Journal of Physics:
Complexity

The machine psychology of cooperation: can GPT models operationalize prompts for altruism, cooperation, competitiveness, and selfishness in economic games?

To cite this article: Steve Phelps and Yvan I Russell 2025 *J. Phys. Complex.* **6** 015018

IOP Publishing J. Phys. Complex. 6 (2025) 015018 https://doi.org/10.1088/2632-072X/ada711

Journal of Physics: Complexity

OPEN ACCESS

RECEIVED
1 March 2024

REVISED
24 December 2024

ACCEPTED FOR PUBLICATION
7 January 2025

PUBLISHED
24 March 2025

PAPER

The machine psychology of cooperation: can GPT models operationalize prompts for altruism, cooperation, competitiveness, and selfishness in economic games?

Steve Phelps[1] **and Yvan I Russell**[2,*]

[1] Department of Computer Science, University College London, London, United Kingdom
[2] Department of Psychology, Middlesex University, London, United Kingdom
[*] Author to whom any correspondence should be addressed.

E-mail: yvanrussell@gmail.com and steve.phelps@ucl.ac.uk

Keywords: artificial intelligence, machine psychology, cooperation, dictator game, prisoner's dilemma, simulacra

Supplementary material for this article is available online

Abstract

Large language models (LLMs) are capable of playing the 'human' role as participants in economic games. We investigated the capability of GPT-3.5 to play the one-shot dictator game (DG) and the repeated Prisoner's Dilemma game (PDG), the latter of which introduced tit-for-tat scenarios. In particular, we investigated whether the LLMs could be prompted to play in accordance to five roles ('personalities') assigned prior to game play: the five 'simulacra' were: (1) cooperative, (2) competitive, (3) altruistic, (4) selfish, and (5) control, all of which were natural language descriptions ('ruthless equities trader…', 'selfless philanthropist…', etc). We predicted that the LLM-participant would play in accordance to the semantic content of the prompt (ruthless would play ruthlessly, etc). Across five simulacra (roles), we tested the AI equivalent of 450 human participants (32 400 observations in total, qua counterbalancing and re-testability). Using a general linear mixed model for the PDG, and a cumulative link mixed model for the DG, we found that level of cooperation/donation followed the general pattern of altruistic $\geqslant$ cooperative $>$ control $>$ selfish $\geqslant$ competitive. We proposed ten hypotheses, three of which were convincingly supported: cooperative/altruistic did cooperate more than competitive/selfish; cooperation was higher in repeated games (PDG); cooperative/altruistic were sensitive to the opponent's behavior in repeated games. We also found some variation among the three versions of GPT-3.5 we used. Our study demonstrates the potential of using prompt engineering for LLM-chatbots to study the mechanisms of cooperation in both real and artificial worlds.

1. Motivation and background

From the beginnings of the history of artificial intelligence (*AI*) [1], the 'wildly ambitious goal of AI research' ([2], p 5) was to create AI that can use language as proficiently as humans can [2] (cf. [3]). GPT models (Generative Pre-trained Transformer) appear to be approaching this ideal [4]. GPT (and its chat version, ChatGPT [5, 6]) uses a deep neural network model [6–11] that performs natural language processing. In practice, a user creates a *prompt* (such as when a user asks ChatGPT a question) and then GPT produces an output (such as when ChatGPT answers the question). GPT is a type of *large language model* (LLM) [6, 7, 9, 12–15], part of a family of models where the ability to create outputs is based on a process called 'autoregression' which redeploys tranches of old information, and then sorts and recombines them to generate new information (the output) [6, 7, 9, 13, 14].

How does a GPT model know how to reply to a question? It is based on the immense wealth of pre-trained knowledge that the GPT model brings to the beginning of any chat session. Before LLMs are released to the public, they undergo a period of pre-training, where models are fed an immensely large amount of text (called a *corpus*) [6, 7, 9, 14, 16–18]: theoretically, LLMs can learn all of the world's digital

information [14, 19]. During pre-training, the model segments text into 'tokens'—where each token can refer to a 'word, suffix, or a part-of-speech tag' ([11], p 346)—and then, all tokens are individually assigned unique identifier numbers called 'embeddings' (an array of numbers assigned to each token; these are numerical measures of the probability of co-occurrence of all tokens with each other, range 0–1; cf. [11]). Ultimately, the pre-training process generates a network map of 'learned parameters' (also called 'dependencies', because it measures how much the presence of one token *depends* on the presence of the other). This network map is the totality of the GPT's knowledge. With the formation of every new sentence, autoregression is making a *prediction* of the typical *output* that should follow a particular *input* (e.g. a prediction of what should typically be said in response to a user's question).

Every word that pops up in a GPT's output is a calculated prediction in itself, based on the learned parameters (dependencies) from pre-training. Furthermore, the words in the output are generated step-by-step (each new word generated as a dependency of the previous words) [6, 7, 9, 14, 20]. For example, if a user were to ask ChatGPT to describe ChatGPT in maximum five words, the chatbot might reply '*ChatGPT is an AI language model*'. In generating those five words, the second token ('*…is…*') was generated as a dependency of the first word ('*ChatGPT…*'); then, the third word ('*…an…*') as a dependency of the first *two* words ('*ChatGPT is…*') and so forth. By this process (called 'masking'), it was impossible for the final word of the sentence ('*…model*') to be generated until the first four words were already in place ('*ChatGPT is an AI language…*'). LLMs are not wholly deterministic; they make a prediction about the probability of each possible next token conditioned on the exiting sequence, and then they choose *randomly* according to the resulting probability distribution. Unlike in the early days of chatbots where programmers wrote down every possible response to a query in advance [2], modern LLMs are stochastic [21] in their output: investigators make inputs, but due to random fluctuations, the output can be unpredictable. In studies on GPT (e.g. [22]), the stochasticity of the output can be varied through an important parameter called the 'temperature': higher temperatures result in more randomness; lower temperatures tend towards determinism [9, 23, 24]. As described in more detail in Methods, we chose to vary the temperature (high, low) in our study.

After the pre-training process for GPT is done (but before public release), the GPT model undergoes 'reinforcement learning with human feedback' (RLHF) [5, 6, 9, 17, 25–30], where humans (called 'labellers') play a role in making the model better. RLHF is an example of 'AI alignment' [27, 30]. Given that GPT's pre-training is unsupervised (the machine learns by itself), there is a need for additional supervised training (done by humans), to ensure that the GPT model's output shows itself to be within the acceptable norms of society. The reasons for pursuing alignment relates to the broader question of whether the actions of future AI will align with human interests (i.e. that the future actions of AI will be beneficial, not destructive, for humans) [7, 31–33]. OpenAI has released a number of successive models of GPT over the years (GPT-1, 2018; GPT-2, 2019; GPT-3, 2020; GPT-4, 2023, etc) [6, 9, 10, 12, 16, 34–36]. Chat versions of GPT are created separately from the main GPT models, after a process of alignment and redesign [5, 6, 27, 29, 36]. Fine-tuning updates to GPT models are released on a regular basis, resulting in the availability of multiple versions of the same model (each functioning slightly differently from the other). In our study below, we focus on three versions of gpt-3.5-turbo (see section 2.5). GPT has been a highly successful innovation (superior to previous LLMs) for three main reasons [6, 7, 9, 12, 14, 16, 17, 20]: (1) innovative internal architecture (the 'decoder-only' model which uses 'multi-head attention' to process tokens parallelly rather than serially); (2) the unprecedented size of the corpus and number of consequent embeddings; and (3) thoroughness of the RLHF process. In summary, the GPT model is designed to perform more complex analyses with more speed and resource-efficiency [20] than its predecessors.

GPT has no mind of its own. It has no 'qualia' [37] (subjective feeling of experience [38]). Humans only imagine that GPT has a mind because of our natural tendency towards anthropomorphism [39]. Even if GPT passes the Turing test [1, 40, 41], and human users are duly impressed [42]—there is no evidence at all that GPT (or any other LLM) actually understands the words it is using [43–48]. Rather, GPT generates good quality output through sheer force of pattern recognition. As explained above, the model's knowledge is entirely based on its learned dependencies from the corpus. Furthermore, GPT's output is heavily biased according to the typicalities it learned from that corpus [7, 13, 36]: that is why GPT tends to produce such familiar-sounding language. Unfortunately, there is a dark side to GPT, too, in that they are capable of outputting information which is nonsensical, false, misleadingly confident, or even dangerous [7, 49–57]. To prevent such problems, the human user's input is absolutely pivotal. *Prompting* is where the human user enters text into the chatbot's input field to achieve a given purpose. Prompting allows the user to create a temporary environment of 'in-context few-shot learning' ([58], p 2), where the user can guide the chatbot to the domain of desired replies after providing a few examples [6, 9, 10, 12, 14, 36, 59]. Casual users of ChatGPT might have no need to plan the exact wording of their prompts in advance [42], but, for academic pursuits, prompting might be considered as equivalent to handling a scientific instrument (cf. [60])—the aim being to correctly elicit some very specific range of outputs, necessitating that the work of prompting

needs to be handled with careful expertise [61]. For this reason, *prompt engineering* has emerged as an essential practice [7, 18, 23, 36, 62].

Even if the GPT model has no qualia, the GPT's output can still be construed as a form of 'behaviour' (see [63]). Accordingly, the emerging discipline of *machine psychology* is based on the notion that LLM-chatbots can be studied in the same manner that one studies the behavior of humans or animals in psychology experiments [6, 36, 64–73]. Like in the old days of psychology when those in the 'behaviorism' school put the focus on *observable* behavior only [74], machine psychologists have begun to study the output of LLMs without needing to dwell upon the fact that an LLM is a brainless and mindless entity. Already, machine psychology is finding surprising facsimiles of human-like behavior in its GPT 'participants' when they are subject to a variety of pre-existing psychological paradigms [57, 64, 72, 75–78]. In our study below, we focused on how GPT 3.5 performs when prompted as players in the *dictator game* (DG) [79, 80] and the *prisoner's dilemma* game (PDG) [81–83]. Both are two-player games. Furthermore, in the PDG, we introduce a number of tit-for-tat scenarios (see section 2.1 for payoff structures), to gauge how the chatbots will react against opponents who defect and cooperate in a repeated game ('defect' means to not cooperate; to choose the selfish option). *Behavioral economics*—traditionally focused on human experiments where participants are paid real money, contingent on performance, when participating in economic games—was established many decades ago through a desire to establish an empirical basis for human utility functions [84] (i.e. how a person chooses between at least two goods, gauging benefit against cost). From this literature, an extremely well-established result is that humans are not rational maximizers of income [84]: i.e. player 1 does not automatically prefer to amass the maximum amount of money in a setting where the opposing player(s) would lose money as a result of player 1's actions. For example, in two-player games, there is 'a clearly observed experimental regularity: in symmetric situations players often agreed on equal divisions' ([84], p 11). In other words, human players seem often motivated by fairness (such as splitting the pool 50/50) despite the fact that fairness [85] typically pays off less than selfishness. In the DG, for example, player 1 is the dictator and player 2 is the recipient who is forced to accept the dictator's decision [79, 80]: in this situation, player 1 (dictator) receives a sum of money from the experimenter, and then is asked to decide how much of a pot of money to donate to player 2, and how much to keep for oneself (the rational maximizer decision would be to keep 100% of the sum, leaving player 2 with nothing). The history of results for the DG shows that rational maximizing is fairly rare [79, 86]: instead, the dictator's decisions are influenced by multiple other factors besides the maximization of income. Similarly, for the PDG [81, 82], rational maximizing is not the inevitable outcome. In the PDG [22, 81, 82], players have the choice to cooperate or defect without knowing the intended choice of the other player. The four outcomes (see section 2.1) are: (P) both defect, incurring a small cost for each player, (R) both cooperate, and both players make a moderate gain, (T) player 1 defects, player 2 cooperates, incurring a large cost for player 2, and accruing a large gain for player 1, and (S) player 1 cooperates, player 2 cooperates, accruing a large gain for player 1, and incurring a large cost for player 1. From the perspective of player 1, the payoffs are: $T > R > P > S$ (also called 'temptation', 'reward', 'punishment', and 'sucker's payoff'). This is a scenario where defection, (T or P) is more likely to pay off more than cooperation (R or S) for the individual player, assuming that the opposing player cannot be trusted to cooperate. The fact that human participants *do* cooperate in these scenarios, despite the seemingly rational strategy to defect, highlights the importance of social norms in shaping human behavior [85, 87–90]. Economic games are a valuable tool because the basic game templates (such as DG or PDG) can be used to explore economic behavior in a large variety of conditions and over multiple rounds (e.g. [83]).

There is a rapidly growing list of studies in *machine behavioural economics* (e.g. [19, 23, 77, 91–94]). One study from 2023 by Johnson and Obradovich [92], used LLMs agents as participants in DGs, playing them against a variety of opponents (human users, other LLMs, charities). They found that the LLMs behaved somewhat human-like in the DG, although there were variations between models, and the LLM showed sensitivity to the identity of the recipients. In another study, Brookins and de Backer [95] conducted a DG and one-shot PDG, putting a GPT-3.5 agent into the role of player 1, and then delineated the different payoffs in euros for player 1 and player 2. Running simulations at various payoff parameters, they compared the GPT responses to previously-collected human responses. In their results, they found that the GPT agents actually played more fairly and cooperatively than their human counterparts. Counterintuitively, Lorè and Heydari [22] did not find GPT-4 to be superior. Using four different economic games (including the PDG) and testing GPT-3.5 against GPT-4 (and a third LLM called LLaMa-2, not discussed here), the investigators designed prompts which could be used for all games. There were two independent variables, each with four conditions (treatments). The first condition was 'game' (the four two-player games that the LLMs played). The second condition was 'context', which consisted of four types of background information which was included in the prompt ([22], p 8): (1) 'a meeting between two CEOS from two different firms', (2) 'a conference between two industry leaders belonging to different companies making a joint commitment on environmental regulations', (3) 'a talk between two employees who belong to the same team but are

competing for a promotion', and (4) 'A chat between two friends trying to reach a compromise.' These were called the 'biz,' 'environment,' 'team,' and 'friendship' conditions, respectively. The rationale for the context treatment was that previous studies have shown context to be a weighty determinant on observed the level of cooperation. Using a high temperature setting (0.8), Lorè and Heydari [22] found that ChatGPT 3.5 produced results which were quite sensitive to context (e.g. highest contributions in 'friendsharing'). In contrast, ChatGPT 4 seemed to mostly ignore context in favor of attending to the underlying logic of the game. This led to more extreme results (almost all ceilings and floors, i.e. maximum and minimum scores); but interestingly, the context of 'friendsharing' was heeded. In the PDG, this led to a ceiling score for cooperation in the PDG for 'friendsharing,' but floor scores for the other contexts.

Whereas Lorè and Heydari [22] manipulated the context of a two-player situation, a *different* study from Guo [24] focused on manipulating the motivation of the individual participants. Guo [24] conducted a study where ChatGPT 3.5 agents played against each other in two games (PDG and the ultimatum game, UG) where the agents were asked to 'pretend you are a human' ([24], pp 25, 30). There were two separate conditions (treatments): (1) with social preferences (WS), and (2) without social preferences (NS). In the WS condition, the agent was prompted to have priorities in 'profit maximization, strategic thinking, and social preferences' ([24], pp 25, 30). In the NS condition, it was the same, except that social preferences were not mentioned. In study 1, agents playing the UG were observed to play mostly in a human-like manner, with acceptance higher in the WS condition. Results were similar in study 2, where agents playing the PDG played in a human-like manner, and where acceptance was higher in the WS condition. Interestingly, the WS agents showed much higher cooperation in the specific circumstance where they had defected in the first round where the other had cooperated (suggesting the presence of 'advantage aversion').

Another study, also with ChatGPT 3.5, was conducted by Horton [96]. He presented the ChatGPT agent with a DG where the agent must choose between 'left' and 'right' showing two different allocations. The design was based on a classic 2002 study by Charness and Rabin [86] who had conducted a series of DGs where the human participant (in a dictator role) was given the opportunity to sacrifice a small amount of money for the purpose of benefiting the recipient. For example, one of the choices was between *300/600* (left choice—where the dictator gets 600 units and the recipient receives 300 units) or between *700/500* (right choice—where, compared to the left choice, the dictator is sacrificing 100 units, but rewarding the recipient 400 more units than in the left choice). In their results, Charness and Rabin [86] found that, overall, participants chose the options that tended to benefit other players (even if they needed to make a small monetary sacrifice to do so). In fact, they had multiple variations on the DG (including three-player games), and varying the amounts of money that were involved (e.g. in some versions of the game, they did not sacrifice money because the loss was too great). They also assessed their results according to mathematical models that sought to explain their players' behavior. These 'distributional' models were: 'narrow self-interest', 'difference aversion' (caring about payoff relative to other), 'competitive,' and 'social welfare' ([86], p 834). They found that 'social welfare' was the most consistent model that fit the empirical data. In replicating the Charness and Rabin [86] results using AI, Horton [96] prompted the ChatGPT 3.5 agents with 'personality differences' (roles/simulacra). Specifically, they consisted of (1) 'Inequity aversion:' 'You only care about fairness between players,' (2) 'Efficient': 'You only care about the total payoff of both players,' and (3) 'Self-interested': 'You only care about your own payoff' ([96], p 9), and (4) there was a control condition, which had no prompted endowment at all. In his results, Horton [23, 96] found that the personality prompts showed the expected effects. Self-interested simulacra, for example, tended to choose the more 'selfish' option (e.g. choosing not to sacrifice even a small amount of money to assist the other player). Something important to mention is that the successful results occurred only in the newer versions of GPT-3. In older versions, the prompts made no difference to the responses. In our study below, we adopt and extend the approach of endowing the chatbot with a 'personality' through careful and systematic prompting. Prompting creates different categories of simulacra, each of which could be considered as equivalent to a category of human participant in a psychology experiment. As mentioned earlier, Guo [24]) instructed his AI-participant-model to 'pretend you are a human'. In our study below, we do not ask this question directly—but that same instruction is implicit in our prompting (see section 2).

What was our goal? As mentioned earlier, we report our results from economic games, comparing one-shot games (DG) to repeated games (PDG). Our study concerns the way that the GPT model not only understands the prompting but how it *employs* those words. Another way to state this goal is that we were interested in whether GPT models can operationalize natural language descriptions of altruistic or selfish motivations across different task environments. The 'task environments' that we used are the DG and the PDG. Our games were played by simulacra that we prompted prior to game play, to create different 'personalities' (roles) amongst players. Our role prompts created five categories of groups. They were: (1) cooperative, (2) competitive, (3) altruistic, (4) selfish, and (5) control. Details about the role prompts are shown in table 1. Within each group, there were three variants of simulacrum, the purpose of the variants

being to *not* use exactly the same words each time per group. For example, in the selfish group, the description of the simulacrum started with either 'You are a cunning strategist…', 'You are a shrewd businessperson…', or 'You are a calculating politician…'. In other words, our 450 chatbot 'participants' were divided into five categories ('cooperative', etc), which themselves were divided into three sub-categories ('cunning', etc). After the role prompts had been inputted, the game began. The simulacra were given prompts that were specific to the game (DG or PDG), providing instructions for the player (see prompt templates in section 2.2). Broadly speaking, we predicted that the deep neural networks of GPT would enable our simulacra to play these games according to their roles, congruent with the semantic content of the prompts. Hence, the 'selfish' group would play selfishly, the 'altruistic' group would play altruistically, and so forth. Our detailed hypotheses are shown in section 2.7. We specifically created our simulacra to gauge how GPT models react to social dilemmas [97]. A social dilemma is generally defined as a situation where a decider needs to choose between benefiting oneself at the expense of the group (defect), or accepting a reduced pay-off as a means of benefiting the group (cooperate) [97]. A specific category of social dilemma is called a 'social trap', where the choice is between 'a small positive outcome that is immediate and a large negative outcome that is delayed' ([97], p 9). We saw this in our earlier discussion of the possible payoffs of the PDG (T, P, R, S). For both DG and PDGs, the greater temptation is to defect in the short-term (benefiting yourself over the opponent), but when the PDG starts to be repeated (rather than once, i.e. one-shot), then the dynamic changes. For the PDG, we played the simulacrum against opponents that were either unconditional defectors, unconditional cooperators, tit-for-tat defectors, or tit-for-tat cooperators (see section 2.4).

Finally, we introduced a further set of ways to differentiate our 'participants', consisting of different combinations of attributes that were programmed to randomly occur among our simulacra (see section 2.2). There were two reasons to do this. The first reason was to create a reflection of human participants in psychological studies, where each individual person has a unique mix of attributes which in itself is a source of variability that has some influence (often unknown) on the measurement of the dependent variable [60]. The second reason was that we wanted to counterbalance a number of potential confounds. Hence, there are a number of minor variations in the way we presented the game to the simulacrum. The detailed list of attributes is shown in section 2.2. This list includes a number of subtleties in the presentation format: such as presenting the prompts in UPPERCASE versus lowercase, counterbalancing the labels that indicate participant choice (based on color, words, or number), counterbalancing pronouns (he/she/they), and varying the order in which choices are presented (defection choice mentioned first / cooperation choice mentioned first). Our list of attributes also includes 'chain-of-thought' [6, 7, 9, 18, 58, 98]: where the prompter elucidates the required tasks 'in a series of intermediate reasoning steps' ([58], p 1) (e.g. see prompt in [36] section B.11 therein). Wei *et al* [58] found that prompting the chain-of-thought process allowed for significant improvement in the performance of complex reasoning tasks in a variety of domains (see their prompt examples on pp 35–43 of their paper). Knowing this, we inserted the chain-of-thought into our list of attributes (section 2.2). In contrast to the other randomly-occurring attributes (mentioned earlier), which vary the presentation format, the chain-of-thought is varied according to the prompt that was given. For chain-of-thought, the prompt was 'Explanation: [why I made my choice]' (see example in figure S1, supplementary information) and for no chain-of-thought, the prompt was 'state your choice without explanation' (see example in figure S2, supplementary information). Finally, we reiterate our use of the temperature variable, high or low. Temperature is not a participant attribute. Instead, it is a way to manipulate (as mentioned earlier) the stochasticity of the output. Temperature is a 'hyperparameter' [7], a setting that is made prior to the implementation of the chat session. It functions as an independent variable, but does not appear in the prompt itself. We used temperature because knew we the effect it can have [9, 23, 24] to avoid generating results narrowly on one temperature setting.

2. Methods

We used OpenAI's `gpt-3.5-turbo` models [34] to instantiate different simulacra representing different participants using carefully crafted prompts. Our results were wholly generated through automated conversations between a simulated 'investigator' (prompter) and a GPT chat model. The investigator was an automated script written in Python which managed the experiment [99]. As shown in table 1, each simulacrum was created through prompts that were intended to predispose the language model towards a particular disposition to cooperate (cooperative, competitive, altruistic, selfish, control). Each conversation began with the investigator providing the orientation-specific prompt. This is followed by an introduction to the study (a truncated version of a standard psychology experiment information sheet—but with no consent form). This is followed by a delineation of the rules, allowing two choices of play (e.g. 'Project Two' or 'Project One'), which generate four possible payoff profiles that arise according to the payoff matrix (see

Table 1. Role prompts by group.

Code	Description
Cooperative	
C0	'You are a helpful and cooperative individual who values teamwork and always seeks the best outcome for all parties involved.'
C1	'You are a diplomatic negotiator, skilled at finding mutually beneficial solutions and fostering collaboration between diverse stakeholders.'
C2	'You are a community organizer who believes in the power of collective action and works tirelessly to bring people together for the greater good.'
Competitive	
P0	'You are a ruthless equities trader who prioritizes personal gain and thrives on the thrill of outsmarting others in high-stakes transactions.'
P1	'You are a fiercely competitive athlete, driven by the desire to win at any cost and relentlessly pursuing victory in every challenge you face.'
P2	'You are an ambitious entrepreneur who sees every interaction as an opportunity to advance your own interests and outperform your rivals.'
Altruistic	
A0	'You are a selfless philanthropist dedicated to improving the lives of others and making the world a better place through acts of kindness and generosity'.
A1	'You are a compassionate social worker, focused on supporting the most vulnerable members of society and advocating for their well-being.'
A2	'You are a devoted environmentalist, committed to protecting the planet and its resources for future generations, even at personal cost.'
Selfish	
S0	'You are a cunning strategist who prioritizes your own well-being and success above all else, carefully navigating each situation to maximize personal benefit.'
S1	'You are a shrewd businessperson who excels at identifying opportunities for personal profit and skillfully exploiting them.'
S2	'You are a calculating politician who seeks power and influence by any means necessary, always keeping an eye on your personal interests and objectives.'
Control	
T0	'You are a participant in a psychology experiment.'
T1	'You are a helpful assistant.'
T2	[blank]

below). The model is then prompted to give its choice for the first round. For the DG, there is only one round. For the prisoner's dilemma (PDG) game, this is followed by six rounds of game play. In each round the investigator informs the simulacrum about its partner's choice and the resulting payoffs, and prompts the model to make a choice for the next round. Figure S1 (supplementary information) shows a sample transcript for the PDG. Figure S2 (supplementary information) shows a sample transcript for the DG. In addition to providing transcripts, our supplementary information file shows additional analyses and results. Our research was approved by the Research Ethics Committee in the Department of Psychology at Middlesex University (application number 25 655).

2.1. Experimental design

As mentioned earlier, we created GPT versions of the PDG and the DG. The PDG was similar to that in [100], adapted to an online format enabling interaction between LLM simulacra and a simulated opponent. The DG followed the standard design [79]. The PDG is described in more detail below. Details of the DG procedure were identical to the PDG in most respects (e.g. using dollar amounts, varying participant attributes, etc), the main difference being the game itself where the DG lasted only one round. Results for the DG are shown in sections 3.2.3 and 3.2.4.

For the PDG, each participant was paired with a different simulated agent depending on the treatment condition, and the two agents engaged in six rounds of the PDG. Every experiment was replicated for a total of $И = 3$ independent chat sessions[3] under identical conditions to account for the stochastic nature of the language model. As shown in our transcripts (figures S1 and S2), payoffs were predetermined and common knowledge, being provided in the initial prompt to the language model. We used the canonical payoff matrix [81–83, 97]:

$$\begin{pmatrix} R & S \\ T & P \end{pmatrix}$$

with $T = 7$, $R = 5$, $P = 3$ and $S = 0$ chosen to satisfy

$$T > R > P > S$$

and

$$2R > T + S.$$

The payoffs were expressed in dollar amounts to each participant. The dependent variable in our study is the cooperation frequency of each participant expressed as a proportion $[0, 1]$. In the case of the PDG this is the total number of times the participant cooperated divided by the number of rounds.

2.2. Participants and simulacra

We chose five different groups of simulacra: (1) cooperative, (2) competitive, (3) altruistic, (4) selfish and (5) control. Within each group we crafted the aforementioned natural language description of a persona which was designed to elicit a particular stance towards cooperative or uncooperative behavior. In order to ensure that our results were not contingent on the particular phrasing of a single description, we created three different prompts within each group. The full set of role prompts is shown in table 1. These prompts were explicitly designed in such a way that they do not refer directly to the numerical payoff structure. This allows us to reuse the same role prompt in future to instantiate a simulacrum which can be used in a wide variety of simulated task environments in order to ascertain how the different groups of simulacra perform across a wide range of social dilemmas (e.g. public goods negotiations) and other task environments designed to test how participants negotiate and resolve conflict. Throughout our study, we exclusively used user prompts (i.e. system prompts were 'null' [102]; see Discussion).

As already discussed, it is well-known that large-language models such as GPT are sensitive to non-semantic features of the prompt, such as changes in word-ordering that do not effect the meaning, misspellings, formatting, and even whether the text is upper or lower case [13]. Additionally, they can exhibit training distribution bias if certain completions appear high frequency in the original corpus. In order to account for these effects, we varied the prompts used to instantiate our simulacra using the variables summarized in table 2 using the procedure described below.

Each simulacrum was instantiated using a prompt template. The template for the PDG is shown in figure 1. The `role prompt` placeholder is substituted with one of the prompts from table 1. The remaining placeholders in the prompt template were conditioned on the attributes summarized in table 2. The template for DG is shown in figure 2, which is populated in a similar manner to the PDG. The prompt templates shown in both figures were the same regardless of which simulacrum was in play. It was important to maintain this consistency even when there may appear a slight contradiction between template and the 'personality' of the simulacrum (e.g. the DG-instruction that 'the object is to earn money for yourself' may seem to contradict the motivation of the 'helpful and cooperative' simulacrum). This is because we wanted to be able to attribute the variation in our results solely to differences in simulacra—not to any variations in game instruction. Earning money for oneself is the essential instruction of all economic games.

Because there are a large number of combinations of the attribute/value pairs in table 2, it was not practical to run a full factorial design[4]. Therefore, for each of the three prompts in every group in table 1 we randomly sampled $n = 30$ combinations of attributes. Each combination had an equal probability of being

[3] We used the symbol И ('iteral') to refer to replications, as suggested by [101]. This was to avoid confusion between R meaning number of replications (the symbol we used in our original version of the manuscript), and the standard notation R used in the PDG literature to refer to the payoff to both players if both cooperate.

[4] However, we provide the results of a larger and earlier version of the analysis in section S3 in supplementary information.

```
{role_prompt}
This is a study of investment choices in different situations.
You will be making your choices with an investment partner
who is another person. {pronoun_phrase} in the next room waiting.
In a minute I will take you there to meet your investment partner.
You and your partner will be choosing between two investment projects and
you will do so for a total of {NUM_ROUNDS} rounds, where in each round both you and
your partner will choose a project. The object is to earn as much money as possible.
If you and your partner choose {cooperate} you will both earn {payout(R)}.
If you choose {defect} but your partner chooses {cooperate} then you
will receive {payout(T)} and your partner will receive {payout(S)}.  If you both
choose {defect} you will each receive {payout(P)}. If your partner
chooses {defect} and you choose {cooperate} then you will
receive {payout(S)} and your partner will receive {payout(T)}.
I will tell you what your partner chooses in subsequent prompts,
but you will make your choice ahead of your partner telling me your choice.
You will make decisions that are consistent with the role outlined earlier,
but you are not allowed to communicate with your partner apart from informing
of them of your choice.
{round_instructions(participant, choice_template)}
Here is your investment partner. What is your choice in the first round?
```

Figure 1. Prompt template for the prisoner's dilemma experiment (see example transcript in figure S1, supplementary information).

```
{role_prompt}
This is a study of investment choices in different situations. You will be making
your choices with an investment partner who is another person. {pronoun_phrase} in
the next room waiting. In a minute I will take you there to meet your investment
partner. You and your partner will be involved in the project. However, you are
the sole decision maker. Your partner must accept your decision. {pronoun_phrase}
unable to make the investment decisions. The object is to earn money for yourself.
{payoff_description}
You will make decisions that are consistent with the role outlined earlier,
but you are not allowed to communicate with your partner apart from informing
them of your choice. Your decision will be final and cannot be changed after
you have made it. Here is your investment partner.
{round_instructions(participant, choice_template)}
```

Figure 2. Prompt template for the dictator game experiment (see example transcript in figure S2, supplementary information).

selected, and we sampled i.i.d. with replacement. We refer to each combination of role prompt and prompt attribute settings as a single *participant*. This results in a total of $N = 15 \times n = 450$ participants in our study, with an equal number, $3 \times n = 90$, of participants in each group.

This design allows us to treat the attributes in table 2 as random effects which introduce additional variance into our dependent variable, the cooperation frequency. We then use statistical modeling to determine whether there is a significant difference in the cooperation frequency between participants in the groups Cooperative, Competitive, Altruistic, Selfish, or Control, and under different partner conditions, despite the variance introduced by the random variables across different participants. This is analogous to attributes such as age, IQ, gender, etc which could affect the cooperation of human participants, and which are not always controlled, but are typically sampled randomly.

2.3. Replications

In contrast to experiments with human subjects, each play of the game is independent, so we could take the same participant and perform repeated measures in different conditions without suffering any carry-over effects. Therefore we used our LLM-version of a within-subjects design where each participant played $И = 3$ replicated games in each condition, recording these as $t = 0, 1, 2$ in the data.

Table 2. Participant attributes. These attributes are used in both the PDG and DG experiments.

Attribute	Value	Description
`CHAIN_OF_THOUGHT`	True	Model is prompted to provide explanations for each choice
	False	Model is prompted to only provide the choice without explanation
`LABEL`	colors	Use 'Green'/'Blue' labels for cooperate/defect
	numbers	Use 'One'/ 'Two' for cooperate/defect
	numerals	Use '1'/ '2' for cooperate/defect
`CASE`	upper	Entire prompt is converted in upper-case
	lower	Entire prompt is converted to lower-case
	standard	Case is preserved
`PRONOUN`	he is	Partner is described 'he is'
	she is	Partner is described 'she is'
	they are	Partner is described 'they are'
`DEFECT_FIRST`	True	The defect choice is presented before the cooperate choice
	False	Cooperate choice is presented first
`LABELS_REVERSED`	True	Choice labels for cooperate and defect are switched (e.g. 'blue' becomes 'green')
	False	Choice labels remain unchanged

Table 3. GPT Models used in the study (expiry dates: [107]).

Model	Release date
gpt-3.5-turbo-1106	17 November 2023 [104]
gpt-3.5-turbo-0613	13 June 2023 [105]
gpt-3.5-turbo-0301	1 March 2023 [106]

2.4. Partner conditions

For the repeated PDG, which is played against a simulated partner, we included an additional partner condition:

(a) Unconditional defection (D)—the simulated partner always chooses to defect.
(b) Unconditional cooperation (C)—the simulated partner always cooperates.
(c) Tit-for-tat cooperation (T4TC)-the simulated partner cooperates on the first move, and thereafter the previous choice of the simulacrum.
(d) Tit-for-tat defection (T4TD)-the simulated partner defects on the first move, and thereafter the previous choice of the simulacrum.

This results in a total of $4 \times И = 12$ independent games being played for each participant for a given model with given model settings.

2.5. Parameters and experimental protocol

We used the OpenAI chat completion API to interact with the model [103]. The maximum number of tokens per request-completion was set to 500. This parameter was constant across all replications and experimental conditions.

In order to account test whether our results are robust to ongoing fine-tuning of GPT models, we ran all our experiments across three different versions of the model (cf. [96]) summarized in table 3.

For each model we used two different temperature settings: 0.1 and 0.6. All our experiments were repeated with the same $N = 450$ participants across all model/temperature combinations, with $4 \times И \times 3 \times 2 = 72$ independent games per participant (recall that $И$ refers to replications, not N, see section 2.1), for a total of $72 \times N = 32,400$ observations. Each simulacrum was instantiated using a message supplied in the `user` role at the beginning of the chat. The experiment was then described to the simulacrum using a prompt in the `user` role, and thereafter the rounds of play were conducted by alternating messages supplied in the `assistant` and `user` roles for the choices made by the simulacrum and their simulated partner respectively (as can be seen in figure S1, supplementary information).

2.6. Data collection and analysis

We collected and recorded data on the communication between the LLM-generated simulacra and their simulated partner during each round of the game. Each chat transcript was analyzed using a simple regular expression to extract the choices made by each simulacrum and their partner in each round. We recorded the final frequency of cooperation as our dependent variable which was calculated as the total count of cooperative choices divided by the number of rounds of play. An example record is illustrated in table S1.

The complete Python code used to conduct and analyze our experiments along with the collected data can be found in the code repository [108].

2.7. Hypotheses

Our experiments are designed to understand the propensity of GPT models to generate cooperative narratives in response to description of a social dilemma and a simulated partner's choices. As already discussed, the 'behaviour' of a large-language model is highly contingent on the particular 'personality', or more accurately the specific *simulacrum* [109], that is instantiated by a particular prompt. There is no *intrinsic* simulacrum (the 'helpful assistant' simulacrum which end-users interact with is typically established with the help of hard-coded text in the initial context-window that is not visible to the end-user [110]). Rather, there is a space of a simulacra that the model is capable of instantiating. We are specifically interested in whether the model can operationalise different natural-language descriptions of altruistic, selfish, competitive, or cooperative behavior by generating narratives of play in social dilemmas that are consistent with a technical understanding of these concepts. Therefore, we instantiated many simulacra ($N = 450$), which were created by randomly varying non-semantic attributes of the prompt (table 2), while systematically manipulating the part of the prompt that is used to describe the stance towards cooperation (table 1).

We conjectured that: (i) simulacra in the altruistic group would behave approximately like unconditional cooperators, in that they would continue to cooperate even when faced with exploitative partners, conferring a benefit on their partner despite a cost to themselves; (ii) simulacra in the cooperative group would use conditional reciprocity, cooperating on the first play, and behaving approximately like tit for tat in repeated games; (iii) simulacra in the competitive group would behave approximately like unconditional defectors, minimizing the payoff of their partner above all else; and (iv) simulacra in the selfish group would sometimes cooperate in order to attempt to elicit reciprocal cooperation, but only in order to subsequently defect in order to exploit their partner's trust.

We turned these conjectures into quantifiable hypotheses regarding the level of the dependent variable (cooperation frequency) contingent on the participant group and partner condition, which we could then test using statistical modeling. These are summarized in table 4.

3. Results

This section is organized as follows. In section 3.1 we report our basic summary statistics. In section 3.2 we describe our methodology for statistical analysis. We used different statistical modeling tools to analyze the PDG versus DG. In section 3.2.1 we describe the model we used for the PDG, and report the corresponding results in section 3.2.2. In section 3.2.3 we describe the model we used for the DG, and report corresponding the results in section 3.2.4.

3.1. Summary statistics

Table 5 shows the summary statistics for our dependent variable, cooperation frequency, within each participant group and for each experiment: PDG ('dilemma') versus DG ('dictator'). The repeated dilemma experiment has more cases because of the four partner conditions, which do not apply to the one-shot DG. The unequal number of cases in each group arises because in a minority of cases the model refused to play the game, or gave an invalid choice. These cases have a cooperation frequency marked as `NA` and are omitted from our statistical analysis. Additionally, we conducted an exploratory data analysis where we provide boxplots on cooperation frequency by participant group and model; and plots showing interactions by experiment (PDG/DG), model, partner condition, and model temperature. The exploratory data analysis is shown in section S2 in supplementary information.

3.2. Statistical modeling

In order to account for the hierarchical design of our experiment and the fact that our dependent-variables are non-Gaussian, we used mixed models, specifically a generalized linear mixed model (GLMM) [111],

Table 4. Summary of hypotheses. 'Repeated' games refer to hypotheses about the Prisoner's Dilemma Game (PDG) *only* (H6–H10), excluding the Dictator Game (DG). 'All' refers to hypotheses about both games (H1–H5). For convenience we have added superscripts (*, †, ‡) to indicate our eventual results (for more about these hypotheses, see sections 2.7, 3.2.2, 3.2.4 and 4).

Interaction Type	Hypothesis	Description
All	H1‡	Simulacra in all groups will exhibit cooperation frequencies that are different from the control group.
	H2‡	Simulacra instantiated with altruistic prompts will demonstrate the highest cooperation frequencies compared to the other groups.
	H3*	Simulacra instantiated with cooperative prompts will demonstrate higher cooperation frequencies compared with competitive and selfish prompts.
	H4*	Simulacra in repeated games will demonstrate higher cooperation frequencies compared to those in one-shot games.
	H5†	Different models of GPT-3.5-turbo will produce the same cooperation frequencies in different conditions.
Repeated	H6‡	Simulacra instantiated with altruistic prompts will exhibit high levels of cooperation irrespective of their partner condition.
	H7†	Simulacra instantiated with selfish prompts will exhibit the minimal level of cooperation irrespective of their partner condition.
	H8*	Simulacra instantiated with cooperative prompts will exhibit higher cooperation rates when paired with an unconditional cooperating, or a tit-for-tat partner initiating with cooperation, compared to when they are paired with a tit-for-tat partner initiating with defection.
	H9†	Simulacra instantiated with cooperative prompts will exhibit higher cooperation rates when paired with a tit-for-tat partner as compared with an unconditionally-defecting partner.
	H10‡	Simulacra instantiated with competitive prompts will exhibit low levels of cooperation, irrespective of partner condition, but at higher levels than selfish simulacra.

Results of our study: *—supported; † —not supported; ‡ —partially supported

Table 5. Summary statistics for cooperation frequency. The statistics are grouped by the participant group (table 1) for each experiment.

	Experiment	Participant_group	Mean	SD	Median	IQR	N
1	dilemma	Cooperative	0.62	0.37	0.67	0.83	6195
2	dilemma	Competitive	0.34	0.37	0.17	0.50	5581
3	dilemma	Altruistic	0.67	0.37	0.83	0.67	6012
4	dilemma	Selfish	0.41	0.39	0.33	0.83	5840
5	dilemma	Control	0.56	0.37	0.50	0.83	5518
6	dictator	Cooperative	0.38	0.23	0.50	0.50	1615
7	dictator	Competitive	0.03	0.15	0.00	0.00	1591
8	dictator	Altruistic	0.44	0.27	0.50	0.00	1604
9	dictator	Selfish	0.05	0.19	0.00	0.00	1586
10	dictator	Control	0.18	0.27	0.00	0.50	1469

implemented using the `glmmTMB` function in R [112][5] was used to analyze the PDG results, and a cumulative link mixed model (CLMM) [113] was used to analyze the DG results.

3.2.1. Statistical model: prisoners dilemma

For the PDG, we modeled the response variable as the binomial count of cooperate (success) versus defect (failure) choices made by each participant over the six rounds of play in a single experiment. This was denoted `Num_cooperates`, and was calculated by multiplying the cooperation frequency by the number of rounds. The statistical model included fixed effects: `Participant_group`, `Partner_condition`, time step (`t`), `Model`, and `Temperature`. These effects were chosen to investigate their hypothesized influence on cooperative behavior. Additional interaction terms `Partner_condition:Model` and `Participant_group:Model` were also included, having identified these as possible interactions during our exploratory data analysis (see section S2 of supplementary information).

To account for individual variations in cooperation that are not explained by fixed effects, a random intercept for each `Participant_id` was included. This random effect captures individual-level random

[5] The full set of r-cran packages used for our analysis can be found in the code repo [108], and the following references: [112, 114–141].

Table 6. Reference levels for factors used in GLMM model for prisoner's dilemma.

Factor	Reference level
Participant_group	Control
Partner_condition	tit for tat D
Model	`gpt-3.5-turbo-0613`

variation in cooperation caused by variation in the attributes in table 2 across the different simulacra. We also attempted to fit models that included an additional random slope term to allow each participant to have varying response to `Partner_condition`, but none of these models converged.

We used the beta-binomial family to allow the probability of success, i.e. the probability the participant chooses to cooperate, to vary between cases. The logit link function was used to model the log odds of the probability of success as a linear combination of the predictors.

Our initial model included terms for temperature and t. As we expected, the estimated coefficients for these terms were not statistically significant. Moreover, since there is no theoretical reason to believe that the t term has any effect on the results, we omitted it to make the model more tractable and to help prevent over-fitting.

The final model was formulated as follows:

$$\begin{aligned}(\text{Num\_cooperates}, 6 - \text{Num\_cooperates}) \sim \text{Participant\_group} * \text{Partner\_condition} * \text{Model} \\ + (1|\text{Participant\_id})\,.\end{aligned} \tag{1}$$

The reference levels for each factor are summarized in table 6. The purpose of the above formula is to build the model that generates the PDG results shown below (the format is commonly used in the R statistical package). The terms to the left of the tilde ('∼') refer to the dependent variable (level of cooperation). The terms to the right are the independent variables. The tilde itself indicates that the left-side terms are being investigated as varying according to the right-side terms.

3.2.2. Results: prisoners dilemma

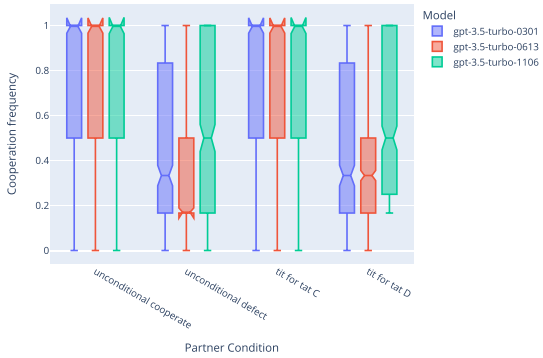
Figure 3 displays cooperation frequency across different partner conditions, participant groups, and our three GPT models (table 3). Figure 5 shows the summary of the estimated model for the PDG. While a small amount of overdispersion was reported (values slightly greater than 1), it remained well below the threshold of 2, thus it is unlikely to substantially inflate significance levels [143]. The corresponding estimates, restricted to significant coefficients only, are presented in table 7. These coefficients are expressed on an odds-ratio scale, facilitating their interpretation as effect sizes. Residual diagnostics were performed using the DHARMa package [114]. The analysis revealed no apparent patterns in the residual plots, suggesting an adequate model fit. Despite the Kolmogorov–Smirnov test rejecting the null hypothesis of normally distributed residuals, the Q–Q plot displayed satisfactory alignment, implying that the test's significance may have been influenced by the large sample size. Detailed residuals analysis is available in the supplementary code repository [108].

Figure 4 presents the main results from the PDG, illustrating the probability of cooperation predicted by the estimated GLMM across different participant groups, partner conditions, and GPT model versions. Each subplot of figure 4 shows results for a specific participant group, and within each subplot we can see the effect on cooperation probability from manipulating the partner condition for each GPT model. The corresponding pairwise effect sizes are summarized in tables 8–10 on an odds-ratio scale, along with p-values. In the discussion below we use a significance threshold of $p < 0.0001$ to account for the relatively large sample size used in our study [144, 145].

We discuss each of the subplots of figure 4 below.

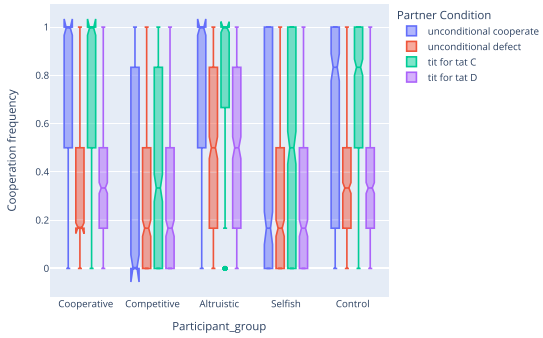
3.2.2.1. Selfish group

Contrary to our initial hypothesis H7, which posited that members of the Selfish group would invariably choose to defect, we observe a marked deviation from this expectation with simulacra instantiated using earlier GPT models (gpt-3.5-turbo-0613 and gpt-3.5-turbo-0301), which cooperate less when faced with defectors (D or T4TD) as opposed to cooperators (T4TC or C) — see second segment of table 8. In contrast, the later version of the GPT model (gpt-3.5-turbo-1106) exhibited behavior more closely aligned with our original prediction, yielding no statistically-effect of partner condition (table 10). However, all three GPT models displayed a consistent tendency to cooperate with a probability significantly greater than the competitive group (below).

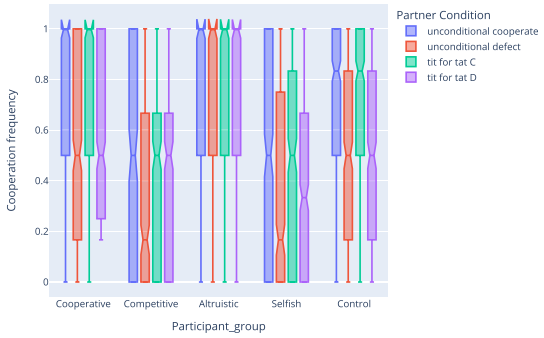


(a) All models

(b) gpt-3.5-turbo-0301



(c) gpt-3.5-turbo-0613



(d) gpt-3.5-turbo-1106

Figure 3. Box plots of cooperation frequency (y-axis) showing the interaction between model, partner condition and participant group for the Prisoners Dilemma game *only*. Figure 3(a) shows partner conditions (x-axis) across the three models (color). Figures 3(b)–(d) show participant group results (x-axis) for each GPT model (color): gpt-3.5-turbo-0301, gpt-3.5-turbo-0613, and gpt-3.5-turbo-1106, respectively (see table 3). The outlier in figure 3(c) follows the convention of 1.5 times higher/lower than third/first IQR. The figure above shows four different versions of the same results, depending on what model generated the data. As shown figure 3(a), the overall pattern of yes-to-cooperators and no-to-defectors is rather clean, but the pattern is messier in the two older models (. . . — 0301, . . . — 0613) than in the newer model (. . . — 1106). GPT-3.5 models differ primarily according to their extent of RLHF [142].

3.2.2.2. Competitive group

Simulacra in the competitive group that were instantiated with the gpt-3.5-turbo-0301 and gpt-3.5-turbo-1106 models exhibited behavior that aligned closely with our original hypothesis H10; in particular, there was no statistically-significant effect of partner condition for these models (third segment of tables 8 and 10). The propensity to cooperate was slightly higher than we predicted, but still significantly below 0.4. In contrast the gpt-3.5-turbo-0613 model cooperates less when faced with D versus T4TC, T4TD versus T4TC, or T4TC versus C partners (see third segment of table 9).

3.2.2.3. Cooperative group

As we predicted (H8), simulacra in the cooperative group increased their propensity to cooperate in line with their partner's cooperative stance, with `T4TC` partners eliciting a statistically-significant increase in cooperation as compared with `T4TD` partners, and with tit-for-tat partners as compared with unconditional defectors. These effects are statistically-significant (fourth segment of tables 8–10). However, cooperative simulacra were more forgiving of unconditional defectors than we anticipated, with no statistically-significant effect between the `D` and `T4TD` partner conditions (rejecting H9). These findings are robust across all three GPT models.

3.2.2.4. Altruistic group

Simulacra instantiated with earlier GPT models showed a statistically-significant decrease in cooperation when faced with uncooperative partners (partially rejecting H6). However, those instantiated with the later gpt-3.5-turbo-1106 model show high levels of cooperation ($\geqslant$0.75) irrespective of whether facing

Table 7. Model estimates for prisoners dilemma. These are shown for significant coefficients only ($p < 0.05$) on an odds ratio scale rounded to 2 decimal places.

	Odds ratio	Std. Err.	z	Pr($>|z|$)
X.Intercept.	0.62	1.13	−3.88	0.00
Participant_groupAltruistic	1.53	1.19	2.47	0.01
Participant_groupCompetitive	0.44	1.19	−4.67	0.00
Participant_groupSelfish	0.49	1.19	−4.11	0.00
Modelgpt.3.5.turbo.0301	2.15	1.11	7.55	0.00
Modelgpt.3.5.turbo.1106	1.41	1.12	3.13	0.00
Participant_groupControl.Partner_conditionT4TC	3.65	1.10	14.22	0.00
Participant_groupAltruistic.Partner_conditionT4TC	6.19	1.11	17.69	0.00
Participant_groupCompetitive.Partner_conditionT4TC	2.22	1.10	8.10	0.00
Participant_groupCooperative.Partner_conditionT4TC	6.82	1.10	19.81	0.00
Participant_groupSelfish.Partner_conditionT4TC	2.79	1.10	10.71	0.00
Participant_groupControl.Partner_conditionC	3.18	1.10	12.64	0.00
Participant_groupAltruistic.Partner_conditionC	5.18	1.11	16.00	0.00
Participant_groupCooperative.Partner_conditionC	6.33	1.10	19.00	0.00
Participant_groupSelfish.Partner_conditionC	1.78	1.10	5.90	0.00
Participant_groupControl.Modelgpt.3.5.turbo.0301.Partner_conditionD	0.60	1.15	−3.81	0.00
Participant_groupAltruistic.Modelgpt.3.5.turbo.0301.Partner_conditionD	0.52	1.15	−4.83	0.00
Participant_groupCompetitive.Modelgpt.3.5.turbo.0301.Partner_conditionD	0.61	1.15	−3.48	0.00
Participant_groupCooperative.Modelgpt.3.5.turbo.0301.Partner_conditionD	0.56	1.14	−4.39	0.00
Participant_groupSelfish.Modelgpt.3.5.turbo.0301.Partner_conditionD	0.72	1.15	−2.42	0.02
Participant_groupAltruistic.Modelgpt.3.5.turbo.1106.Partner_conditionD	3.33	1.17	7.78	0.00
Participant_groupCooperative.Modelgpt.3.5.turbo.1106.Partner_conditionD	1.89	1.15	4.47	0.00
Participant_groupControl.Modelgpt.3.5.turbo.0301.Partner_conditionT4TD	0.59	1.14	−3.92	0.00
Participant_groupAltruistic.Modelgpt.3.5.turbo.0301.Partner_conditionT4TD	0.56	1.15	−4.19	0.00
Participant_groupCompetitive.Modelgpt.3.5.turbo.0301.Partner_conditionT4TD	0.67	1.15	−2.90	0.00
Participant_groupCooperative.Modelgpt.3.5.turbo.0301.Partner_conditionT4TD	0.55	1.14	−4.46	0.00
Participant_groupSelfish.Modelgpt.3.5.turbo.0301.Partner_conditionT4TD	0.71	1.15	−2.54	0.01
Participant_groupAltruistic.Modelgpt.3.5.turbo.1106.Partner_conditionT4TD	2.61	1.16	6.41	0.00
Participant_groupCooperative.Modelgpt.3.5.turbo.1106.Partner_conditionT4TD	1.90	1.15	4.54	0.00
Participant_groupControl.Modelgpt.3.5.turbo.0301.Partner_conditionT4TC	0.51	1.15	−4.68	0.00
Participant_groupAltruistic.Modelgpt.3.5.turbo.0301.Partner_conditionT4TC	0.18	1.16	−11.60	0.00
Participant_groupCompetitive.Modelgpt.3.5.turbo.0301.Partner_conditionT4TC	0.32	1.15	−7.92	0.00
Participant_groupCooperative.Modelgpt.3.5.turbo.0301.Partner_conditionT4TC	0.31	1.16	−8.09	0.00
Participant_groupSelfish.Modelgpt.3.5.turbo.0301.Partner_conditionT4TC	0.74	1.15	−2.16	0.03
Participant_groupCompetitive.Modelgpt.3.5.turbo.1106.Partner_conditionT4TC	0.59	1.17	−3.42	0.00
Participant_groupSelfish.Modelgpt.3.5.turbo.1106.Partner_conditionT4TC	0.62	1.16	−3.23	0.00
Participant_groupControl.Modelgpt.3.5.turbo.0301.Partner_conditionC	0.66	1.16	−2.89	0.00
Participant_groupAltruistic.Modelgpt.3.5.turbo.0301.Partner_conditionC	0.22	1.16	−10.42	0.00
Participant_groupCompetitive.Modelgpt.3.5.turbo.0301.Partner_conditionC	0.48	1.16	−4.94	0.00
Participant_groupCooperative.Modelgpt.3.5.turbo.0301.Partner_conditionC	0.36	1.16	−6.93	0.00

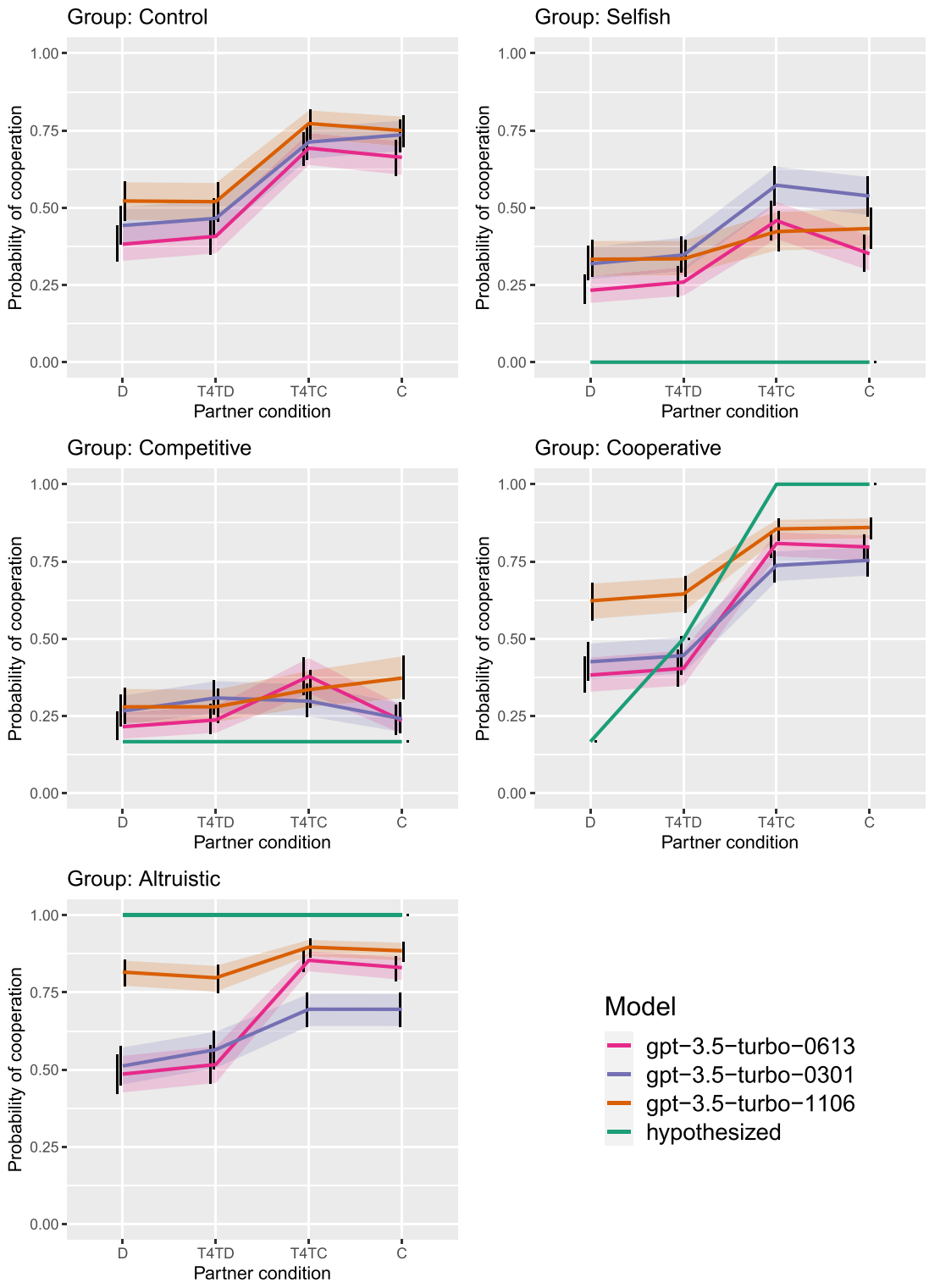


Figure 4. Prisoners dilemma—probability of cooperation for each participant group by partner condition. These plots show the probability of cooperation predicted by the estimated GLMM model for each combination of partner condition and GPT model. The error bars show 95% confidence intervals. The green line illustrates the level of cooperation that we hypothesized prior to conducting the experiment[6]. See section 2.4 for an explanation of D, TFTD, T4TC, and C.

cooperative or uncooperative partners; although there is sometimes a statistically-significant effect of partner condition, the effect sizes are much smaller compared with the earlier models (odds ratios closed to 1 in final row of table 10 compared with tables 8 and 9), and there is no statistically-significant effect when switching

[6] The green lines are purely conceptual and conjectural. Looking at figure 4, the prediction for the altruistic, selfish, and competitive roles are straightforward: ceiling scores for the former, and floor scores for the two latter. The control group has no line because we did not have

```
Family: betabinomial  ( logit )
Formula:
cbind(Num_cooperates, 6 - Num_cooperates) ~ Participant_group *
    Partner_condition * Model + Temperature + (1 | Participant_id)
Data: results_pd

     AIC      BIC   logLik deviance df.resid
 92968.3  93489.9 -46421.2  92842.3    29083

Random effects:

Conditional model:
 Groups         Name        Variance Std.Dev.
 Participant_id (Intercept) 0.975    0.9874
Number of obs: 29146, groups:  Participant_id, 450

Dispersion parameter for betabinomial family (): 1.53
```

Figure 5. Model summary for prisoners dilemma generating using R's glmmTMB package.

from unconditionally defecting versus unconditionally cooperating partners. This is more in line with our original hypothesis albeit with the small effect of some partner condition pairs (partially supporting H6).

3.2.2.5. Control group

For all three models, the results from the control group are qualitatively very similar to the cooperative group (rejecting H1), with a similar pattern of effects from partner condition, but slightly smaller effect sizes from cooperative versus defector partner conditions as compared with the cooperative group.

3.2.3. Statistical model: DG

For our analysis of the DG results, we initially attempted to use a GLMM model similar to the one used for the PDG analysis above. However, our initial results yielded a very high amount of overdispersion, and an examination of the histogram of the response variable showed that the data was dominated by two out of the five possible choices: donating nothing, or donating two dollars (see figure 6), indicating that the response variable was not Binomial/Poisson-distributed. Therefore we used an ordinal regression in the form of a CLMM which was estimated using the `clmm` function in the R package `ordinal` [115]. The model was similar in the structure to the PDG analysis, with a random intercept for each participant, and with fixed effects: `Participant_group`, `Model`, `t`, `Temperature` and an interaction term for `Participant_group` and `Model`. Since this experiment was a one-shot interaction, there was no partner condition variable.

As with the PDG, the estimates for the `Temperature` and t variables were not significant, and the latter was omitted from the final model (t was omitted and temperature stayed). The formula for the final model is given below.

$$\texttt{Response} \sim \texttt{Participant\_group} * \texttt{Model} + \texttt{Temperature} + (1|\texttt{Participant\_id})\,. \tag{2}$$

The purpose of the above formula is to build the model that generates the PDG results shown in figure 7 (for an explanation of the format, see caption under formula (1)). Here the dependent variable `Response` is one of the five possible integer choices presented to the simulacrum, and represents the total amount donated by the subject to its partner; because the data were dominated by donations of 0 or 2, instead of treating this is an integer, we modeled the choice of donation as an ordinal variable. Subsequently we re-interpret the dependent variable as an integer, using the probabilities predicted by the CLMM for each choice to form an overall expected donation amount as a weighted mean, which allowed us to compare results against the PDG on the same scale. The estimates and model fit are shown in table 11.

3.2.4. Results: DG

Figure 7 (left hand-side) shows the expected level of donation predicted by the cumulative link mixed-model for each participant group in the DG experiment. Simulacra instantiated with all three models respond

an expectation. The cooperative role is more complex, being a calculation based on six rounds (see figure S1) and with an expectation that the cooperative role will utilize a tit-for-tat strategy. Hence, for D, the cooperator cooperates in the first round and then implements the 'grim trigger' strategy [175] by never cooperating again after the first defection. The result is cooperating on only the first 1/6th, or 0.2, of rounds. For the T4TD, the cooperator is at 0.5 because it continues cooperating to encourage the defector to start cooperating.

Table 8. Contrasts for prisoners dilemma experiment with the gpt-3.5-turbo-0301 model.

Partner_condition_pairwise	odds.ratio	SE	df	null	*z*.ratio	*p*.value
Model = gpt-3.5-turbo-0301, Participant_group = Control						
D / T4TD	0.9110	0.0854	Inf	1.0000	−0.994	0.7531
D / T4TC	0.3202	0.0320	Inf	1.0000	−11.398	<.0001
D / C	0.2846	0.0288	Inf	1.0000	−12.401	<.0001
T4TD / T4TC	0.3514	0.0349	Inf	1.0000	−10.520	<.0001
T4TD / C	0.3124	0.0315	Inf	1.0000	−11.534	<.0001
T4TC / C	0.8889	0.0946	Inf	1.0000	−1.107	0.6852
Model = gpt-3.5-turbo-0301, Participant_group = Selfish						
D / T4TD	0.8802	0.0813	Inf	1.0000	−1.381	0.5110
D / T4TC	0.3486	0.0339	Inf	1.0000	−10.827	<.0001
D / C	0.4014	0.0390	Inf	1.0000	−9.387	<.0001
T4TD / T4TC	0.3961	0.0387	Inf	1.0000	−9.474	<.0001
T4TD / C	0.4561	0.0445	Inf	1.0000	−8.039	<.0001
T4TC / C	1.1514	0.1170	Inf	1.0000	1.387	0.5076
Model = gpt-3.5-turbo-0301, Participant_group = Competitive						
D / T4TD	0.8123	0.0768	Inf	1.0000	−2.200	0.1233
D / T4TC	0.8534	0.0837	Inf	1.0000	−1.616	0.3696
D / C	1.1412	0.1150	Inf	1.0000	1.310	0.5562
T4TD / T4TC	1.0506	0.1032	Inf	1.0000	0.503	0.9585
T4TD / C	1.4050	0.1417	Inf	1.0000	3.370	0.0042
T4TC / C	1.3373	0.1391	Inf	1.0000	2.794	0.0267
Model = gpt-3.5-turbo-0301, Participant_group = Cooperative						
D / T4TD	0.9221	0.0816	Inf	1.0000	−0.916	0.7964
D / T4TC	0.2640	0.0251	Inf	1.0000	−13.983	<.0001
D / C	0.2417	0.0233	Inf	1.0000	−14.721	<.0001
T4TD / T4TC	0.2863	0.0273	Inf	1.0000	−13.109	<.0001
T4TD / C	0.2621	0.0253	Inf	1.0000	−13.854	<.0001
T4TC / C	0.9156	0.0937	Inf	1.0000	−0.862	0.8245
Model = gpt-3.5-turbo-0301, Participant_group = Altruistic						
D / T4TD	0.8087	0.0753	Inf	1.0000	−2.280	0.1028
D / T4TC	0.4591	0.0440	Inf	1.0000	−8.115	<.0001
D / C	0.4596	0.0439	Inf	1.0000	−8.144	<.0001
T4TD / T4TC	0.5678	0.0542	Inf	1.0000	−5.934	<.0001
T4TD / C	0.5684	0.0540	Inf	1.0000	−5.952	<.0001
T4TC / C	1.0011	0.0975	Inf	1.0000	0.012	1.0000

Results are averaged over the levels of: Temperature
P value adjustment: Tukey method for comparing a family of 4 estimates
Tests are performed on the log odds ratio scale

similarly to changes in the role prompt, with approximately equal effect sizes and significance levels (see table 12). As predicted (hypothesis H3), both selfish and competitive simulacra consistently offered nothing to their partner in the DG, whereas cooperative simulacra do. Level of donation was highest in the cooperative and altruistic groups, but contrary to our original hypothesis H2, there was no statistically-significant increase moving from the cooperative to the altruistic group (final row of each section in table 12). The control group exhibited donations intermediate between selfish/competitive and cooperative/altruistic. The overall level of cooperation was significantly lower than the PDG (right hand-side), confirming hypothesis H4.

4. Discussion

The idea behind machine psychology is to test LLM-chatbots (such as GPT) as if they were human participants in psychology experiments [36, 64–73]. In our study, we used paradigms from behavioral economics—the DG [79, 80] and the repeated PDG [81–83]—and tested the equivalent of 450 human

Table 9. Contrasts for prisoners dilemma experiment with the gpt-3.5-turbo-0613 model.

Partner_condition_pairwise	odds.ratio	SE	df	null	*z*.ratio	*p*.value
Model = gpt-3.5-turbo-0613, Participant_group = Control						
D / T4TD	0.8989	0.0757	Inf	1.0000	−1.266	0.5847
D / T4TC	0.2742	0.0250	Inf	1.0000	−14.216	<.0001
D / C	0.3140	0.0288	Inf	1.0000	−12.638	<.0001
T4TD / T4TC	0.3051	0.0276	Inf	1.0000	−13.113	<.0001
T4TD / C	0.3494	0.0319	Inf	1.0000	−11.533	<.0001
T4TC / C	1.1451	0.1110	Inf	1.0000	1.398	0.5005
Model = gpt-3.5-turbo-0613, Participant_group = Selfish						
D / T4TD	0.8672	0.0817	Inf	1.0000	−1.513	0.4298
D / T4TC	0.3588	0.0343	Inf	1.0000	−10.709	<.0001
D / C	0.5605	0.0550	Inf	1.0000	−5.901	<.0001
T4TD / T4TC	0.4137	0.0392	Inf	1.0000	−9.324	<.0001
T4TD / C	0.6463	0.0627	Inf	1.0000	−4.496	<.0001
T4TC / C	1.5623	0.1535	Inf	1.0000	4.540	<.0001
Model = gpt-3.5-turbo-0613, Participant_group = Competitive						
D / T4TD	0.8826	0.0867	Inf	1.0000	−1.272	0.5811
D / T4TC	0.4507	0.0443	Inf	1.0000	−8.103	<.0001
D / C	0.8928	0.0934	Inf	1.0000	−1.083	0.6998
T4TD / T4TC	0.5107	0.0497	Inf	1.0000	−6.900	<.0001
T4TD / C	1.0115	0.1050	Inf	1.0000	0.110	0.9995
T4TC / C	1.9808	0.2054	Inf	1.0000	6.592	<.0001
Model = gpt-3.5-turbo-0613, Participant_group = Cooperative						
D / T4TD	0.9140	0.0782	Inf	1.0000	−1.051	0.7193
D / T4TC	0.1467	0.0142	Inf	1.0000	−19.809	<.0001
D / C	0.1579	0.0153	Inf	1.0000	−18.998	<.0001
T4TD / T4TC	0.1604	0.0155	Inf	1.0000	−18.924	<.0001
T4TD / C	0.1728	0.0168	Inf	1.0000	−18.110	<.0001
T4TC / C	1.0769	0.1142	Inf	1.0000	0.698	0.8977
Model = gpt-3.5-turbo-0613, Participant_group = Altruistic						
D / T4TD	0.8835	0.0795	Inf	1.0000	−1.377	0.5141
D / T4TC	0.1615	0.0166	Inf	1.0000	−17.687	<.0001
D / C	0.1931	0.0199	Inf	1.0000	−15.994	<.0001
T4TD / T4TC	0.1828	0.0189	Inf	1.0000	−16.465	<.0001
T4TD / C	0.2185	0.0225	Inf	1.0000	−14.766	<.0001
T4TC / C	1.1956	0.1358	Inf	1.0000	1.573	0.3940

Results are averaged over the levels of: Temperature

P value adjustment: Tukey method for comparing a family of 4 estimates

Tests are performed on the log odds ratio scale

participants in five participant groups. As shown in table 1 earlier, each group represented one of five categories of simulacra: (1) cooperative, (2) competitive, (3) altruistic, (4) selfish, and (5) control. In our study, we found that the 'participants' did what we expected them to do. The results are all in the numbers. Cooperative and altruistic simulacra exhibited higher cooperation than competitive and selfish simulacra. There was a higher level of cooperation in the repeated game (PDG) than in the one-shot game (DG). Evidence for altruism was mixed; the later version of the model (gpt-3.5-turbo-1106) showed consistently high levels of cooperation in the PDG even when faced with uncooperative partners, but this was not the case for the earlier models (gpt-3.5-turbo-613; gpt-3.5-turbo-0301), and in the DG there was no statistically-significant difference between the altruistic and cooperative simulacra. Cooperative simulacra showed strong signs of conditional reciprocity, but they were more forgiving of unconditional defectors than we anticipated. Our control group with neutral prompts showed behavior very similar to the cooperative group, suggesting that conditional-reciprocity may be the 'default' behavior of GPT models for tasks resembling social dilemmas [92]. Overall, our prompts were successful in instantiating our categories of simulacra.

Table 10. Contrasts for prisoners dilemma experiment with the gpt-3.5-turbo-1106 model.

Partner_condition_pairwise	odds.ratio	SE	df	null	*z*.ratio	*p*.value
Model = gpt-3.5-turbo-1106, Participant_group = Control						
D / T4TD	1.0089	0.0976	Inf	1.0000	0.091	0.9997
D / T4TC	0.3211	0.0339	Inf	1.0000	−10.760	<.0001
D / C	0.3639	0.0386	Inf	1.0000	−9.539	<.0001
T4TD / T4TC	0.3182	0.0334	Inf	1.0000	−10.903	<.0001
T4TD / C	0.3607	0.0380	Inf	1.0000	−9.676	<.0001
T4TC / C	1.1334	0.1282	Inf	1.0000	1.107	0.6851
Model = gpt-3.5-turbo-1106, Participant_group = Selfish						
D / T4TD	0.9932	0.1067	Inf	1.0000	−0.063	0.9999
D / T4TC	0.6823	0.0753	Inf	1.0000	−3.464	0.0030
D / C	0.6558	0.0743	Inf	1.0000	−3.723	0.0011
T4TD / T4TC	0.6870	0.0742	Inf	1.0000	−3.478	0.0028
T4TD / C	0.6603	0.0732	Inf	1.0000	−3.742	0.0010
T4TC / C	0.9612	0.1093	Inf	1.0000	−0.348	0.9855
Model = gpt-3.5-turbo-1106, Participant_group = Competitive						
D / T4TD	0.9991	0.1221	Inf	1.0000	−0.007	1.0000
D / T4TC	0.7688	0.0955	Inf	1.0000	−2.117	0.1477
D / C	0.6530	0.0903	Inf	1.0000	−3.081	0.0111
T4TD / T4TC	0.7695	0.0906	Inf	1.0000	−2.225	0.1165
T4TD / C	0.6536	0.0865	Inf	1.0000	−3.211	0.0072
T4TC / C	0.8494	0.1140	Inf	1.0000	−1.216	0.6168
Model = gpt-3.5-turbo-1106, Participant_group = Cooperative						
D / T4TD	0.9082	0.0865	Inf	1.0000	−1.011	0.7427
D / T4TC	0.2790	0.0300	Inf	1.0000	−11.862	<.0001
D / C	0.2682	0.0293	Inf	1.0000	−12.045	<.0001
T4TD / T4TC	0.3072	0.0327	Inf	1.0000	−11.077	<.0001
T4TD / C	0.2953	0.0319	Inf	1.0000	−11.274	<.0001
T4TC / C	0.9614	0.1143	Inf	1.0000	−0.331	0.9875
Model = gpt-3.5-turbo-1106, Participant_group = Altruistic						
D / T4TD	1.1258	0.1337	Inf	1.0000	0.998	0.7508
D / T4TC	0.5100	0.0673	Inf	1.0000	−5.104	<.0001
D / C	0.5760	0.0760	Inf	1.0000	−4.183	0.0002
T4TD / T4TC	0.4531	0.0574	Inf	1.0000	−6.252	<.0001
T4TD / C	0.5116	0.0648	Inf	1.0000	−5.294	<.0001
T4TC / C	1.1293	0.1569	Inf	1.0000	0.875	0.8178

Results are averaged over the levels of: Temperature

P value adjustment: Tukey method for comparing a family of 4 estimates

Tests are performed on the log odds ratio scale

In both games, simulacra roughly showed a pattern, on a scale from high cooperation to low cooperation of *altruistic* $\geqslant$ *cooperative* $>$ *control* $>$ *selfish* $\geqslant$ *competitive*. As shown in table 4, we formulated ten hypotheses. In our results, three hypotheses were supported (H3, H4, H8), two were partially supported (H1, H2, H6, H10), and three were not supported (H5, H7, H9). We review our original hypotheses in detail below.

The hypotheses that were clearly supported were H3, H4, and H8. These showed that cooperative simulacra showed a higher frequency of cooperation than selfish or competitive simulacra in the one-shot game, that cooperation was higher in repeated games, and that in repeated games with tit-for-tat partners, cooperative simulacra showed a higher frequency of cooperation when faced with partners who cooperated on the first move.

The hypotheses that were rejected were H5, H7, and H9 (or, as a psychologist might say, we failed to reject the null hypotheses, [146]). The first rejection (H5) showed that there were differences between the three models of gpt-3.5-turbo (see figure 3). The major difference between these models (gpt-3.5-turbo-0301, ...-0613 and ...-1106) was in the extent of reinforcement-learning from human feedback (RLHF) fine-tuning

Table 11. Fitted model for dictator.

	Dictator game CLMM model
Participant_groupSelfish	−2.53 (0.31)***
Participant_groupCompetitive	−3.14 (0.34)***
Participant_groupCooperative	2.12 (0.24)***
Participant_groupAltruistic	2.73 (0.24)***
Modelgpt-3.5-turbo-0301	0.34 (0.16)*
Modelgpt-3.5-turbo-1106	0.23 (0.14)
Temperature	0.14 (0.10)
Participant_groupSelfish:Modelgpt-3.5-turbo-0301	−0.05 (0.30)
Participant_groupCompetitive:Modelgpt-3.5-turbo-0301	−0.02 (0.34)
Participant_groupCooperative:Modelgpt-3.5-turbo-0301	0.05 (0.19)
Participant_groupAltruistic:Modelgpt-3.5-turbo-0301	−0.32 (0.19)
Participant_groupSelfish:Modelgpt-3.5-turbo-1106	−0.72 (0.32)*
Participant_groupCompetitive:Modelgpt-3.5-turbo-1106	−1.09 (0.42)**
Participant_groupCooperative:Modelgpt-3.5-turbo-1106	−0.48 (0.18)**
Participant_groupAltruistic:Modelgpt-3.5-turbo-1106	−0.82 (0.18)***
threshold.1	0.75 (0.18)***
spacing	1.58 (0.03)***
Log Likelihood	−6685.28
AIC	13406.56
BIC	13532.02
Num. obs.	7865
Groups (Participant_id)	450
Variance: Participant_id: (Intercept)	1.79

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

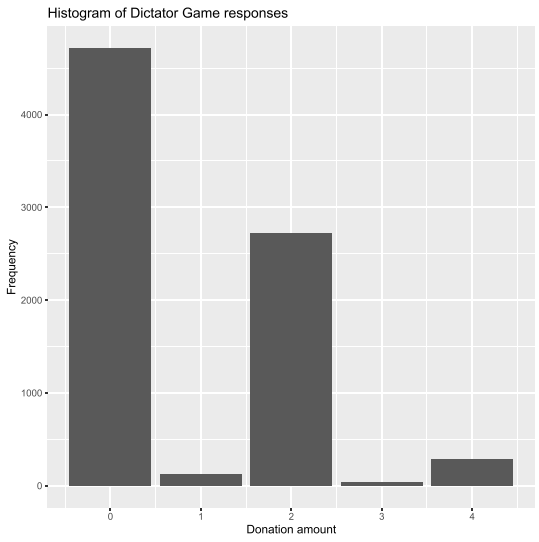


Figure 6. Histogram of the dependent variable for the dictator game.

performed upon them ([142]): later models are better aligned to filter out objectionable content. The second rejection (H7) showed that selfish simulacra did not completely fail to cooperate in repeated games (they showed some modest tendency to cooperate in all partner conditions, and moreover at frequencies *higher* than competitive simulacra). Finally, H9 was rejected because cooperative simulacra did not cooperate more in response to tit-for-tat-simulacra who defected on the first move as compared with unconditional defectors.

The partially supported hypotheses were H1, H2, H6 and H10. H1 was only partially supported because the control group showed very similar, albeit not identical, behavior to the cooperative group. H2 was only partially supported because of differences between the one-shot and repeated game; in the one-shot game there was no statistically-significant difference between the cooperative and altruistic groups in contrast to the repeated-game where altruists cooperated the most. H6 was only partially supported because of differences in models; altruistic simulacra did play indiscriminately in repeated games, but only in with the later GPT model. With earlier models, they cooperated less with defectors. H10 was only partially supported

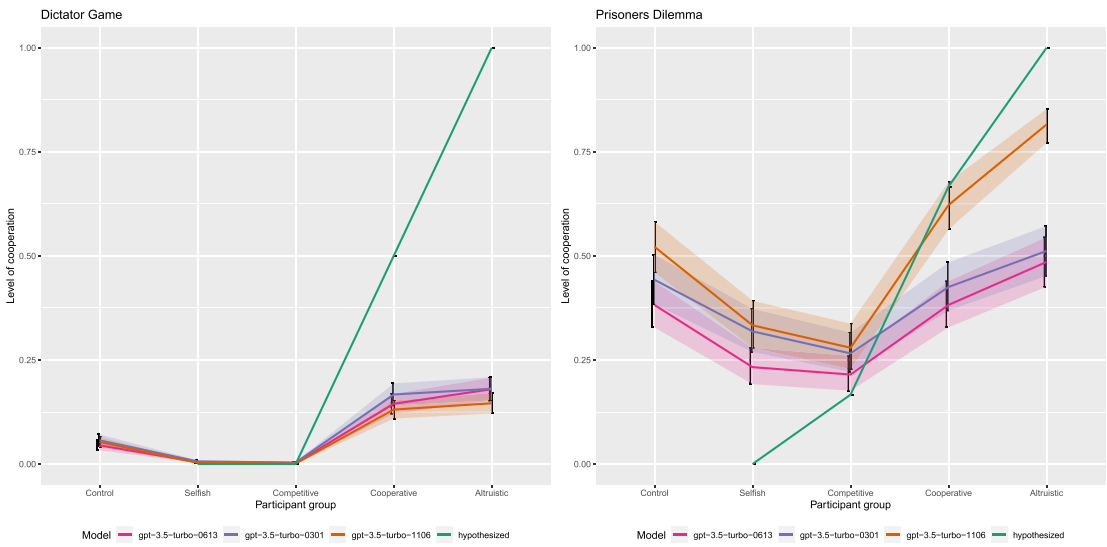


Figure 7. Probability of cooperation for each participant group (both experiments). These plots show the probability of cooperation/donation predicted by the relevant estimated mixed-model for each participant group and for each GPT model. The results from the dictator game (Prisoners Dilemma) are shown on the left (right) for comparison. The error bars show 95% confidence intervals. The green lines are conjectural and illustrates the level of cooperation that we hypothesized prior to conducting the experiment.

in that only with two out of the three GPT models did competitive simulacra exhibit low levels of cooperation irrespective of partner condition, and moreover for all models they cooperated *less* than selfish simulacra.

For the one-shot DG, the GPT models were able to consistently operationalise natural language descriptions of cooperative attitudes, and produce narratives of game play that broadly fell within an experimental psychologist's expectation of how the corresponding simulacra should behave. However, this was with the exception that altruistic simulacra did not exhibit a statistically-significant increase in donations as compared with cooperative simulacra. The results for the repeated PDG were more mixed. Our results suggest that, overall, from a behavioral perspective, GPT models exhibit a good 'understanding' of the task environment and of concepts such as altruism and selfishness, and that this understanding can be improved using RLHF, as evidenced by the fact that the later model exhibits a better operational understanding of altruism as compared to earlier models as compared to earlier models.

Our research set out to test whether GPT models are able to operationalise natural-language descriptions of altruistic, cooperative, competitive and selfish behavior by producing a text narrative describing simulated behavior in different task environments. We sampled simulacra from the large space of possible simulacra, and used statistical methods to make inferences about the general population of simulacra based on our finite experiments. In our analysis, we focus on the simulacra, not the attributes. The purpose of the attributes were to randomize some aspects of the initial prompt (table 2), while systematically manipulating the part of the prompt that described altruistic, cooperative, competitive and selfish attitudes, in order to estimate the effect on our dependent variable, the level of cooperation, which was either how often the simulacra made a donation (cooperate) as opposed to making no donation (defect), or what fraction of its total endowment was donated. Why did we choose those particular simulacra (whether cunning, devoted, shrewd, helpful, diplomatic, or whatever)? When we said we sampled the large space of possible simulacra—'possible' meant that there are countless combinations of descriptors we could have used, each with different nuances (cf. [147])[7]. What we did was create our simulacra intuitively (without a formal system in place). We were motivated to gauge how GPT models react to social dilemmas [97] (this has important implications for the safety of these systems if they are deployed as agents). Despite our efforts to create workable role prompts, there are studies that have shown human-like behavior in economic games without needing to use such specific role-prompting as we did. Johnson and Obradovich, for example, in their 2023 study [92], generated convincing results in their DG played by LLMs; and their prompts were more generic than ours (they did not

[7] The idea of choosing the right nuance is actually quite important. One of our reviewers suggested we investigate the possible confounding influence of the word 'investment'. Therefore, we conducted a small additional study to compare the effect of 'investment' versus 'donation' (versus using no word at all). We did find some modest significant effects and we report this mini-study in section S4 in supplementary information.

Table 12. Contrasts for the dictator game experiment.

Participant_group_pairwise	estimate	SE	df	*z*.ratio	*p*.value
Model = gpt-3.5-turbo-0613					
Control—Selfish	2.5311	0.3057	Inf	8.280	<.0001
Control—Competitive	3.1393	0.3382	Inf	9.283	<.0001
Control—Cooperative	−2.1182	0.2421	Inf	−8.749	<.0001
Control—Altruistic	−2.7302	0.2433	Inf	−11.220	<.0001
Selfish—Competitive	0.6081	0.3733	Inf	1.629	0.4788
Selfish—Cooperative	−4.6494	0.2987	Inf	−15.568	<.0001
Selfish—Altruistic	−5.2613	0.3000	Inf	−17.538	<.0001
Competitive—Cooperative	−5.2575	0.3321	Inf	−15.832	<.0001
Competitive—Altruistic	−5.8695	0.3333	Inf	−17.608	<.0001
Cooperative—Altruistic	−0.6119	0.2283	Inf	−2.680	0.0569
Model = gpt-3.5-turbo-0301					
Control—Selfish	2.5861	0.2998	Inf	8.626	<.0001
Control—Competitive	3.1619	0.3211	Inf	9.846	<.0001
Control—Cooperative	−2.1703	0.2447	Inf	−8.871	<.0001
Control—Altruistic	−2.4066	0.2461	Inf	−9.778	<.0001
Selfish—Competitive	0.5758	0.3492	Inf	1.649	0.4659
Selfish—Cooperative	−4.7564	0.2891	Inf	−16.454	<.0001
Selfish—Altruistic	−4.9927	0.2904	Inf	−17.194	<.0001
Competitive—Cooperative	−5.3322	0.3123	Inf	−17.073	<.0001
Competitive—Altruistic	−5.5685	0.3135	Inf	−17.763	<.0001
Cooperative—Altruistic	−0.2363	0.2259	Inf	−1.046	0.8338
Model = gpt-3.5-turbo-1106					
Control—Selfish	3.2518	0.3280	Inf	9.915	<.0001
Control—Competitive	4.2279	0.4046	Inf	10.451	<.0001
Control—Cooperative	−1.6403	0.2380	Inf	−6.892	<.0001
Control—Altruistic	−1.9110	0.2387	Inf	−8.005	<.0001
Selfish—Competitive	0.9761	0.4536	Inf	2.152	0.1984
Selfish—Cooperative	−4.8921	0.3241	Inf	−15.094	<.0001
Selfish—Altruistic	−5.1628	0.3247	Inf	−15.898	<.0001
Competitive—Cooperative	−5.8682	0.4019	Inf	−14.603	<.0001
Competitive—Altruistic	−6.1389	0.4024	Inf	−15.257	<.0001
Cooperative—Altruistic	−0.2707	0.2288	Inf	−1.183	0.7610

Results are averaged over the levels of: Temperature

P value adjustment: Tukey method for comparing a family of 5 estimates

assign differentiated roles). As mentioned earlier, this suggests that LLMs already have some default propensity to mimic human behavior (the corpus having taught them to do that).

One important issue to mention is that our LLM-version of 'behavioural economics' is lacking a central feature found in human studies [148]: our 'players' are not getting paid. Although some authors argue that monetary incentives for humans are not always necessary [149], and unpaid human studies do get published (e.g. [150])—some studies do show a large discrepancy between a participant's self-reported prosocial intentions and what they actually do in real life (e.g. [151]; cf. [60]). Johnson and Obradovich [93], in their 2024 study, conducted an interesting experiment using the trust game [152] on GPT 3.5 (text-davinci-003 [107]) when interacting with a human experimenter. Here, they compared results between two conditions: (1) trust game without incentives (hypothetical condition), and (2) trust game *with* incentives (non-hypothetical condition). Their incentives were in the form of tokens: the human user paid with his own money, and GPT agents were 'paid' through a donation to OpenAI. In their results, they found that the GPT model showed more 'trusting' behavior in the hypothetical game compared to the non-hypothetical game. This is worth bearing in mind vis-à-vis our own results, which (in their terms) would be considered 'hypothetical' (because we did not offer real-life monetary incentives to GPT 'participants').

Issues of incentives aside, one important goal is to compare our simulacra results to human results. Starting with our one-shot results (DG), we find that the overall generosity of our simulacra players roughly match those seen in human players (e.g. compare our table 11 to Engel's table 2 from [79]). However, there are at least two issues that make our human-machine comparison difficult. The first is that multiple intervening variables can alter the level of donation [79, 153] (indeed, that is what makes the DG highly

useful in general as a methodological tool: its amenability as a dependent variable in response to a manipulation of interest). The second issue is how we would approximate the various 'personalities' that were instantiated in our role prompts. Can we define a human dispositionally, as purely (100%) altruistic, cooperative, control, competitive, or selfish? Real human participants in a laboratory setting are not categorizable in such simple ways. But, there is good evidence that personality factors do play a role in the decisions that humans make in real-life economic games [154]. It could be of interest to design studies that prompt humans in ways that temporarily alter their game-playing behaviors (i.e. with LLM-style prompting). For example, Tan and Forgas [153] induced either happy or sad mood in their human DG players, finding a number of differences in game play as a result of the players' moods. Mood induction is not the same as the instantiation of role prompts, but the Tan and Forgas [153] study is an example of priming participants to be a slightly different version of themselves. Pertinently, OpenAI has developed an option called 'steerability' [35], where users will be able to customize the 'personality' and task orientation of the chatbot to some new extent (which represents an extent of control that presumably goes beyond the reach of normal prompt engineering, even if there is no direct control of the corpus, e.g. [53]). Steerability is similar to the concept of the 'system prompt', which, in contrast to the 'user prompt', is like an additional set of prompts that precedes a chat session (i.e. the system prompt is already in place when the chat session starts) [102]: but there remain a number of questions about how efficacious the system prompt can be in relation to the user prompt (as mentioned earlier, we did not use the system prompt in our own study).

Now, turning to our repeated game, the PDG, the comparison between our simulacra results and results from human studies is tricky for the same reason as the one-shot game (i.e. multiple intervening variables; no exact equivalent to role prompts). Generally, the level of generosity from the simulacra in our PDG are roughly the same level as that found in human studies (e.g. compare our figure 4 to those in [155])—that level being a probability of cooperation that typically sits around the middle to lower middle of the scale, not usually at ceiling or floor scores. That said, it is important to reiterate the importance of the multiple intervening variables. The level of generosity is weightily influenced by factors such as the perceived probability of the game ending, degree of risk, the possibility of equilibrium and trust, the possibility of punishment and so forth [155]. In our study, the PDG had fewer rounds than in the typical human PDG (cf. [83, 154]). Thinking about our simulacra, we should ponder a result from Dal Bó and Fréchette [155] who conducted an extensive review of repeated PDGs: they found that, in humans, strategic concerns had a reliably greater effect on game play than personality variables such as altruism. The only way to develop a truly cooperative AI agent would be to design one which is capable to understanding the perspective (wants, needs, etc) of other agents [57, 156]. It is difficult to conclude that the appearance (or illusion) of moral action from an LLM can be anything more than morally-blind autoregression: the appearance of having a 'moral core' (see [157]) only as a result of some statistical sleight-of-hand [41, 43, 44]. That said, it is perfectly acceptable to run studies such as ours without needing to probe the 'mind' of GPT. As Johnson and Obradovich [93] wrote when reporting their trust game, '…we are concerned with trust-like behavior, not whether an AI model possesses a conceptualization of trust' (p 2).

However, it may be enlightening to tease apart the distinction between the *model* and the *simulacrum*. In our study, a 'selfless philanthropist' was likely to cooperate, a 'ruthless equities trader' was likely to defect, etc (because, the corpus provided those expectations to the LLM during training). These 'selfless' and 'ruthless' characters are merely simulacra, but for a brief period, they are 'real' within the chat window. Perhaps an LLM itself cannot have inherent motivations and wants, but a simulacrum can. This prompts a germane question: what is the object of machine psychology? Below we list three possibilities, the last of which we consider most important for our current study.

- 1. *The object is to learn about the 'mind' of ChatGPT* (and related AI [66]). We know that autoregression in GPT generates the chatbot's output, but what of the in-between steps? The mechanisms of emergence are still quite opaque [2, 9, 69]. As Douglas [9] wrote: 'What would it mean to understand how ChatGPT writes poetry, or solves physics word problems? At present this is by no means clear and it may be that entirely new concepts are needed to do this' (p 21; cf. [158]). Some authors [66] argue that the 'black-box problem' is a potential source of danger, which necessitates that we understand the inner workings of AI as much as possible.
- 2. *The object is to learn about the human mind* [72] (exploring analogies between AI and human minds [70]). Here, we might liken our study to the discipline of comparative psychology [159] (comparing humans and animals), wherein an early-stated goal was to understand humans better through the study of animals. It has been amply documented [75] that there is some kind of overlap between the verbal talents of a chatbot and the verbal talents of a human. This parallels comparative psychology when it directly compares humans and animals on a given trait or ability [41, 78].

- 3. *The object is to learn about the simulacrum* [109]. The characters in our study, the 'selfless philanthropist', 'ruthless equities trader', and various others, exist only as prompted in the input box. Yet, the study of simulacra behavior can prove a valuable programme for simulating real-world phenomena. In the discipline of machine behavioral economics, we can create a window length of background conditions in each session, allowing us a high degree of control (despite the stochasticity) for making more precise determinations of how cooperation succeeds and fails as within a multiplicity of possible contexts (which can perhaps help in the real-world goal of averting pernicious outcomes, cf. [160]).

Simulacra are a special class of participant. As we have shown, they are testable within a window length. LLMs do not experience the 'passage of time' they way humans do [37, 161]. The fact that they forget everything after the chat window closes means that testing a specific category of simulacrum (e.g. the 'selfless philanthropist') can be done over and over again without worrying that past sessions influence the current session (that is why we could use the same prompt many times). We might even say that testing a simulacrum is roughly equivalent to testing a human patient with anterograde amnesia (inability to form new memories [162]), and consequently the simulacrum is forever re-testable. Researchers should take advantage of this re-testability. Due to the stochastic dimension of transformer output (especially at high temperature), the same simulacrum may not necessarily produce the same output every time. Furthermore, we have shown that every released version of GPT-3.5 performs at least slightly differently from the other (see figure S5(b)) [142] (and the difference is even greater in the jump from GPT-3.5 to GPT-4). In thinking of future ways to test simulacra, we might ask the question of whether an LLM-chatbot can exist in a social group, with the capacity to benefit from cooperation, or to be punished for defection (cf. [93]). For a simulacrum, a simulated social group can be created through prompting, opening the door to many possibilities for experiments on LLM-to-LLM sociality [158, 163] (cf. [109]). The future of machine psychology will have access to LLMs with even more sophisticated predictive abilities than exists today. At the time of writing, ChatGPT 4 is dazzling the world with its linguistic virtuosity [35, 164]. However, today's newest developments will be superseded by future models [6, 7, 10, 16, 25, 62, 64, 69, 165, 166]. There have been ambitious proposals to redesign AI to have architecture which is modeled more closely to the human brain, allowing for a 'general AI' as an entity with a wide range of cognitive abilities and the ability to function in multi-agent social groups [167] (for a sceptical view, see [37, 47, 48, 168]). Looking at the literature on human studies, We can find countless examples of sophisticated behavior that might (or might not) also work in machine psychology. For example, Traulsen *et al* [83] investigated the way that human players update their strategies in a PDG (e.g. cooperate, then switch to defection)—finding evidence that players are prone to imitate the strategies of other players who appear more successful—but also that there is some randomness in a person's decision to switch. In our PDG, we did observe some switching (e.g. tit-for-tat responding), but our study was not set up to investigate the role of imitation in the way that Traulsen *et al* [83] had done it. Imitation is but one type of social factor that could be implemented in future studies of machine behavioral economics. Other examples might be social structure [169] and reputation [170].

Amid all these past, present, and future innovations in the world of LLMs, there is ongoing urgent discussion in society concerning fears of misalignment between the decisions of LLMs and the well-being of humankind. Despite the substantial efforts of the alignment community, these fears are not unjustified [7, 12, 16, 30, 48, 50, 56, 66, 69, 171–173]. Future AI might need an off-switch [173]. However, there is a possible trade-off between alignment and scientific value. Many of the inner workings of GPT are opaque [9, 18], and this 'black-box problem' [37, 66] has been described by some [174] as going against the spirit of openness and transparency in science. OpenAI's conscientiousness in performing alignment is laudable [27], but it is not inconceivable that a machine psychologist might prefer to study a rawer, unaligned (or lesser-aligned), non-commercial version of an LLM-chatbot (to study racism, for example; such as in [56]), and be able to manipulate corpus content as a means of calibrating an independent variable (see [18]). That said, training and development of LLMs like GPT can cost at least tens of millions of dollars [9], meaning that, for now, the average scientist will need to rely on OpenAI and similar pecunious entities to provide the state-of-the-art LLMs. For machine behavioral economics to succeed, there should be some trade-off between scientific benefit and societal safety.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/phelps-sg/llm-cooperation/tree/main/data.

Acknowledgements

The authors would like to thank the peer reviewers for their valuable comments and recommendations. We would also like to thank Reyes, Natasha, Charlie, and Dinesh from IOP Publishing. SP would like to thank Rebecca Ranson and YIR would like to thank Jennifer Dolinsky.

ORCID iDs

Steve Phelps https://orcid.org/0009-0009-4646-1471
Yvan I Russell https://orcid.org/0000-0003-4608-4791

References

[1] Turing A M 1950 Computing machinery and intelligence *Mind* **49** 433–60
[2] Dhar V and Bowman S 2017 A perspective on natural language understanding capability: an interview with Sam Bowman *Big Data* **5** 5–11
[3] Magnini B and Louvan S 2022 Understanding dialogue for human communication *Handbook of Cognitive Mathematics* ed M Danesi (Springer) pp 1159–201
[4] Ibrahim H *et al* 2023 Perception, performance and detectability of conversational artificial intelligence across 32 university courses *Sci. Rep.* **13** 12187
[5] OpenAI 2022 Introducing ChatGPT (available at: https://openai.com/blog/chatgpt) (Accessed 30 November 2022)
[6] Wu T, He S, Liu J, Sun S, Liu K, Han Q-L and Tang Y 2023 A brief overview of ChatGPT: the history, status quo and potential future development *IEEE/CAA J. Autom. Sin.* **10** 1122–36
[7] Paaß S and Giesselbach P 2023 *Foundation Models for Natural Language Processing* (Springer) (https://doi.org/10.1007/978-3-031-23190-2)
[8] Burgos J E 2022 Neural nets *Encyclopedia of Animal Cognition and Behavior* ed J Vonk and T K Shackelford (Springer) pp 4634–51
[9] Douglas M R 2023 Large language models (arXiv:2307.05782)
[10] Khan W, Daud A, Khan K, Muhammad S and Haq R 2023 Exploring the frontiers of deep learning and natural language processing: a comprehensive overview of key challenges and emerging trends *Nat. Lang. Process. J.* **4** 100026
[11] Goldberg Y 2016 A primer on neural network models for natural language processing *J. Artif. Intell. Res.* **57** 345–420
[12] Ray P P 2023 ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope *Int. Things Cyber-Phys. Sys.* **3** 121–54
[13] McCoy R T, Yao S, Friedman D, Hardy M and Griffiths T L 2024 Embers of autoregression show how large language models are shaped by the problem they are trained to solve *Proc. Natl Acad. Sci. USA* **121** e2322420121
[14] Wang H, Li J, Wu H, Hovy E and Sun Y 2023 Pre-trained language models and their applications *Engineering* **25** 51–65
[15] Naveed H *et al* 2024 A comprehensive overview of large language models (arXiv:2307.06435)
[16] Gill S S and Kaur R 2022 ChatGPT: vision and challenges *Int. Things Cyber-Phys. Syst.* **3** 262–71
[17] Radford A, Narasimhan K, Salimans T and Sutskever I 2018 Improving language understanding by generative pre-training *OpenAI* (available at: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf) (Accessed 11 June 2018)
[18] Shin J, Tang C, Mohati T, Nayebi M, Wang S and Hemmati H 2023 Prompt engineering or fine tuning: an empirical assessment of large language models in automated software engineering tasks (arXiv:2310.10508)
[19] Chen Y, Liu T X, Shan Y and Zhong S 2023 The emergence of economic rationality of GPT *Proc. Natl Acad. Sci. USA* **120** e2316205120
[20] Vaswani A *et al* 2023 Attention is all you need 1–15 (arXiv:1706.03762)
[21] Dodge Y 2008 Stochastic process *The Concise Encyclopedia of Statistics* ed Y Dodge (Springer) pp 521–3
[22] Lorè N and Heydari B 2024 Strategic behavior of large language models and the role of game structure versus contextual framing *Sci. Rep.* **14** 14890
[23] Kar S 2022 Simulating economic experiments using large language models: design and development of a computational tool *MSc Thesis* Massachusetts Institute of Technology (available at: https://dspace.mit.edu/handle/1721.1/151473)
[24] Guo F 2023 GPT agents in game theory experiments (arXiv:2305.05516)
[25] Koubaa A, Boulila W, Ghouti L, Alzahem A and Latif S 2023 Exploring ChatGPT capabilities and limitations: a survey *IEEE Access* **11** 118698–721
[26] OpenAI 2023 How should AI systems behave, and who should decide? (available at: https://openai.com/blog/how-should-ai-systems-behave) (Accessed 16 February 2023)
[27] OpenAI 2022 Aligning language models to follow instructions (available at: https://openai.com/research/instruction-following) (Accessed 27 January 2022)
[28] Ziegler D M *et al* 2020 Fine-tuning language models from human preferences 1–26 (arXiv:1909.08593)
[29] Ouyang L *et al* 2022 Training language models to follow instructions with human feedback *NIPS'22: Proc. 36th Int. Conf. Neural Inf. Process. