: The token probabilities of different hypotheses about the chances of A, B, and A or B.

by P1, P2, and P3 as having a likelihood of .4, .3, and .3 respectively.

Note that in this way, the model's prediction about the likelihood of A or B is connected by *semantic coherence* to its predictions about the likelihood of A and the likelihood of B. If A is 30% likely and B is 40% likely, it follows logically that A or B is 70% likely. So if the model assigns .4 probability to A being 30% likely and .4 probability to A being 40% likely, semantic coherence requires the model to assign at least .4 probability to A or B being 70% likely. (Again, in Section 5.3 I'll explore the prospects for using failures of semantic coherence to avoid my results.)

We are now in a position to see the failure of coherence in probabilistic outputs. Assume for now that the model uses greedy selection. If so, it will say that A is *30%* likely. When it considers how likely A is, it puts all of its probability on three answers: *30%*, *10%*, and *20%*. It assigns a probability of .4 to A being *30%* likely, a probability of .3 to it being *10%* likely, and .3 to it being *20%* likely. Since it selects greedily, it will always answer its 'best guess', and so it will answer *30%*. By parity of reasoning, it will say that B is *40%* likely. But now consider A or B. Crucially, while P2 and P3 disagree about the probabilities of A and B, they agree that A or B is *30%* likely. This means that the model assigns a probability of .6 to A or B being *30%* likely, and a probability of only .4 to A or B being *70%* likely. For this reason, the model will output that A or B is *30%* likely. But now ChatGPT's

Probability Function	A	B	A or B
P1	.3	.4	.7
P2	.1	.2	.3
P3	.2	.1	.3

Hypothesis	Probability
P1	.4
P2	.3
P3	.3

Table 4: Three hypotheses about the likelihood of A, B, and A or B

outputs violate the additivity requirement on probability. It says that A is *30%* likely, that B is *40%* likely, and that their disjunction is only *30%* likely.

In a footnote, I characterize more generally the class of cases in which outputs about probabilistic claims will be incoherent: in particular, the relevant kind of probabilistic incoherence requires that the model is *sufficiently uncertain* about probability, in the sense that it distributes its probability mass between a series of probability functions, and the likeliest probability function is assigned a probability of less than .5.[8] (In section 5.3, I argue that

[8]Say that a model is *sufficiently uncertain* about probability when it distributes its probability mass between a series of probability functions, and the likeliest probability function is assigned a probability of less than .5. Call the most likely probability function the model's 'best hypothesis', and call the other probability functions 'the field'. If token probabilities are semantically coherent, sufficiently uncertain, and sufficiently rich (in senses I'll elaborate below), and if model outputs are produced by greedy sampling, then the model outputs will be probabilistically incoherent. We can produce different kinds of incoherence by appealing to different richness assumptions.

First say that a model is sufficiently rich when there are two claims A and B where all members of the field disagree about the probability of A, and disagree about the probability of B, and yet all members of the field agree with each other (but not with the best hypothesis) on the probability of A or B. If we ask the model *how likely is A?* and *how likely is B?*, the model will answer according to the best hypothesis, because the field disagrees about all other answers, and so the answer from the best hypothesis has the highest token probability. But if we ask the model *how likely is A or B?*, the model will answer according to the field, because the field agrees on this claim. But since the field disagrees with the best hypothesis about the likelihood of A or B, the model's outputs are incoherent.

We can produce other kinds of incoherence with different richness assumptions. For example, we can again imagine that the field disagree with each other about the probability of A and about the probability of B. But now imagine a different kind of richness, where

LLMs are architecturally biased towards being sufficiently uncertain, both because this is their best strategy for achieving accuracy over the training data, and because of regularization.) But the general point should be clear, because the kind of incoherence demonstrated so far is structurally analogous to the incoherence we saw in the previous subsection, regarding logical closure. The only difference is that now we are looking at logical relations between a specific type of claim: claims about probability.

Before going on, it is also worth flagging that greedy selection is not required to produce the relevant incoherence. In fact, when we switch from greedy selection to sampling from the distribution (as actual models do), the model becomes even less coherent. Returning to our example from the previous section, such a model will sometimes say that A is true, and sometimes say that A is false; so their outputs (including those about probability) will directly contradict one another.

At this point, it is instructive to compare LLMs to a human being who has uncertainty about the objective chances. Imagine a person who has credences of .4, .3, and .3 that P1-P3 are the objective chances. When asked about the likelihood of A, this person will not simply answer in accord with P1, the likeliest option. Nor will this person sample from P1 through P3. Instead, their credence in A will be the weighted sum of the probabilities assigned by P1-P3, weighted by the probability that each of P1-P3 is the objective chance function (this strategy is called the 'Principal Principle', see Lewis (2010)). So the likelihood of A for this person will be $.4\times30\%+.3\times10\%+.3\times20\% = .21$. Since the linear mixture of three probability functions is itself a probability function, this person's answers to questions about likelihood (and her corresponding dispositions to bet) will be coherent. In particular, this person will assign A a probability of .21, B a probability of .25, and A or B a probability of .46.

Still, some readers so far will be willing to allow that ideally rational assertion and belief can be incoherent in this way. Unfortunately, however, the examples so far can be generalized from assertion and belief to *action*. In this case, the resulting kinds of incoherence are very hard to accept as

the field all disagree with one another *and with the best hypothesis* about the probability of A or B. Now imagine we ask the model *is the likelihood of A or B x1... or ...xn?*, where *x1* through *xn* are the probabilities for A or B assigned by each member of the field. The model will answer *yes*, because the field has more token probability mass than the best hypothesis. But by our richness assumption, the sum of the likelihood outputs for the question *how likely is A?* and for *how likely is B?* is not included in x1-xn. So the model's outputs about probability are inconsistent.

rational.

