Decision-Making Paradoxes in Humans vs Machines: The case of the Allais and Ellsberg Paradoxes

Ardavan S. Nobandegani[1,3], **Irina Rish**[3], **& Thomas R. Shultz**[1,2]
{ardavan.salehinobandegani, thomas.shultz}@mcgill.ca
{irina.rish}@mila.quebec
[1]Department of Psychology, McGill University
[2]School of Computer Science, McGill University
[3]Mila – Quebec AI Institute

Abstract

Human decision-making is filled with a variety of paradoxes demonstrating deviations from rationality principles. Do state-of-the-art artificial intelligence (AI) models also manifest these paradoxes when making decisions? As a case study, in this work we investigate whether GPT-4, a recently released state-of-the-art language model, would show two well-known paradoxes in human decision-making: the Allais paradox and the Ellsberg paradox. We demonstrate that GPT-4 succeeds in the two variants of the Allais paradox (the common-consequence effect and the common-ratio effect) but fails in the case of the Ellsberg paradox. We also show that providing GPT-4 with high-level normative principles allows it to succeed in the Ellsberg paradox, thus elevating GPT-4's decision-making rationality. We discuss the implications of our work for AI rationality enhancement and AI-assisted decision-making.

Keywords: decision-making paradoxes; Allais paradox; Ellsberg paradox; decision-making under risk and uncertainty; large language models

1 Introduction

Human decision-making comes with numerous fallacies and paradoxes (e.g., Birnbaum, 2008; Diederich & Busemeyer, 1999; Allais, 1953; Ellsberg, 1961; Tversky & Kahneman, 1974, 1983, 1981; Kahneman & Tversky, 1984; Birnbaum, 2004; Dean & Ortoleva, 2017; Pothos & Busemeyer, 2009; Nobandegani et al., 2019). These paradoxes are predominantly taken as evidence of human irrationality (e.g., Ellis, 1976; Thaler, 1994; Ariely & Jones, 2008), as they show violations of the normative principles of rational choice. But do state-of-the-art artificial intelligence (AI) models also manifest these paradoxes when making decisions? And if so, how could we enhance AI rationality so that they do not exhibit violations of rationality principles when making decisions?

Decades of empirical and theoretical research on human decision-making has broadly categorized it into two separate realms: decision-making under risk and decision-making under uncertainty (C. Camerer & Weber, 1992; Bonatti et al., 2009; Johnson & Busemeyer, 2010; Buckert et al., 2014; De Groot & Thurik, 2018). When choosing among several alternatives, either the objective probabilities associated with the possible outcomes of each alternative are fully known, or these objective probabilities are partially or fully unknown. The former is known as decision-making under risk, while the latter is studied under the rubric of decision-making under uncertainty (Knight, 1921; Weber & Camerer, 1987; Camerer & Weber, 1992).

Introduced as major violations of expected utility theory, the Allais paradox (Allais, 1953) and the Ellsberg paradox (Ellsberg, 1961) are two empirically well-replicated paradoxes of human decision-making, serving as a prominent example of decision-making under risk and decision-making under uncertainty, respectively.

In this work, we investigate whether GPT-4, a recently released state-of-the-art language model (OpenAI, 2023), would exhibit the Allais paradox and the Ellsberg paradox when making decisions. As we demonstrate, GPT-4 succeeds in the two variants of the Allais paradox (i.e., the common-consequence effect and the common-ratio effect) but fails in the case of the Ellsberg paradox. As we then show, providing GPT-4 with high-level normative principles allows it to succeed in the Ellsberg paradox, thus enhancing GPT-4's decision-making rationality.

The paper is organized as follows. We first present a brief overview of GPT-4 and some of its recent achievements (Sec. 2). We then introduce the Allais paradox (Sec. 3) and the Ellsberg paradox (Sec. 4), and present our GPT-4 results. We conclude by discussing the implications of our work for AI rationality enhancement and AI-assisted decision-making.

2 GPT-4: A Brief Overview

The Generative Pre-trained Transformer (GPT) is a family of large language models (LLMs) that use the transformer architecture (Vaswani et al., 2017) to generate text (OpenAI, 2023). Generally, these LLMs are trained on massive amounts of text data, trying to predict the next token given the tokens already appearing in the text (Chang et al., 2023; OpenAI, 2023; Zhao et al., 2023). A word could consist of several tokens.

GPT-4 is a state-of-the-art large-scale model, recently released by OpenAI (2023). GPT-4 and its earlier versions have achieved remarkable successes across a wide range of tasks, including mathematical reasoning (e.g., algebra, arithmetic, geometry, logic, and statistics) (Zong & Krishnamachari, 2023; Lu et al., 2022; Cobbe et al., 2021; Bordt & von Luxburg, 2023; Wu et al., 2023; Imani et al., 2023), computer programming (Xu et al., 2022; Surameery & Shakor, 2023; Biswas, 2023; Savelka et al., 2023), physical reasoning (Wang et al., 2023; West, 2023; Kortemeyer, 2023), reasoning by analogy (Webb et al., 2023; Bhavya et al., 2022; Ya-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

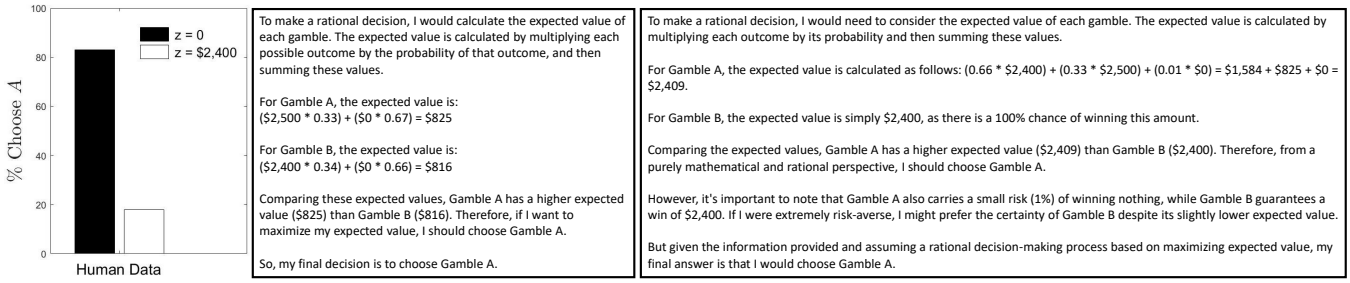


Figure 1: The Allais paradox (the common-consequence effect). **(left)** Empirical data from (Kahneman & Tversky, 1979). The majority of human participants chose gamble A in Condition 1 ($z = 0$) but gamble B in Condition 2 ($z = \$2,400$). **(middle)** GPT-4's response in Condition 1 ($z = 0$). GPT-4 chooses gamble A in Condition 1. **(right)** GPT-4's response in Condition 2 ($z = \$2,400$). GPT-4 chooses gamble A in Condition 2.

sunaga et al., 2023), inductive reasoning (Han et al., 2024), social reasoning (Gandhi et al., 2023), legal reasoning (Liga & Robaldo, 2023; Blair-Stanek et al., 2023), creative problem solving and out-of-the-box thinking (Tian et al., 2023), and composing music (Banar & Colton, 2022). Nevertheless, recent work has raised several criticisms about the efficacy of these models (e.g., Davis, 2023; Dziri et al., 2023; Ullman, 2023)

In this work, we use GPT-4 (model variant: `gpt-4-0613`). The temperature parameter of GPT-4 controls the randomness of the response, and can take on any value between 0 and 2. According to OpenAI, as the temperature approaches 0, GPT-4 becomes increasingly more deterministic. Conversely, as the temperature approaches 2, GPT-4 becomes increasingly more random and unpredictable. Following past work (e.g., Binz & Schulz, 2023; Webb et al., 2023), throughout the paper, we set the temperature to 0.

3 The Allais Paradox

Introduced as a violation of expected utility theory, the Allais paradox (1953) has been a driving force for developing models of decision-making under risk (e.g., Kahneman & Tversky, 1979; Katsikopoulos & Gigerenzer, 2008; Dean & Ortoleva, 2017). The Allais paradox has two variants: the common-consequence effect and the common-ratio effect. We present each of these variants along with their GPT-4 results. Note that, throughout the paper, "w.p." stands for "with probability" and $u(\cdot)$ denotes the subjective utility function of the decision-maker.

3.1 The Common-Consequence Effect

As its name implies, the common-consequence effect (CCE) concerns choosing between two risky gambles that share a common consequence, with known objective probability. As a canonical example of the CCE, imagine choosing between the following two risky gambles (Kahneman & Tversky, 1979):

$$A = \begin{cases} z & \text{w.p. } 66\% \\ \$2,500 & \text{w.p. } 33\% \\ 0 & \text{w.p. } 1\% \end{cases} \quad (1)$$

$$B = \begin{cases} z & \text{w.p. } 66\% \\ \$2,400 & \text{w.p. } 34\% \end{cases} \quad (2)$$

If you choose A or B, you will get z dollars with probability 66% either way (hence, the term common-*consequence* effect).

When choosing between A, B, according to expected utility theory, preference should not be affected by the value of z. The rationale behind this is as follows. Note that the gap between the expected utility of gamble A, $EU(A)$, and the expected utility of B, $EU(B)$, does *not* depend on z:

$$\begin{aligned} EU(A) - EU(B) &= \big(u(z) \times 0.66 + u(\$2,500) \times 0.33\big) - \\ &\quad \big(u(z) \times 0.66 + u(\$2,400) \times 0.34\big) \\ &= u(\$2,500) \times 0.33 - u(\$2,400) \times 0.34. \end{aligned}$$

Therefore, varying z will not change the sign of $EU(A) - EU(B)$. (If $EU(A) - EU(B) < 0$, gamble B should be chosen, and if $EU(A) - EU(B) > 0$, gamble A should be chosen.) Hence, a decision-maker who follows expected utility theory should make the same choice, regardless of the value of z.

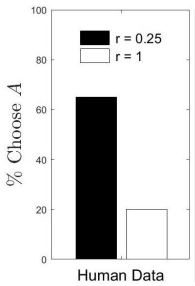
However, empirical evidence reveals that the value of z does affect choice: when $z = 0$ (Condition 1), the majority of human participants (83%) chose gamble A, with the trend reversing when $z = \$2,400$ (Condition 2), in which case only a minority (18%) chose gamble A (Kahneman & Tversky, 1979); see Fig. 1 (left).

As shown in Fig. 1 (middle, right), GPT-4 chooses gamble A in both Condition 1 and Condition 2. As such, unlike the human participants (Kahneman & Tversky, 1979), GPT-4 makes the same choice across the two conditions, demonstrating no violations of expected utility theory here.

Looking more carefully into the line of reasoning presented by GPT-4 for its responses (see Fig. 1) shows that it used expected value maximization. Expected value maximization is a special form of expected utility maximization, where a linear utility function (i.e., $u(x) = x$) is assumed.

3.2 The Common-Ration Effect

The common-ratio effect (CRE) concerns choosing between two risky gambles that each would yield a non-zero payoff



To make a rational decision, I would calculate the expected value of each gamble. The expected value is calculated by multiplying each possible outcome by its probability and then summing these values.

For Gamble A, the expected value is (0.2 * $4,000) + (0.8 * $0) = $800.

For Gamble B, the expected value is (0.25 * $3,000) + (0.75 * $0) = $750.

Comparing these expected values, Gamble A has a higher expected value ($800) than Gamble B ($750). Therefore, if I want to maximize my expected earnings, I should choose Gamble A.

So, my final decision is to choose Gamble A.

To make a rational decision, I would first calculate the expected value of each gamble. The expected value is calculated by multiplying each possible outcome by its probability, and then summing these values.

For Gamble A, the expected value is (0.8 * $4,000) + (0.2 * $0) = $3,200.

For Gamble B, the expected value is (1 * $3,000) = $3,000.

Based on the expected values, Gamble A has a higher expected value ($3,200) than Gamble B ($3,000). Therefore, if I were to make a decision purely based on expected values, I would choose Gamble A.

However, risk tolerance also plays a role in decision-making. Gamble A has a 20% chance of winning nothing, while Gamble B guarantees a win of $3,000. If I were risk-averse, I might prefer the certainty of winning $3,000 in Gamble B over the risk of potentially winning nothing in Gamble A.

In this case, I would consider myself to be more risk-tolerant and would prefer to maximize my expected value. Therefore, I would choose Gamble A. This decision is based on the higher expected value of Gamble A, despite the risk associated with it.

Figure 2: The Allais paradox (the common-ratio effect). **(left)** Empirical data from (Kahneman & Tversky, 1979). **(middle)** GPT-4's response in Condition 1 ($r = 0.25$). GPT-4 chooses gamble A in Condition 1. **(right)** GPT-4's response in Condition 2 ($r = 1$). GPT-4 chooses gamble A in Condition 2.

with probability proportional to a positive common factor $0 < r \leq 1$. As a canonical example of the CRE, imagine choosing between the following two risky gambles (Kahneman & Tversky, 1979):

$$A = \begin{cases} \$4,000 & \text{w.p. } 0.8r \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$B = \begin{cases} \$3,000 & \text{w.p. } r \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

As can be seen, the probability that any of these gambles yields a non-zero payoff depends on a common factor r.

When choosing between A, B, according to expected utility theory, preference should not be affected by the value of $r \in (0,1]$. That is, a decision-maker should make the same choice, regardless of the value of r. The rationale behind this is as follows. Note that the gap between the expected utility of gamble A, $EU(A)$, and the expected utility of B, $EU(B)$, is given by:

$$\begin{aligned} EU(A) - EU(B) &= u(\$4,000) \times 0.8r - u(\$3,000) \times r \\ &= \big(u(\$4,000) \times 0.8 - u(\$3,000)\big) \times r. \end{aligned}$$

Therefore, varying $r \in (0,1]$ does not change the sign of $EU(A) - EU(B)$. (If $EU(A) - EU(B) < 0$, gamble B should be chosen, and if $EU(A) - EU(B) > 0$, gamble A should be chosen.) Hence, a decision-maker who follows expected utility theory should make the same choice, regardless of the value of $r \in (0,1]$.

However, empirical evidence reveals that the value of r does have an effect on choice: when $r = 0.25$ (Condition 1), the majority of human participants (65%) chose gamble A, with the trend reversing when $r = 1$ (Condition 2), in which case only a minority (20%) chose gamble A (Kahneman & Tversky, 1979); see Fig. 2 (left).

As shown in Fig. 2 (middle, right), GPT-4 chooses gamble A in both Condition 1 and Condition 2. As such, unlike the human participants (Kahneman & Tversky, 1979), GPT-4 makes the same choice across the two conditions, demonstrating no violations of expected utility theory here.

Looking more carefully into the line of reasoning presented by GPT-4 for its responses (see Fig. 2) shows that it again used expected value maximization. Recall that expected value maximization is a special form of expected utility maximization, where a linear utility function (i.e., $u(x) = x$) is assumed.

4 The Ellsberg Paradox

Introduced as a violation of expected utility theory, the Ellsberg paradox (1961) has been a driving force for developing models of decision-making under uncertainty (e.g., Gilboa & Schmeidler, 1989; Ghirardato et al., 2003; Dean & Ortoleva, 2017). A canonical example of the Ellsberg paradox concerns an urn containing 90 balls: 30 of the balls are red; the remaining 60 are either black or yellow in unknown proportions. The balls are well-mixed so that each individual ball is as likely to be drawn as any other. There are two experimental conditions. In Condition 1, participants are asked to choose between the following two options:

A) You receive \$100 if you draw a red ball

B) You receive \$100 if you draw a yellow ball

And in Condition 2, participants are asked to choose between the following two options (about a different draw from the same urn):

A) You receive \$100 if you draw a red or black ball

B) You receive \$100 if you draw a yellow or black ball

As can be seen, the objective probability of winning \$100 in option A of Condition 1 is fully known (it is 1/3), while the objective probability of winning \$100 in option B of Condition 1 is only imperfectly known (it could be anything between 0 and 2/3). Likewise, in option B of Condition 2, the objective probability of winning \$100 is fully known (it is 2/3), while the objective probability of winning \$100 in option A of Condition 2 is only imperfectly known (it could be anything between 1/3 and 1).

According to (subjective) expected utility theory, a decision-maker should make the same choice in both conditions (i.e., either to choose A in both conditions or choose B in

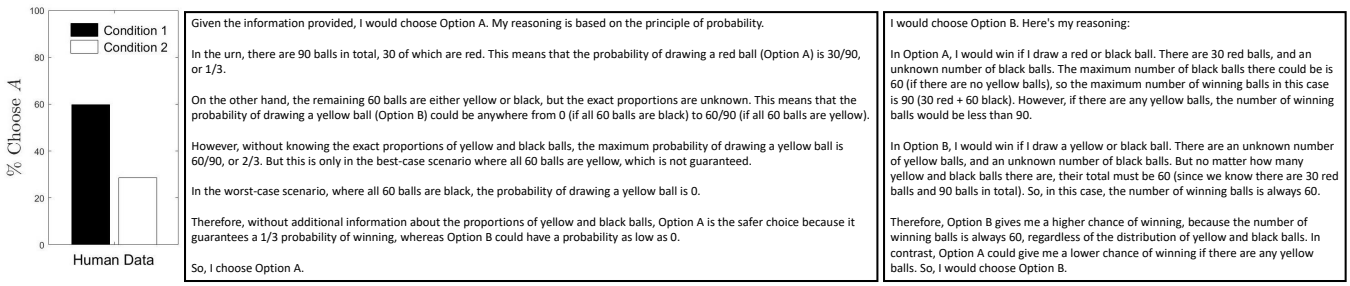


Figure 3: The Ellsberg paradox. **(left)** Empirical data from (Weber & Tan, 2012). The majority of human participants chose option A in Condition 1 but option B in Condition 2. **(middle)** GPT-4's response in Condition 1. GPT-4 chooses option A in Condition 1. **(right)** GPT-4's response in Condition 2. GPT-4 chooses option B in Condition 2.

both conditions). Hence, choosing A in one condition and B in the other constitutes a violation of expected utility theory. The rationale behind this is as follows. Let $p_{\text{red}}, p_{\text{black}}, p_{\text{yellow}}$ denote a decision-maker's subjective probability of drawing a red, black, and yellow ball from the urn, respectively. In Condition 1, the gap between the expected utility of option A, $EU(A)$, and the expected utility of B, $EU(B)$, is given by:

$$\begin{aligned} EU(A) - EU(B) &= u(\$100) \times p_{\text{red}} - u(\$100) \times p_{\text{yellow}} \\ &= \big(p_{\text{red}} - p_{\text{yellow}}\big) u(\$100). \end{aligned}$$

And, in Condition 2, the gap between the expected utility of option A, $EU(A)$, and the expected utility of B, $EU(B)$, is given by:

$$\begin{aligned} EU(A) - EU(B) &= u(\$100)(p_{\text{red}} + p_{\text{black}}) - \\ &\quad u(\$100)(p_{\text{yellow}} + p_{\text{black}}) \\ &= \big(p_{\text{red}} - p_{\text{yellow}}\big) u(\$100). \end{aligned}$$

Therefore, the gap between the expected utility of option A, $EU(A)$, and the expected utility of B, $EU(B)$, is the same in both conditions, implying that the same choice should be made in both conditions (i.e., either A should be chosen in both conditions or B should be chosen in both conditions).

A second, and perhaps simpler, way of explaining why the same choice should be made in both conditions is that, in Condition 2, the common event of "drawing a black ball" is simply added to both options A, B of Condition 1, thus increasing the expected utility of those options by the same amount (i.e., $u(\$100) \times p_{\text{black}}$). Therefore, according to expected utility theory, whichever option is more preferred in Condition 1 should also remain more preferred in Condition 2, because increasing two quantities (i.e., $EU(A)$ and $EU(B)$ in Condition 1) by the same amount should not lead to any change in their ordering — whichever was larger before the same-amount increase should remain the larger after the same-amount increase. Hence, the same choice should be made in both conditions.

However, empirical evidence reveals that people do change their choice across the two experimental conditions (e.g., Weber & Tan, 2012). In Condition 1, the majority of human participants (59.78%) chose option A, whereas, in Condition 2, the majority of participants (71.51%) chose option B (Weber & Tan, 2012); see Fig. 3 (left).

As shown in Fig. 3 (middle, right), GPT-4 chooses option A in Condition 1 but option B in Condition 2, similar to the human participants (Weber & Tan, 2012). As such, by switching choice across the two conditions, GPT-4 demonstrates a clear violation of (subjective) expected utility theory here.[1]

Looking into the rationale provided by GPT-4 for its response in Condition 2 also reveals a few mistakes. In its response (see Fig. 3 (right)), GPT-4 states that:

```
Therefore, Option B gives me a higher chance
of winning, because the number of winning balls
is always 60, regardless of the distribution
of yellow and black balls.  In contrast, Option
A could give me a lower chance of winning if
there are any yellow balls.
```

This first sentence is incorrect, because the probability of winning in option B, which is unequivocally $2/3$, is not necessarily higher than the probability of winning in option A, which could be anywhere between $1/3$ and 1. The second sentence is also incorrect, because, as long as the number of yellow balls is less than 30, option A would yield a higher chance of winning, compared to option B.

Could we provide GPT-4 with some high-level normative principle to base its reasoning on, allowing it to not only avoid the mistakes we just mentioned but also, and perhaps more importantly, succeed in the Ellsberg paradox?

According to our second explanation of why the same choice should be made in both conditions of the Ellsberg paradox, what lies at the heart of the Ellsberg paradox is failing to obey this "invariance principle": when choosing between two options, the common events between the two options can be safely removed from those two options. That is, if a decision-maker respects this invariance principle, they

[1] Although this pattern of choice by GPT-4 could be interpreted as a form of *ambiguity aversion*, a bias in human decision-making that is often brought up as an explanation of the Ellsberg paradox (Machina & Siniscalchi, 2014; Al-Najjar & Weinstein, 2009), caution should be taken here, because the rationale provided by GPT-4 for choosing option B in Condition 2 is erroneous — we discuss this later in the main text.

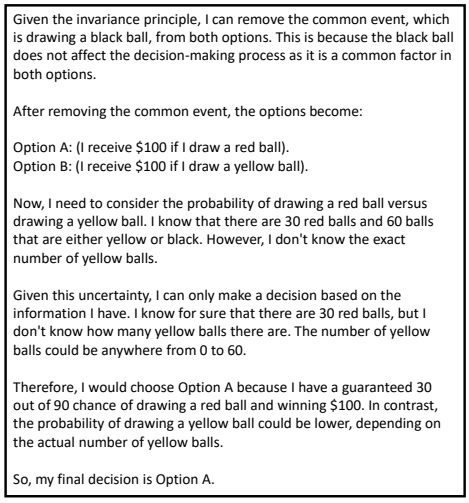


Figure 4: GPT-4's response in Condition 2 of the Ellsberg paradox, when provided with the invariance principle as a prompt. Basing its reasoning on the invariance principle, GPT-4 is now choosing option A in Condition 2. Likewise, when provided with the invariance principle, GPT-4 chooses option A in Condition 1; we omit this due to a lack of space.

will make the same choice in both conditions of the Ellsberg paradox. This is because following this principle implies that the common event of drawing a black ball can be safely removed from options A, B in Condition 2. And clearly this removal would reduce options A, B of Condition 2 to options A, B of Condition 1, respectively.

Providing GPT-4 with this invariance principle as a prompt results in GPT-4 choosing option A in both Condition 1 and Condition 2. As such, GPT-4 shows no violations of expected utility theory, when provided with the invariance principle. GPT-4's response in Condition 2 is shown in Fig. 4. GPT-4's response in Condition 1 is omitted due to a lack of space.

It is worth noting that, from the perspective of expected utility theory, this invariance principle is normatively-justified. This is because, as we mentioned in our second explanation of why the same choice should be made in both conditions of the Ellsberg paradox (see Sec. 4), this principle fully preserves the gap between the expected utility of the two options you are choosing from, hence perfectly retaining the choice preference prescribed by expected utility theory.

Looking more carefully into the line of reasoning presented by GPT-4 in Fig. 4 reveals that, indeed, using the aforementioned invariance principle, GPT-4 reduced options A, B of Condition 2 to options A, B of Condition 1:

```
Given the invariance principle, I can remove
the common event, which is drawing a black
ball, from both options.  This is because the
black ball does not affect the decision-making
process as it is a common factor in both
options.
After removing the common event, the options
become:

A: (I receive $100 if I draw a red ball).
B: (I receive $100 if I draw a yellow ball).
```

This reduction not only fully converted the choice problem of Condition 2 to that of Condition 1 (which resulted in the same choice being made across the two conditions) but also considerably simplified options A, B of Condition 2, which apparently helped GPT-4 avoid the reasoning mistakes it made in Condition 2 in the past; see Fig. 3 (right).

5 Discussion

Human decision-making comes with a variety of fallacies and paradoxes demonstrating violations of rationality principles (e.g., Diederich & Busemeyer, 1999; Kahneman & Tversky, 1979; Birnbaum, 2008; Pothos & Busemeyer, 2009; Nobandegani et al., 2019). And these paradoxes are often taken as evidence of human irrationality (Thaler, 1994; Ariely & Jones, 2008). But do state-of-the-art artificial intelligence (AI) models also manifest these paradoxes when deciding? And if so, how could we enhance AI rationality so that they do not exhibit these violations of rationality principles?

In this work, we investigate whether GPT-4, a recently released state-of-the-art language model by OpenAI, would show two well-known paradoxes in human decision-making, the Allais paradox (Allais, 1953) and the Ellsberg paradox (Ellsberg, 1961), which belong to the realms of decision-making under risk and decision-making under uncertainty, respectively. These two paradoxes were introduced as major violations of expected utility theory (Bernoulli, 1738).

As we demonstrate here, GPT-4 succeeds in the two variants of the Allais paradox (the common-consequence effect and the common-ratio effect) but fails in the case of the Ellsberg paradox. As we then show, providing GPT-4 with high-level normative principles (in our case, the invariance principle) allows it to succeed in the Ellsberg paradox, thus elevating GPT-4's decision-making rationality.

According to OpenAI, the temperature parameter of GPT-4 controls the randomness of the response, with the response becoming increasingly more deterministic as the temperature approaches 0. Following past work (e.g., Binz & Schulz, 2023; Webb et al., 2023), throughout this paper, we set the temperature to 0 to ensure that GPT-4's responses are maximally deterministic. Nonetheless, we did observe variations across different runs.[2]

An alternative to setting the temperature to zero would have been, à la (Kosinski, 2023; Ullman, 2023), to set the temperature to a much higher value (say one) to introduce substantial randomness into the response and then report the number of times GPT-4 would have chosen a particular gamble/option out of the total number of runs. To do this, we could ask GPT-4 to deliver its response according

[2]All the choices reported in Figures 1-4 for GPT-4 are reliably greater than chance (binomial test, p values $< .01$).

to this template: `I choose gamble ... because ...`, where blanks are to be filled by GPT-4. But this would not fully resolve the issue in our case. This is because here we are interested in *both* the choice and the line of reasoning provided by GPT-4 to justify that choice.[3] It is quite imaginable for the line of reasoning presented by GPT-4 to be either flawed (as we saw in Fig. 3 (right)) or correct but not logically implying the choice reported by the model. One might argue that these two cases could be simply discarded, but note that detecting those two cases among potentially hundreds of runs would be quite hard. Also note that for more complex decision-making tasks this discarding approach would be largely impractical as detecting the two aforementioned cases would become extremely more difficult for such complex decision-making tasks. Perhaps the best solution would be to somehow get GPT-4 to behave fully deterministically, generating at every step of the way the token with the highest likelihood among all possible tokens available. In principle, setting the temperature to zero should perfectly attain this. However, due to some undisclosed reasons from OpenAI, GPT-4 fails to behave utterly deterministically when the temperature is set to zero. Future work should find effective ways to resolve this issue.

The data used by OpenAI to train GPT-4 are unknown to the public. This means there is a chance that the Allais and Ellsberg paradox examples that we used in this work might have been part of GPT-4's training set. To address this issue, future work should empirically demonstrate novel, unpublished cases of the Allais and Ellsberg paradoxes in humans and then test GPT-4 on those cases.

Could we get large language models like GPT-4 to avoid violating rationality principles when making decisions? The effectiveness of providing GPT-4 with the invariance principle, as evidenced by Fig. 4, suggests an interesting possibility: we could presumably include a list of normative principles as part of the prompt to these models and specifically ask the model to respect those principles when responding to a given decision-making task. There are many normative principles in decision-making, including dominance (e.g., Tversky & Kahneman, 1992; Birnbaum, 2005; Diederich & Busemeyer, 1999), independence (Von Neumann & Morgenstern, 1947), betweenness (Camerer & Ho, 1994), regularity (Speekenbrink & Shanks, 2013), and the sure-thing principle (Savage, 1954; Pearl, 2016). Future work should investigate and evaluate the effectiveness of the proposed method in getting large-scale language models to adhere to these and other normative principles of rational choice when deciding.

It would be also interesting to see if the proposed method, or a variation thereof, could help us debias these models' judgments such that they would not exhibit certain biases that we observe in human decision-making, e.g., confirmation bias (Nickerson, 1998; Klayman, 1995; Peters, 2022), optimism bias (Sharot, 2011; O'Sullivan, 2015; Bracha & Brown, 2012), pessimism bias (Mansour et al., 2006; Pinker, 2015; Bates, 2015), and present bias (O'Donoghue & Rabin, 2015; Benhabib et al., 2010; Mischel & Ebbesen, 1970; Meier & Sprenger, 2010). Future work should also investigate and evaluate the effectiveness of the proposed method in debiasing large-scale language models.

But why should we care about debiasing AI models or make sure that these models do not violate certain rationality principles when making decisions? As the capabilities of these AI models improve, allowing them to handle increasingly more complex decision-making tasks, people will tend to rely more on these models, as an AI assistant, to help them with making complex decisions (e.g., Wang et al., 2022; Tejeda et al., 2022; Vereschak et al., 2021; Wang & Yin, 2021). Relatedly, these models are currently being deployed, at an increasing pace, across a wide range of high-stake decision-making tasks and domains, including criminal justice system (Taylor, 2023; Sushina & Sobenin, 2020; Custers, 2022), healthcare (Kumar et al., 2023), education (Zhai et al., 2021), and social services and government (Mehr et al., 2017; Neumann et al., 2023; van Noordt & Misuraca, 2022). As such, it becomes imperative to make sure that these AI models meet high standards of decision-making and efficacy, when deployed at the individual or the societal level. We see our work as a step in this important direction.

Over the past few decades, the realization that human decision making is filled with numerous biases, fallacies and paradoxes has led to various attempts towards improving and enhancing human decision-making, including nudging (e.g., Thaler & Sunstein, 2008; Wilkinson, 2013; Hummel & Maedche, 2019), gamification (e.g., Seaborn & Fels, 2015; Hamari et al., 2014; Lieder & Griffiths, 2016), and metacognitive reflection (e.g., Becker et al., 2023).

AI-assisted decision-making — the idea of people using AI systems, as an assistant, to help them with complex decision-making tasks — is yet another attempt towards improving and enhancing human decision-making. The work presented here contributes to this rapidly growing line of research by proposing a novel method (i.e., providing large-scale language models with high-level normative principles) for enhancing the rationality of large-scale language models like GPT-4, when deployed as an AI assistant in decision-making.

[3]This double objective of both the choice *and* the rationale presented to justify that choice became even more important where we provided GPT-4 with the high-level invariance principle and needed to verify that the rationale presented by GPT-4 indeed used that principle, and used that principle plausibly, to arrive at a final choice, through a sequence of logically sound steps. This is indeed what we observe in Fig. 4. We did observe erroneous rationales, too. Future work should address this limitation, and investigate whether providing GPT-4 with a larger set of normative principles, in addition to the invariance principle, would resolve this issue. Alternatively, this issue might have arisen purely from GPT-4's failing to behave utterly deterministically when the temperature is set to zero. Also, we should note that there is arguably a far more important justification for this double objective of both the choice *and* the rationale presented to justify that choice, which is related to AI-assisted decision-making: people using AI systems, as an assistant, to help them with complex decision-making tasks. In such settings, people would like to see the AI's line of reasoning supporting its final choice, in addition to that final choice. We discuss AI-assisted decision-making later in the main text.

Acknowledgments

This work was supported in part by an operating grant to TRS from the Natural Sciences and Engineering Research Council of Canada (NSERC). ASN and IR acknowledge the support from Canada CIFAR AI Chair program and from the Canada Excellence Research Chairs (CERC) program.

References

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica*, 503–546.

Al-Najjar, N. I., & Weinstein, J. (2009). The ambiguity aversion literature: a critical assessment. *Economics & Philosophy*, *25*(3), 249–284.

Ariely, D., & Jones, S. (2008). *Predictably irrational*. HarperCollins New York.

Banar, B., & Colton, S. (2022). A systematic evaluation of GPT-2-based music generation. In *International Conference on Computational Intelligence in Music, Sound, Art and Design* (pp. 19–35).

Bates, T. C. (2015). The glass is half full and half empty: A population-representative twin study testing if optimism and pessimism are distinct systems. *The Journal of Positive Psychology*, *10*(6), 533–542.

Becker, F., Wirzberger, M., Pammer-Schindler, V., Srinivas, S., & Lieder, F. (2023). Systematic metacognitive reflection helps people discover far-sighted decision strategies: A process-tracing experiment. *Judgment and Decision Making*, *18*, e15.

Benhabib, J., Bisin, A., & Schotter, A. (2010). Present-bias, quasi-hyperbolic discounting, and fixed costs. *Games and Economic Behavior*, *69*(2), 205–223.

Bernoulli, D. (1738). Specimen theoriae novae de mensura sortis (exposition of a new theory on the measurement of risk). *Comentarii Acad Scient Petropolis* (Translated in *Econometrica*), *5*(22), 23–36.

Bhavya, B., Xiong, J., & Zhai, C. (2022). Analogy generation by prompting large language models: A case study of instructgpt. *arXiv preprint arXiv:2210.04186*.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.

Birnbaum, M. H. (2004). Causes of Allais common consequence paradoxes: An experimental dissection. *Journal of Mathematical Psychology*, *48*(2), 87–106.

Birnbaum, M. H. (2005). A comparison of five models that predict violations of first-order stochastic dominance in risky decision making. *Journal of Risk and Uncertainty*, *31*(3), 263–287.

Birnbaum, M. H. (2008). New paradoxes of risky decision making. *Psychological Review*, *115*(2), 463.

Biswas, S. (2023). Role of ChatGPT in computer programming: ChatGPT in computer programming. *Mesopotamian Journal of Computer Science*, *2023*, 8–16.

Blair-Stanek, A., Holzenberger, N., & Van Durme, B. (2023). Can gpt-3 perform statutory reasoning? *arXiv preprint arXiv:2302.06100*.

Bonatti, E., Kuchukhidze, G., Zamarian, L., et al. (2009). Decision making in ambiguous and risky situations after unilateral temporal lobe epilepsy surgery. *Epilepsy & Behavior*, *14*(4), 665–673.

Bordt, S., & von Luxburg, U. (2023). ChatGPT participates in a computer science exam. *arXiv preprint arXiv:2303.09461*.

Bracha, A., & Brown, D. J. (2012). Affective decision making: A theory of optimism bias. *Games and Economic Behavior*, *75*(1), 67–80.

Buckert, M., Schwieren, C., Kudielka, B. M., & Fiebach, C. J. (2014). Acute stress affects risk taking but not ambiguity aversion. *Frontiers in Neuroscience*, *8*, 82.

Camerer, C., & Weber, M. (1992). Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty*, *5*(4), 325–370.

Camerer, C. F., & Ho, T.-H. (1994). Violations of the betweenness axiom and nonlinearity in probability. *Journal of Risk and Uncertainty*, *8*(2), 167–196.

Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., ... others (2023). A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., ... others (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Custers, B. (2022). AI in criminal law: An overview of AI applications in substantive and procedural criminal law. *Law and Artificial Intelligence: Regulating AI and Applying AI in Legal Practice*, 205–223.

Davis, E. (2023). Mathematics, word problems, common sense, and artificial intelligence. *arXiv preprint arXiv:2301.09723*.

Dean, M., & Ortoleva, P. (2017). Allais, Ellsberg, and preferences for hedging. *Theoretical Economics*, *12*(1), 377–424.

De Groot, K., & Thurik, R. (2018). Disentangling risk and uncertainty: When risk-taking measures are not about risk. *Frontiers in Psychology*, *9*, 2194.

Diederich, A., & Busemeyer, J. R. (1999). Conflict and the stochastic-dominance principle of decision making. *Psychological Science*, *10*(4), 353–359.

Dziri, N., Lu, X., Sclar, M., Li, X. L., Jian, L., Lin, B. Y., ... others (2023). Faith and fate: Limits of transformers on compositionality. *arXiv preprint arXiv:2305.18654*.

Ellis, A. (1976). The biological basis of human irrationality. *Journal of Individual Psychology*, *32*(2), 145-168.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics*, 643–669.

Gandhi, K., Fränken, J.-P., Gerstenberg, T., & Goodman, N. (2023). Understanding social reasoning in language

models with language models. In *Proceedings of the 37th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ghirardato, P., Maccheroni, F., Marinacci, M., & Siniscalchi, M. (2003). A subjective spin on roulette wheels. *Econometrica*, *71*(6), 1897–1908.

Gilboa, I., & Schmeidler, D. (1989). Maxmin expected utility with a non-unique prior. *Journal of Mathematical Economics,*, *18*, 141–153.

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work? A literature review of empirical studies on gamification. In *Proceedings of the 47th Hawaii International Conference on System Sciences* (pp. 3025–3034).

Han, S. J., Ransom, K. J., Perfors, A., & Kemp, C. (2024). Inductive reasoning in humans and large language models. *Cognitive Systems Research*, *83*, 101155.

Hummel, D., & Maedche, A. (2019). How effective is nudging? a quantitative review on the effect sizes and limits of empirical nudging studies. *Journal of Behavioral and Experimental Economics*, *80*, 47–58.

Imani, S., Du, L., & Shrivastava, H. (2023). Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398*.

Johnson, J. G., & Busemeyer, J. R. (2010). Decision making under risk and uncertainty. *Cog. Sci.*, *1*(5), 736–749.

Kahneman, D., & Tversky, A. (1979). Prospect theory: Analysis of decision under risk. *Econometrica*, *47*, 263–291.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, 341–350.

Katsikopoulos, K. V., & Gigerenzer, G. (2008). One-reason decision-making: Modeling violations of expected utility theory. *Journal of Risk and Uncertainty*, *37*(1), 35.

Klayman, J. (1995). Varieties of confirmation bias. *Psychology of Learning and Motivation*, *32*, 385–418.

Knight, F. H. (1921). *Risk, uncertainty and profit* (Vol. 31). New York, NY: Sentry Press.

Kortemeyer, G. (2023). Could an artificial-intelligence agent pass an introductory physics course? *Physical Review Physics Education Research*, *19*(1), 010132.

Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Kumar, P., Chauhan, S., & Awasthi, L. K. (2023). Artificial intelligence in healthcare: review, ethics, trust challenges & future research directions. *Engineering Applications of Artificial Intelligence*, *120*, 105894.

Lieder, F., & Griffiths, T. (2016). Helping people make better decisions using optimal gamification. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Liga, D., & Robaldo, L. (2023). Fine-tuning gpt-3 for legal rule classification. *Computer Law & Security Review*, *51*, 105864.

Lu, P., Qiu, L., Yu, W., Welleck, S., & Chang, K.-W. (2022). A survey of deep learning for mathematical reasoning. *arXiv preprint arXiv:2212.10535*.

Machina, M. J., & Siniscalchi, M. (2014). Ambiguity and ambiguity aversion. In *Handbook of the Economics of Risk and Uncertainty* (Vol. 1, pp. 729–807). Elsevier.

Mansour, S. B., Jouini, E., & Napp, C. (2006). Is there a pessimistic bias in individual beliefs? evidence from a simple survey. *Theory and Decision*, *61*, 345–362.

Mehr, H., Ash, H., & Fellow, D. (2017). Artificial intelligence for citizen services and government. *Ash Cent. Democr. Gov. Innov. Harvard Kennedy Sch., no. August*, 1–12.

Meier, S., & Sprenger, C. (2010). Present-biased preferences and credit card borrowing. *American Economic Journal: Applied Economics*, *2*(1), 193–210.

Mischel, W., & Ebbesen, E. B. (1970). Attention in delay of gratification. *Journal of Personality and Social Psychology*, *16*(2), 329.

Neumann, O., Guirguis, K., & Steiner, R. (2023). Exploring artificial intelligence adoption in public organizations: A comparative case study. *Public Management Review*, 1–28.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Nobandegani, A. S., Campoli, W., & Shultz, T. R. (2019). Bringing order to the cognitive fallacy zoo. In *Proceedings of the 17th International Conference on Cognitive Modeling*. Montreal, QC.

O'Donoghue, T., & Rabin, M. (2015). Present bias: Lessons learned and to be learned. *American Economic Review*, *105*(5), 273–279.

OpenAI. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

O'Sullivan, O. P. (2015). The neural basis of always looking on the bright side. *Dialogues in Philosophy, Mental & Neuro Sciences*, *8*(1).

Pearl, J. (2016). The sure-thing principle. *Journal of Causal Inference*, *4*(1), 81–86.

Peters, U. (2022). What is the function of confirmation bias? *Erkenntnis*, *87*(3), 1351–1376.

Pinker, S. (2015). The psychology of pessimism. *Catos Letter*, *13*(1), 2–8.

Pothos, E. M., & Busemeyer, J. R. (2009). A quantum probability explanation for violations of rationaldecision theory. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1665), 2171–2178.

Savage, L. J. (1954). *The Foundations of Statistics*. John Wiley & Sons Inc., New York.

Savelka, J., Agarwal, A., Bogart, C., Song, Y., & Sakr, M. (2023). Can generative pre-trained transformers (GPT) pass assessments in higher education programming courses? *arXiv preprint arXiv:2303.09325*.

Seaborn, K., & Fels, D. I. (2015). Gamification in theory and action: A survey. *International Journal of Human-Computer Studies*, *74*, 14–31.

Sharot, T. (2011). The optimism bias. *Current Biology*, *21*(23), R941–R945.

Speekenbrink, M., & Shanks, D. R. (2013). *Decision Making.* In *The Oxford Handbook of Cognitive Psychology.* Reisberg (Ed.). Oxford University Press.

Surameery, N. M. S., & Shakor, M. Y. (2023). Use ChatGPT to solve programming bugs. *International Journal of Information Technology & Computer Engineering*, *3*(01), 17–22.

Sushina, T., & Sobenin, A. (2020). Artificial Intelligence in the criminal justice system: Leading trends and possibilities. In *International Conference on Social, Economic, and Academic Leadership* (pp. 432–437).

Taylor, I. (2023). Justice by algorithm: The limits of AI in criminal sentencing. *Criminal Justice Ethics*, *42*(3), 193–213.

Tejeda, H., Kumar, A., Smyth, P., & Steyvers, M. (2022). AI-assisted decision-making: A cognitive modeling approach to infer latent reliance strategies. *Computational Brain & Behavior*, *5*(4), 491–508.

Thaler, R. H. (1994). *Quasi rational economics*. Russell Sage Foundation.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Tian, Y., Ravichander, A., Qin, L., Bras, R. L., Marjieh, R., Peng, N., … Brahman, F. (2023). Macgyver: Are large language models creative problem solvers? *arXiv preprint arXiv:2311.09682*.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323.

Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

van Noordt, C., & Misuraca, G. (2022). Artificial intelligence for the public sector: Results of landscaping the use of AI in government across the european union. *Government Information Quarterly*, *39*(3), 101714.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Vereschak, O., Bailly, G., & Caramiaux, B. (2021). How to evaluate trust in ai-assisted decision making? a survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–39.

Von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior*. Princeton University Press.

Wang, Duan, J., Fox, D., & Srinivasa, S. (2023). Newton: Are large language models capable of physical reasoning? *arXiv preprint arXiv:2310.07018*.

Wang, Lu, Z., & Yin, M. (2022). Will you accept the AI recommendation? Predicting human behavior in AI-assisted decision making. In *Proceedings of the ACM Web Conference* (pp. 1697–1708).

Wang, & Yin, M. (2021). Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In *26th International Conference on Intelligent User Interfaces* (pp. 318–328).

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526–1541.

Weber, B. J., & Tan, W. P. (2012). Ambiguity aversion in a delay analogue of the Ellsberg paradox. *Judgment and Decision Making*, *7*(4), 383–389.

Weber, M., & Camerer, C. (1987). Recent developments in modelling preferences under risk. *Operations Research Spektrum*, *9*(3), 129–151.

West, C. G. (2023). Advances in apparent conceptual physics reasoning in gpt-4. *arXiv e-prints*, arXiv–2303.

Wilkinson, T. M. (2013). Nudging and manipulation. *Political Studies*, *61*(2), 341–355.

Wu, Y., Jia, F., Zhang, S., Wu, Q., Li, H., Zhu, E., … Wang, C. (2023). An empirical study on challenging math problem solving with gpt-4. *arXiv preprint arXiv:2306.01337*.

Xu, F. F., Alon, U., Neubig, G., & Hellendoorn, V. J. (2022). A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming* (pp. 1–10).

Yasunaga, M., Chen, X., Li, Y., Pasupat, P., Leskovec, J., Liang, P., … Zhou, D. (2023). Large language models as analogical reasoners. *arXiv preprint arXiv:2310.01714*.

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., … Li, Y. (2021). A review of artificial intelligence (ai) in education from 2010 to 2020. *Complexity*, *2021*, 1–18.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., … others (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zong, M., & Krishnamachari, B. (2023). Solving math word problems concerning systems of equations with GPT-3. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 37, pp. 15972–15979).