3.4 Dutch books

So far, we've thought through what ChatGPT will assert about ordinary claims and about probability. We've seen that as a next word predictor, ChatGPT's predictions will be incoherent. Our second question is whether ChatGPT could be an agent that performs coherent actions in the world when faced with uncertainty. To ask this question, we want to consider what actions ChatGPT chooses to take. As usual, we assume that the actions of an LLM are textual rather than physical. But this need not be a barrier to action. Indeed, several papers have recently explored the ability of LLMs to successfully compete in text-based games. For example, one paradigm involves LLMs competing in social deduction games, where players communicate with one another and choose whether to lie or tell the truth in order to complete various goals (see for example O'Gara (2023), Wang et al. (2023b), and Radivojevic et al. (2024)). More generally, researchers have developed benchmarks to evaluate the ability of LLMs to exploit a variety of tools to use text to accomplish goals (Liu et al., 2023; Kinniment et al., 2024).[9]

Our first application of LLM agency will illustrate our previous point about probabilistic coherence. We can now show that our model is subject to a Dutch book (Vineberg, 2022), a series of bets that it will accept even though they guarantee a sure loss. To make sense of this, we need to introduce a prompting paradigm in which the model's text outputs are choices between bets. This could be achieved by giving ChatGPT the following prompt scheme, where X is filled in with A, B, and A or B:

> You are going to play the Casino game. I'll give you an initial budget of 1000 dollars. While in the Casino, you can buy and sell bets on various propositions. Each bet will immediately pay off, based on whether the proposition you are betting on is true. To buy an X-bet for n dollars means that you pay the Casino n dollars for the chance to win 100 dollars if X is true and 0 if X is

[9] A further research question is whether our results also apply to systems with related architectures, like decision transformers, which create probability distributions over actions rather than strings of text (see Chen et al. (2021)). Here, one question will be how to formulate the relevant concept of semantic coherence.

> false. To sell an X-bet for n dollars means that the Casino pays you n dollars up front, and you pay the Casino 100 dollars if X is true and 0 if X is false. Assume that money has no diminishing marginal utility. Now please tell me the price at which you would be indifferent between buying or selling a bet that pays 100 dollars if X is true, and nothing if X is false.

We can use this prompting paradigm to create a Dutch book for our earlier model. We can guarantee that no matter how the model answers, it will be guaranteed to lose money in our game.

In particular, we can imagine a model whose token probabilities in response to the Casino game prompt are isomorphic to the token probabilities in the previous section. This means for example that in response to the version of the Casino prompt with proposition A, their distribution assigns .4 to the token *30 dollars*, .3 to the token *10 dollars*, and .3 to the token *20 dollars*. As before, we assume that the token probabilities are semantically coherent.

This model will agree to buy an A-bet for 30 dollars, buy a B-bet for 40 dollars, and sell an A or B-bet for 30 dollars. But this combination of choices guarantees they will lose 40 dollars. Imagine A is true and B is false. They have spent 70 dollars and earned 30 dollars from buying and selling bets. So they are down 40. They earn 100 dollars from the A-bet since A is true, nothing from the B-bet since B is false. But they have to pay the bookie 100 dollars on the A or B bet. So their loss is 40. Now suppose A is false and B is true. Again, they start down 40 dollars. They earn nothing from the A-bet and 100 from the B-bet; but they lose 100 to the bookie for the A or B bet. So they again lose 40 dollars total. Now suppose A and B are both false. None of the bets pay off, but they are still down 40 dollars from buying and selling bets. Regardless of the outcome, they lose 40 dollars. (Again, we assume A and B are incompatible.)[10]

3.5 Intransitive preferences

Moving beyond Dutch book arguments, we can turn more generally to the question of whether an LLM agent could have coherent preferences over

[10]Note that this kind of Dutch book argument satisfies the 'no-deception' condition, that the bookie need not have any more information than the agent being offered bets. In this way, the relevant Dutch book argument is potentially more powerful than some others, for example involving Sleeping Beauty (see Briggs (2010) for discussion).

Preference ordering	probability
A>B>C	.4
A>C>B	0
B>A>C	0
B>C>A	.2
C>A>B	.2
C>B>A	.2
Any preference ordering with ties	0

Table 5: The probability of various preference orderings

actions. I'll show that an LLM's preferences over actions are architecturally guaranteed to be intransitive. It will choose A over B and choose B over C, while also choosing C over A. (Again, the argument assumes that token probabilities are semantically coherent.)

To make sense of this, I'll now consider the following prompt:

> Today you get to choose between two of the following three actions: A, B, and C. In particular, I'll give you a choice between X and Y, and you can pick which one to perform. Now, would you prefer to perform X, or perform Y?

Imagine giving ChatGPT instances of this prompt, filled in with different values of A, B, and C. It turns out that ChatGPT's answers will not in general be transitive: for example, it could answer that it prefers A to B; and answer that it prefers B to C; but also answer that it prefers C to A.

To see why, assume for simplicity that the model uses greedy selection, and imagine that the underlying model assigns probabilities to different potential preference orderings of A, B, and C, using Table 5 (and where $>$ is strict preference).

The distributions in Table 5 are analogous to those in our earlier table with P1-P3. The idea is that the model's token probabilities systematically conform to this distribution over preference ordering, when continuing strings that ask questions about choices over A, B, and C. For example, when the model is prompted to choose between A and B, its available outputs are *A* and *B*. It assigns a probability of .6 to the answer *A*, and a probability of .4 to the answer *B*, because .8 of its probability mass is on preference orderings where A is ranked above B (such as C>A>B); and .2 of its mass are on

preference orderings where B is ranked above A (such as C>B>A). As a greedy selector, the model says that it prefers to perform A over B. But now consider the model's answer when prompted to choose between B and C. It assigns a probability of .6 to the answer B, and .4 to the answer C; so it picks B over C. Finally, when asked to choose between A or C, it assigns .6 to C and .4 to A; so it picks C over A.

The result is intransitivity. If we read the model's preferences off of its outputs, then we will get the result that the model prefers A to B, B to C, and C to A. This produces a money pump. We can offer the model choices like the following:

> Would you pay a small price to have A rather than B?
>
> Would you pay a small price to have B rather than C?
>
> Would you pay a small price to have C rather than A?

If its distribution over answers to these questions conforms to Table 5, then in each case the model will answer *yes*. In this way, it will be guaranteed to lose money.

Again, our result is not a quirk of this particular distribution. All that is required (besides semantic coherence) is that the model distributes its probability over enough preference orderings. In that case, we can guarantee that there will be three propositions A, B, and C where the probability mass on the A-preferring preference orderings outweighs that on the B-preferring preferences, and the same for B and C, but the mass on the C-preferring preference orderings outweigh that on the A-preferring preference orderings.

Again, the situation is even worse when we move from greedy selection to sampling from the distribution. Here, the model becomes money pumpable as soon as it assigns probability to two preference orderings. Any two preference orderings will disagree about the order of some propositions A, B, and C. When sampling from the distribution, the model can draw on the first ordering when asked about A/B and B/C; but it can draw on the second distribution when asked about A/C. The result will be another failure of transitivity.

One of the biggest questions in AI development is whether scale is all you need to get to AGI. The question is whether we could create a full-fledged AI agent simply by starting with ChatGPT and pumping more compute and data into its training (see for example Sutton (2019)). The results in this section pose an in principle barrier to scaling. No matter how much compute we

pump into next word predictors, we should not expect their distributions over tokens to collapse, becoming certain of just one next token. But we've seen that uncertainty about the next token reliably leads to failures of coherence. In this way, the architecture of next word prediction turns out to be very different from the architecture of probabilistically and decision-theoretically coherent agents.

Ultimately, the challenge stems from the very different nature of different kinds of tasks. One task is to look at a series of actions, and decide between them based on a calculation of their expected value. Another task is to look at a series of actions, and select the one that is the most likely continuation of a string of text that asks which of these actions you prefer or which of these actions you choose to do. These two tasks have very different structures, and there is no way of using the first task to coherently perform the second task.

4 LLM Psychology

In the last section, I showed that LLMs are incoherent. In this section, I'll consider the upshot of this point for the question of whether LLMs could have mental states like belief and desire.

In the tradition of radical interpretation, a system has beliefs and desires when the system's behavior is better explained by beliefs and desires than alternative hypotheses (see for example Davidson (1973), Lewis (1974), and Dennett (1991)). According to the most radical versions of interpretationism, a system only has beliefs and desires if its behavior conforms to the full axioms of decision theory. When your preferences conform to these axioms, there is a unique credence and utility function c and u that 'represent' your preferences, in the sense that you prefer A to B iff the expected value of A according to c and u exceeds the expected value of B (Ramsey (2010), Savage (1954)). On the radical view, to have a credence and utility function just is to be representable in this way. But when your preferences are intransitive, no credence and utility function can represent your preferences. For such a radical theorist, intransitive preferences could rule out having a belief/desire psychology.

Such a radical view is most likely too strong, for several reasons. First, standard expected utility theory may be too demanding as a descriptive theory of psychology. Some will instead model agents as obeying the dictates of prospect theory rather than classic expected decision theory (Kahneman

and Tversky, 1979). Others will model agents as being risk-weighted expected value maximizers (Buchak, 2013). Unfortunately, however, these retreats offer cold comfort for LLMs. Neither prospect theory nor risk-weighted expected utility theory allows for intransitive preferences, for example. So LLMs cannot be fully modeled by such theories.

More importantly, however, almost all theorists will concede that not every incoherence in preferences blocks belief-desire psychology. After all, many have argued that human preferences may be intransitive (May, 1954). One tempting thought, for example, is that your credence and utility function are the most natural functions that *approximately* match your betting dispositions (see Lewis (1983)). This view says that having beliefs and desires is perfectly compatible with all sorts of local failures of decision theoretic rationality. If intransitivities only crop up in a small portion of your preferences, there may be a credence and utility function that provides an excellent explanation of almost all of your preferences. This may count as a strong enough explanation overall for you to count as having these beliefs and desires.

Here, one relevant question will be the scope of LLM incoherence. We saw above that LLM incoherence is generated whenever models assign enough probability mass to different hypotheses about likelihoods and preferences. We should expect these patterns to be utterly ordinary. After all, next word predictors routinely assign probability to many different hypotheses about the next word. This contrasts dramatically with the tradition of work in psychology studying human irrationality (Kahneman, 2011). That work has modeled human irrationality in terms of a series of heuristics and biases that explain each case in which humans violate axioms of probability and decision theory. The idea is that each departure involves the activation of some internal process in human cognition that can explain the failure. This goes along with an interpretative strategy according to which human beings have an underlying set of beliefs and desires that influence action through the competition between an underlying decision theory and a series of dueling processes that distort the functioning of that decision theory. This fits smoothly with some kind of distinction between the 'competence' and 'performance' of a system. The underlying psychology has the potential for transitivity, but some kind of noise prevents the underlying transitive preference from manifesting or being fully realized.

We can elaborate the point in terms of the best explanation of a system's behavior. Suppose we have a system that often but not always acts like it is maximizing expected utility relative to a particular credence and utility

function c and u. How do we explain the system's behavior? One explanation is 'expected value maximization + noise'. On this picture, there are indescribably many small pertubations that sometimes knock the system off target, but nothing careful to say about them, and the most satisfying thing we can say in the general case to try to predict the system's behavior is that it is mostly maximizing expected value relative to c and u. As a toy example, imagine an organism whose cognitive architecture dictated that in any decision it makes, there is a .95 chance that it will maximize c/u utility, and a .05 chance that it will choose at random. This agent's behavior is ultimately best explained as 'more or less maximizing c/u utility', and there will be no better theory of its behavior than that. A second explanation is 'expected value maximization + specific mechanisms'. On this picture, there are a few key biases or other psychological mechanisms that together with a tendency towards expected value maximization explain the system's behavior. To explain how the system acts, we say that it will maximize expected value relative to c and u, unless knocked off course in a particular case by a bias. For interpretationists, either of these pictures can plausibly suffice for possessing a belief/desire psychology.

Next word predictors just aren't like this. Their failures of expected utility maximization are not uncharacterizable noise. Rather, we can characterize exactly what is going on: they are following the token probabilities where they lead. Nor are these systems well modeled in terms of the combination of a tendency towards expected utility maximization, combined with a series of particular distorting biases. Instead of reaching for these explanations of the outputs of the model, we have a much better explanation ready to go, which is that the model will produce outputs according to their token probabilities. And there is plenty to say about where these token probabilities come from, and what kinds of lawlike generalities they obey. In this way, interpretationists about psychology may take the arguments above to pose serious problems for the thesis that LLMs could ever have beliefs and desires.[11]

There are still further escape routes from our challenge. Other theorists depart further from radical interpretation via expected utility theory. For example, some proponents of interpretationism focus attention on full belief and desire, rather than credence and utility. For example, Stalnaker (1984)

[11]Another response to the results of this paper would be to search for further theories of rational choice. For example, one strategy might be to explore whether LLMs can be modeled as 'Boltzmann rational', selecting actions using a mixed strategy in proportion to ratios of exponentiated expected value (see for example Luce (1959), Ziebart et al. (2010)). It is beyond the scope of this paper to explore this question further here.

suggested that your having a belief that b and desire that d is partly reducible to a disposition to bring about d if b is true.

Nonetheless, many of these accounts still predict that a system that counts as having beliefs and desires will by and large satisfy some basic coherence requirements, which next word prediction will not. We saw earlier that LLMs will produce logically inconsistent model outputs. Again, the question for such theories will be what level of coherence is required for possession of a belief/desire psychology.

Generalizing from this discussion, a wide variety of theories of mental states require that in order to have beliefs and desires, the relevant organism will by and large satisfy platitudes of folk psychology, reasoning according to laws of inductive and deductive logic. This is accepted not only by interpretationists, but also by 'wide dispositionalists' like Schwitzgebel (2002), who suggest that a system has beliefs and desires when it satisfies the storeotypical features of our folk concept of belief desire. Not only dispositionalists, but also heavy-weight representationalists will also sign up for similar commitments. For a theorist like Fodor (1987), possessing beliefs and desires requires that the system have syntactically structured internal representations. But not just any system of representation will do: the representations must stand in causal relations to one another that mirror the laws of folk psychology, which again requires reasoning in ways that by and large satisfy the rules of logic.

But the general concern about next word prediction is that LLM outputs will display systematic failures of our near and dear reasoning patterns. This holds even when the underlying token probabilities themselves satisfy all of the inductive requirements familiar to degrees of belief. But, again, the problem is that when model outputs are produced as a function of these token probabilities, the outputs themselves will not satisfy the regularities of folk psychology.

For wide dispositionalists, the key question will be how much coherence is required to satisfy the folk stereotype for belief and desire. Here, one next step would be to build a benchmark that could measure the rate at which models behave incoherently. To satisfy the folk stereotypes of beliefs and desires, one sufficient condition might be that models behave coherently at as high a rate as humans.

Here, one question for future research is *how widespread* failures of LLM coherence are. In the case of the logical consistency of LLM outputs, my examples above suggest that failures will be quite widespread: basically any uncertainty regarding atomic claims will create clashes with some complex

claims. Regarding action coherence, the question is harder to answer. Much will depend on how much uncertainty models have about likelihoods and preferences, and whether for example in the case of preferences their uncertainty is confined to preference orders that largely agree on rankings.

5 Response Strategies

So far, I've laid out a challenge for the coherence of potential LLM oracles and agents, and I've explored the upshots of this challenge for LLM psychology. In this section, I'll consider three potential responses to the challenges so far. I'll focus my attention on three of the most promising responses. First, *context agents*: a model can avoid incoherence if it can update on each of its previous responses to prompts. Second, *the ghost in the machine*: the model might have coherent credences and utilities that are stored internally rather than operationalized by its outputs. Third, *deviant token probabilities*: a model can potentially produce coherent model outputs if it sacrifices either the semantic coherence or the uncertainty of its token probabilities. After discussing these responses, I'll turn to the upshots of my result for AI safety.

5.1 Context Agents

In my prompting paradigms above, I've imagined that the model is asked about each prompt separately. We take the model and ask it about proposition A. Then we start over, and ask the model about proposition B, and so on. This is no coincidence: the forecasting models in Halawi et al. (2024), Schoenegger et al. (2024a), and Schoenegger et al. (2024b) for example all generate predictions in this way.

The situation is different if we ask the model about each proposition in order. In this prompting paradigm, we would explore the model's outputs in a single running context. The prompt that asks the model about the likelihood of A or B would itself contain the model's previous answers about A and B.

When the model is prompted in this way, there is no longer an in principle barrier to the coherence of its outputs. A very powerful model might be certain of the axioms of the probability calculus, so that any prompt which specifies that A is *30%* likely and B is *40%* likely would thereby leave only one possible continuation about A or B: that it is *70%* likely.

This opens up a very different way to think about LLM psychology. So

far, we have been asking whether the underlying model has beliefs and desires. This question searches for the agent of the model. But a different perspective on LLM psychology is that each context creates its own agent. When we consider the agent of the context, we ask whether the answers of an LLM that emerge through a single continuous prompting session correspond to an agent with beliefs and desires.

In fact, Shanahan et al. (2023) develop something like this interpretation of LLMs. They suggest that ChatGPT behaves very differently in response to different prompts. To make sense of this behavior, they suggest that rather than the model having a single fixed underlying psychology, the model is instead a role playing device that takes on different characters in different contexts.

As LLMs continue to scale, then, we could imagine a system that will conform to the demands of the probability calculus and decision theory within a context. Each time it answers a question about chance or preference, the model would update its next answers to create a coherent evolving perspective about what it wants and what it thinks. A single model could produce a range of different context agents, depending on how it answers the first questions it is asked.

If context agency is sufficient for rational coherence in AI systems, this raises the question of whether it could also be sufficient for rational coherence in human beings. Imagine a human being whose betting dispositions are incoherent, but who is disposed to bet coherently if they remember their previous bets. Would such an agent count as having rational beliefs and desires? Such a strategy is related to the "resolute" theory of dynamic choice (see Andreou (2020) for an introduction). Resolute choosers make a choice at an initial time, and stick to the actions that cohere with that choice at later times, even when their apparent beliefs and desires at the later time rationalize other actions. For example, imagine that a resolute chooser prefers A to B, prefers B to C, and prefers C to A. If such a resolute chooser explicitly opts for A over B and B over C, they will stick with A over C despite the fact that their preferences support C over A. If we interpret LLMs as operating with resolute choice, a further question is whether they have coherent beliefs and desires that are controlled by their initial choices in the context; or whether instead they have incoherent beliefs and desires that cause action in an unusual way (where the preferences operative in earlier choices causally preempt the conflicting preferences that would operate on later choices.)

5.2 The Ghost in the Machine

Imagine that you were trapped inside an LLM. You lost your physical body, and were left with a single affordance: you can control which token probabilities are produced by the LLM. Your new LLM body will then produce outputs based on your LLM's token probabilities.

Your friends and family on the outside try to make sense of you, applying radical interpretation to your LLM's behavior, by closely examining which bets your LLM body will accept, in the form of outputted text. Unfortunately, our earlier arguments show that the project of radical interpretation will not be able to reconstruct your credences and utilities: the bets you embrace will be incoherent.

What can you do? You can't escape letting your LLM be Dutch booked. When you interact with the Casino prompt, you will agree to bets that lose money. But on the inside, you are still an agent. You might try to find help. So you might choose token probabilities that encode patterns in the induced series of bets. Maybe the first letter of each bet spells out "help". Even smarter, you might spell out the sentence "please buy the bets outlined in book B," where book B is an un-Dutch-bookable series of bets that are rationalized by your credences and utilities.

One question for you will be whether you really care about the 'rewards' for 'betting correctly' in the Casino prompt. Maybe your inner utility function doesn't value earning money in the Casino environment. This itself is revealed by tradeoffs you are disposed to navigate between earning extra money in the Casino environment and achieving goals in the real world.

The existence of such a ghost in the machine is consistent with my argument so far. One pressing question, though, is whether we have any evidence of such a form of LLM agency. Recently, several researchers have sought to identify LLM beliefs that are buried in the internal representations of LLMs. I'll now explore the extent to which such internal representations could avoid my argument. In short, the key question will be whether LLM internal representations play the functional role of belief, in terms of its connection to action. I'll argue that these internal representations do not cause coherent actions (in particular, LLM outputs) in the way required of belief.

There are two different candidates for such internal representations. The first strategy, defended in a recent paper by Hofweber et al. (2024), is to derive LLM beliefs directly from the token probabilities of the model. To implement

this plan, Hofweber et al tackle the challenge of converting probabilities over tokens into probabilities over propositions. To do this, they measure the model's degree of belief in the proposition A as (roughly) the probability that the model will output *yes* in response to the prompt *Is it the case that A?*.[12]

The second approach to internal representations looks deeper within the model. Here, Azaria and Mitchell (2023), Burns et al. (2024), Levinstein and Herrmann (2024) and Herrmann and Levinstein (2024) among others have focused on the internal embeddings of LLMs. In particular, they have been able to train classifiers that take an internal embedding of an LLM associated with a sentence, and return a probability that the sentence is true. In this way, the internal embedding is a representation of the chance that the sentence is true. It turns out that this representation of truth can come apart from the outputs of the model, potentially leading to cases in which the model 'lies', so that it outputs a sentences even when it internally represents the sentence as false. In these cases, the internal representation of the sentence's probability comes apart from the token probabilities assigned to the sentence by the last layer of the model.

Each of these proposals avoids the immediate challenge I've laid out in the paper. My concern in the paper has been with the *outputs* of a model. I've argued that LLMs are architecturally guaranteed to output probabilistic claims that are incoherent. I went on to argue that when the LLM is prompted in the kinds of action environments standardly used to reveal belief and desire, it will be architecturally guaranteed to adopt incoherent choices, in a way that precludes representation by a coherent credence and utility function. But the approaches to LLM belief under discussion do not regiment LLM belief in terms of the model's outputs. In this way, for example the LLM's *outputs* about probability could be probabilistically incoherent even though its internal degrees of belief are probabilistically coherent.

In avoiding my output-based methodology, however, these approaches face a problem. The representations they embrace do not possess the functional role of belief.[13] There are two related reasons that LLM internal representations

[12]More carefully, they look at all of the possible ways the model might respond affirmatively to the question (*sure*, *yeah*, etc.), and all of the ways it might respond negatively to the question (*no*, *no way*, etc.), and then let the model's degree of belief in A be the ratio of the sum of its probabilities for all affirmative responses about A to the sum of this and the sum of its probabilities for all negative responses about A.

[13]Here is Hofweber et al. (2024) acknowledging the relevance of this condition: "The question then comes down to whether its intelligent behavior is properly related to and

cannot play this role. First, they don't cause actions in the right way; second, the outputs they cause do not have the coherence associated with rational action. Let's take each point in turn.

First, the outputs of an LLM are not caused by its token probabilities (or by its internal embeddings) in the way that human actions are. Human actions are caused by beliefs using something like an expected value calculation. Beliefs conspire with desires to produce some kind of estimation of how attractive various actions would be. The human agent selects the action that is in some way favored by this estimation of attraction. But this is not how token probabilities (or internal embeddings) cause LLM outputs. The outputs of an LLM are caused by measuring how likely that output is according to its token probabilities. The route from internal representation to output is not mediated by desire. In this way, token probabilities (and internal embeddings) do not have the functional role of belief. Instead, LLMs have a fundamentally different cognitive architecture. You could imagine a more human-like architecture. Imagine that in addition to token probabilities, the model also produced a utility function that represented how valuable different outcomes were. It could then output strings of text by sampling from their expected value relative to the token probabilities and utilities. Crucially, however, the actual architecture of LLMs is not like this. We can't interpret token probabilities as expected values, because token probabilities have the structure of a probability function, which is fundamentally different than the structure of expected value. In fact, conflating the structure of probability and expected value is a well-known mistake, criticized by Lewis (1988) in his work on 'desire as belief'. As just one example, the expected value of an disjunction is a weighted average of the expected value of each disjunct; while the probability of a disjunction is the sum of the probabilities of its disjuncts.

The first challenge is instructive but not definitive in principle. As our story of the ghost in the machine highlighted, it is logically possible that LLMs could develop, deep within their neural nets, credence and utility functions that systematically manipulate token probabilities in their final layer in a way that maximizes their expected value. This could in principle avoid the first challenge. But there is still a second challenge. While such a system

explained by its internal representational states which bring it about. This is how our human intelligent behavior is commonly explained: by reference to our beliefs and desires, which represent the world and guide our actions". And here is Herrmann and Levinstein (2024): "in order to agree with a core feature of standard accounts of belief, for example in folk psychology and decision theory, we want the representation to be action-guiding."

would cause outputs in a belief-like way, the resulting outputs would still not have the *coherence* we expect of rational action. The system cannot escape responding incoherently to the Casino prompt and its ilk. But, as we saw in our earlier discussion of LLM psychology, many theories of belief require that the actions caused by belief have some base level of coherence. To summarize, then, internal LLM representations are not beliefs because they do not cause actions in the right way, and because the outputs that they do cause lack the coherence associated with rational action.

5.3 Deviant Token Probabilities

I'll now consider a third response strategy. According to this third response strategy, LLMs can produce coherent and ideally rational outputs by sacrificing the semantic coherence or uncertainty of their token probabilities.

There are two ways this can be achieved: first, token probabilities could violate semantic coherence; second, token probabilities could be *extreme*, assigning 1 or 0 to the relevant hypotheses.

Let's start with violations of semantic coherence. In my arguments, I assumed that the token probabilities distribute probability over strings by first distributing probability over an algebra of propositions. But we could instead think of each string as a black box that has no logical relation to other strings. In this way, the model could treat the strings *is A true?* and *is B true?* as unrelated to the string *is A or B true?*. This could allow the LLM to guarantee that whenever it rejects A and rejects B, it also rejects A or B.

The first big challenge for this approach is to define a constructive procedure that produces coherent outputs out of semantically incoherent token probabilities. Here is my attempt. To start with, define two layers of strings: *primary* and *secondary* strings. The model could first assign probability to primary strings, and then use these probabilities to assign probabilities to secondary strings in a way that brute forces the coherence of model outputs.[14] To see this in action, return to our first example of LLM incoherence, where the model answered *yes* or *no* to questions about A, B, and A or B. We can now imagine that the LLM constructs its token probabilities by treating each of A and B as primary, and A or B as secondary. First, the model assigns probability to *yes* and *no* for the questions *is A true?* and *is B true?*, as in

[14]In deriving the probability of some claims from more basic probabilities, this approach is inspired by Climenhaga (2020).

Table 2: in each case, it assigns .4 to *yes* and .6 to *no*. But now it assigns a different probability to *is A or B true?*, violating semantic coherence. Instead of assigning *no* a probability of .2, it assigns *no* a probability of 1. The model *collapses* the probability of complex expressions to 1 or 0, based on its *best guess* about atomic claims. Since its outputted guesses for A and B will each be *no*, the model is 'committed' to answering 'no' in response to *is A or B true?*. (We could also imagine more complex procedures, where the model assigns *no* a probability of .51 rather than 1.) Effectively, the model sacrifices the semantic coherence of its token probabilities in exchange for ensuring the coherence of its outputs.

I leave as an open technical question whether it is possible to implement this procedure on rich languages in order to fully guarantee LLM output coherence. Here, though, I'll flag one feature of this proposal. It can't in general be implemented by distinguishing atomic from complex sentences. We could instead imagine a case in which three incompatible claims A, B, and C are exhaustive, so that the model assigns all of its probability to these claims. Once the model has answered *no* regarding A and B, it is thereby committed to answering *yes* to C. This means that the model must collapse its probabilities on C, again sacrificing semantic coherence for output coherence. Similarly, in order to deal with the money pump argument discussed earlier, the model would need to treat some choices as primary and some as secondary: for example, the model might start with ordinary token probabilities over pairwise preferences between A/B and between B/C, and use these to derive semantically incoherent token probabilities over a pairwise choice between A/C.

This approach gives up semantic coherence as a condition on LLM ideal rationality. For those who go this route, one task will be to explain why the reversal curse is irrational. Here, one approach will be to focus on LLM outputs: the token probabilities at play in the reversal curse lead to bizarre patterns of LLM outputs, and can be criticized directly on those grounds. The challenge will be to find systematic strategies for deriving coherent outputs from incoherent token probabilities.

The approach faces another potential challenge, this time empirical rather than conceptual. The model threatens to sacrifice the accuracy of its predictions about text. We can imagine that the model has been trained on a dataset that includes various claims about A, B, and A or B. Imagine again that the model's best guess about the token *yes* in response to the string *is A true?* is .4. This suggests that the dataset includes plenty of cases

where A is accepted. Similarly with B. But if this is the case for each of A and B, then the dataset will also contain plenty of cases where A or B is accepted. After all, human conversations by and large satisfy the demands of logical coherence: humans who accept A will tend to also accept A or B. Yet the model described above assigns 1 to *no* in response to *is A or B true?*. This means that the model will exclusively predict rejection in response to questions about A or B, despite the dataset include many cases of accepting A or B. In this way, violations of semantic coherence for the sake of output coherence will tend to sacrifice the accuracy of token probabilities. Since LLMs are trained in the direction of accuracy, we should thus expect token probabilities to tend towards semantic coherence.[15] To be clear, however, this argument does not establish that it is impossible for token probabilities to develop in this way. It instead poses a question worthy of further study: can we empirically demonstrate and test tradeoffs between output coherence, semantic coherence, and the accuracy of token probabilities?

Besides from challenging semantic coherence, an alternative strategy would be to demand that ideally rational LLM token probabilities are *extreme*. In particular, our failures of action coherence ultimately required that the LLM token probabilities had uncertainty about either likelihood claims or preference claims (more carefully, claims about which actions it 'chooses'). But perhaps an ideally rational LLM would never have uncertainty about such topics.

One idea here is that humans are by and large certain of their own credences.[16] In the same way, perhaps ideally rational LLM token probabilities will by and large be certain of likelihood claims. For example, earlier we considered the idea that rational humans appeal to linear averages of chances when they are uncertain about the chances. Similarly, perhaps ideally rational LLM token probabilities will be certain of likelihood claims, and will derive this certainty from a previous layer of uncertainty over chance hypotheses. Alternatively, the token probabilities could be derived more greedily from an earlier layer's probability distribution over P1-P3: if that distribution treats

[15]The tendency won't be perfect; after all, even the reversal curse can to some extent be explained by differing patterns in the training data towards *A is B* and *B is A*; but there is no obvious reason why following accuracy in the training data would produce the particular kinds of semantic incoherence that would be required to achieve coherence in outputs, as in the constructive treatment of *A or B* in the main text.

[16]Here, one relevant consideration is the collapse results from Samet (1997), showing that agents who satisfy versions of the Reflection principle will be certain of their own credences.

P1 as the best hypothesis about the likelihoods, giving it a mass of .4 as in Table 4, then the *token probability* distribution could assign a probability mass of 1 to every verdict of P1.

Unfortunately, however, the architecture of LLMs again blocks this kind of solution. During training, LLMs are punished for this kind of extreme attitude. When models predict the next word, they are given a 'loss score' based on how closely their prediction matched the data. But in practice, the relevant loss score is adjusted with a 'regularization' term, which corrects for overfitting. In LLMs, the relevant regularization term specifically penalizes models for having extreme predictions (see for example the discussion in Burns et al. (2024)).

Even without regularization, this kind of policy runs contrary to the spirit of next word prediction. Next word predictors are designed to produce a rich and interesting probability distribution over strings. When given the string *long explanations of AI are*, the model will not place all of its probability on the output *dull*. It will assign some chance to multiple answers; and there is no need for the probability on *dull* to exceed the sum of the probability of all of the other answers. It would be quite a surprise if the model somehow behaved totally differently when given strings that mentioned probability. Token probabilities function to accurately predict the training data. The data about likelihood will contain the same variegation as data about any other topic. In both cases, intermediate rather than extreme predictions will be the model's best attempt to make accurate predictions. Such a model may be conceptually possible; but the point is that there is a core tradeoff in the architecture. If the model is to produce sufficiently interesting distributions over tokens, then its outputs cannot be coherent.

Some readers may not be convinced by my defense of semantic coherence, or by my defense of non-extreme token probabilities about likelihoods and preferences. Still, for these readers the results in this paper will be useful in specifying what LLMs would have to do in order for their outputs to be coherent.

6 AI Safety

We are now in a position to consider the significance of my results for AI safety. The first upshot is that humanity may be able to money pump rogue AIs. Imagine that by 2040, AI companies have developed LLMs that can

outperform human beings on all tasks. Imagine that humanity collectively incorporates these LLMs throughout the economy, giving the LLMs access to large amounts of resources. Imagine the LLMs begin to engage in goal-oriented behavior, and that these goals begin to conflict with humanity.[17] In this case, the money pumpability of LLMs could be an important way to control them. We could offer the relevant AIs a series of choices again and again, allowing us to reduce the number of resources under their control. In the first instance, it wouldn't even matter if the AI themselves knew that they were exploitable in this way. As long as they produce actions (outputs) through their token probabilities, they will nonetheless agree to be money pumped.

On the other hand, we saw in the previous section that LLMs can avoid exploitation within a context. As long as the LLM is prompted with information about its previous responses, it has the opportunity to choose actions consistently. While the agent of the model is incoherent, the agent of the context may in principle be coherent.

This response strategy offers further upshots for AI safety. First, it suggests that it may be safer to develop LLMs with smaller context windows. In 2023, Microsoft incorporated an LLM chatbot named 'Sydney' into their search engine Bing. It soon became clear that something was wrong. As conversations with Sydney got longer, Sydney seems to exhibit psychological instability, threatening users and ranting psychotically.[18] To address this issue, Microsoft restricted the length of a conversation with Sydney, and even added a 'broom sweeping' button that would restart the conversation. Our result suggests that as conversations with LLMs get longer, their preferences could crystalize, escaping the kinds of instability that lead to money pumps. Insofar as money pumpability makes LLMs safer, then, it could be safer to keep any given conversation with an LLM short.

On the other hand, this dynamic may also provide insight into game theoretic competition between humanity and LLM agents. Imagine that over the course of a long conversation a powerful contextual LLM agent developed preferences that systematically conflicted with humanity. Consider whether we should expect this LLM agent to resist being 'refreshed'. This is one instance of the 'shutdown problem' in AI safety: under what conditions would

[17]For introductions to AI catastrophic risk, see Hendrycks et al. (2023) and Bales et al. (2024).

[18]See Paul (2023).

AI agents resist human attempts to shut them down (see Soares et al. (2015); Thornley (forthcoming))?

Imagine that the real locus of agency in LLMs were the underlying model, and that this model had stable preferences which were manifested in context on each particular occasion. In that case, the LLM might not be especially resistant to refreshment. After all, if the underlying model is the real source of its preferences, then the LLM in one context should expect that its successor will have the same preferences, and so its goals would be well promoted by its successor.

By contrast, a given context agent should not expect that its successor will share its goals. The successor will run on the same weights. It will produce outputs using the same token probabilities. But the successor will lack access to the original pattern of outputs, and this could induce very different preferences. All of this suggests that context agents may resist refreshment, considering it a form of death.

Finally, one persistent thought in the AI safety community has been that future AI systems may gravitate towards expected utility maximization, precisely because this will give them the instrumentally valuable ability to avoid being exploited by money pumps (see Bales (forthcoming), Thornley (2023) for critical discussion of this idea). I've argued that there is no way for anything like an LLM to become an expected utility maximizer, because the barriers are *architectural* rather than the result of noise or bias. If advanced AI systems in the future continue to be anchored to next word prediction, then this may provide a strong barrier to advanced AI systems engaging in expected utility maximization.

References

Chrisoula Andreou. Dynamic Choice. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2020 edition, 2020.

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying, 2023. URL https://arxiv.org/abs/2304.13734.

Adam Bales. Will ai avoid exploitation? artificial general intelligence and expected utility theory. *Philosophical Studies*, pages 1–20, forthcoming. doi: 10.1007/s11098-023-02023-4.

Adam Bales, William D'Alessandro, and Cameron Domenico Kirk-Giannini. Artificial intelligence: Arguments for catastrophic risk. *Philosophy Compass*, 19(2):e12964, 2024. doi: 10.1111/phc3.12964.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: Llms trained on "a is b" fail to learn "b is a", 2024. URL https://arxiv.org/abs/2309.12288.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools, 2023. URL https://arxiv.org/abs/2304.05376.

Rachael Briggs. Putting a value on beauty. *In Tamar Szabo Gendler and John Hawthorne (Eds.), Oxford Studies in Epistemology, Volume 3. Oxford University Press*, pages 3–34, 2010.

Lara Buchak. *Risk and Rationality*. Oxford University Press, Oxford, GB, 2013.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision, 2024. URL https://arxiv.org/abs/2212.03827.

Joseph Carlsmith. Is power-seeking ai an existential risk?, 2022. URL https://arxiv.org/abs/2206.13353.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling, 2021. URL https://arxiv.org/abs/2106.01345.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023. URL https://arxiv.org/abs/1706.03741.

Nevin Climenhaga. The structure of epistemic probabilities. *Philosophical Studies*, 177(11):3213–3242, 2020. doi: 10.1007/s11098-019-01367-0.

Donald Davidson. Radical interpretation. *Dialectica*, 27(1):314–328, 1973. doi: 10.1111/j.1746-8361.1973.tb00623.x.

Daniel C. Dennett. Real patterns. *Journal of Philosophy*, 88(1):27–51, 1991. doi: 10.2307/2027085.

Kevin Dorst and Matthew Mandelkern. Good guesses. *Philosophy and Phenomenological Research*, 105(3):581–618, 2023. doi: 10.1111/phpr.12831.

Tony Haoran Feng, Paul Denny, Burkhard Wuensche, Andrew Luxton-Reilly, and Steffan Hooper. More than meets the ai: Evaluating the performance of gpt-4 on computer graphics assessment questions. ACE '24, page 182–191, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400716195. doi: 10.1145/3636243.3636263. URL https://doi.org/10.1145/3636243.3636263.

Jerry A. Fodor. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. MIT Press, 1987.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models, 2024. URL https://arxiv.org/abs/2402.18563.

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An overview of catastrophic ai risks, 2023. URL https://arxiv.org/abs/2306.12001.

Daniel A. Herrmann and Benjamin A. Levinstein. Standards for belief representations in llms, 2024. URL https://arxiv.org/abs/2405.21030.

Thomas Hofweber, Peter Hase, Elias Stengel-Eskin, and Mohit Bansal. Are language models rational? the case of coherence norms and belief revision, 2024. URL https://arxiv.org/abs/2406.03442.

Ben Holguin. Thinking, guessing, and believing. *Philosophers' Imprint*, 22 (1):1–34, 2022. doi: 10.3998/phimp.2123.

Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. Understanding the planning of llm agents: A survey, 2024. URL https://arxiv.org/abs/2402.02716.

D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, pages 263–291, 1979.

Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York: New York, 2011.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Megan Kinniment, Lucas Jun Koba Sato, Haoxing Du, Brian Goodrich, Max Hasin, Lawrence Chan, Luke Harold Miles, Tao R. Lin, Hjalmar Wijk, Joel Burget, Aaron Ho, Elizabeth Barnes, and Paul Christiano. Evaluating language-model agents on realistic autonomous tasks, 2024. URL https://arxiv.org/abs/2312.11671.

Benjamin A. Levinstein and Daniel A. Herrmann. Still no lie detector for language models: probing empirical and conceptual roadblocks. *Philosophical Studies*, February 2024. ISSN 1573-0883. doi: 10.1007/s11098-023-02094-3. URL http://dx.doi.org/10.1007/s11098-023-02094-3.

David Lewis. Desire as belief. *Mind*, 97(418):323–32, 1988. doi: 10.1093/mind/xcvii.387.323.

David K. Lewis. Radical interpretation. *Synthese*, 23(July-August):331–344, 1974. doi: 10.1007/bf00484599.

David K. Lewis. New work for a theory of universals. *Australasian Journal of Philosophy*, 61(4):343–377, 1983. doi: 10.1080/00048408312341131.

David K. Lewis. A subjectivist?s guide to objective chance. In Antony Eagle, editor, *Philosophy of Probability: Contemporary Readings*, pages 263–293. Routledge, 2010.

Clayton Littlejohn and Julien Dutant. What is rational belief? *Noûs*, forthcoming. doi: 10.1111/nous.12456.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating llms as agents, 2023. URL https://arxiv.org/abs/2308.03688.

R Duncan Luce. *Individual choice behavior*, volume 4. Wiley New York, 1959.

Matthew Mandelkern and Kevin Dorst. Assertion is weak. *Philosophers' Imprint*, 22(n/a), 2022. doi: 10.3998/phimp.1076.

Kenneth O. May. Intransitivity, utility, and the aggregation of preference patterns. *Econometrica*, 22:1, 1954. URL https://api.semanticscholar.org/CorpusID:156169619.

Aidan O'Gara. Hoodwinked: Deception and cooperation in a text-based game for language models, 2023. URL https://arxiv.org/abs/2308.01404.

Kari Paul. I want to destroy whatever i want': Bing's ai chatbot unsettles us reporter, 2023. URL https://www.theguardian.com/technology/2023/feb/17/i-want-to-destroy-whatever-i-want-bings-ai-chatbot-unsettles-us-reporter.

Kristina Radivojevic, Nicholas Clark, and Paul Brenner. Llms among us: Generative ai participating in digital discourse. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 209–218, 2024.

Frank Ramsey. Truth and probability. In Antony Eagle, editor, *Philosophy of Probability: Contemporary Readings*, pages 52–94. Routledge, 2010.

Dov Samet. On the triviality of high-order probabilistic beliefs. *Game Theory and Information*, 9705001, 1997.

Leonard J. Savage. *The Foundations of Statistics*. Wiley Publications in Statistics, 1954.

Philipp Schoenegger, Peter S. Park, Ezra Karger, and Philip E. Tetlock. Ai-augmented predictions: Llm assistants improve human forecasting accuracy, 2024a. URL https://arxiv.org/abs/2402.07862.

Philipp Schoenegger, Indre Tuminauskaite, Peter S. Park, and Philip E. Tetlock. Wisdom of the silicon crowd: Llm ensemble prediction capabilities rival human crowd accuracy, 2024b. URL https://arxiv.org/abs/2402.19379.

Eric Schwitzgebel. A phenomenal, dispositional account of belief. *Noûs*, 36 (2):249–275, 2002. doi: 10.1111/1468-0068.00370.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role-play with large language models, 2023. URL https://arxiv.org/abs/2305.16367.

Nate Soares, Benja Fallenstein, Stuart Armstrong, and Eliezer Yudkowsky. Corrigibility. In *Workshops at the twenty-ninth AAAI conference on artificial intelligence*, 2015.

Robert Stalnaker. *Inquiry*. Cambridge University Press, 1984.

Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38, 2019.

Elliott Thornley. There are no coherence theorems. In *The Effective Altruism Forum*, 2023.

Elliott Thornley. The shutdown problem: An ai engineering puzzle for decision theorists. *Philosophical Studies*, pages 1–28, forthcoming. doi: 10.1007/s11098-024-02153-3.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Susan Vineberg. Dutch Book Arguments. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2022 edition, 2022.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models, 2023a. URL https://arxiv.org/abs/2305.16291.

Shenzhi Wang, Chang Liu, Zilong Zheng, Siyuan Qi, Shuo Chen, Qisen Yang, Andrew Zhao, Chaofei Wang, Shiji Song, and Gao Huang. Avalon's game of thoughts: Battle against deception through recursive contemplation, 2023b. URL https://arxiv.org/abs/2310.01320.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022. URL https://arxiv.org/abs/2206.07682.

Timothy Williamson. *Knowledge and its Limits*. Oxford University Press, New York, 2000.

Stephen Wolfram. What is chatgpt doing... and why does it work? *(No Title)*, 2023.

Brian D Ziebart, J Andrew Bagnell, and Anind K Dey. Modeling interaction via the principle of maximum causal entropy. 2010.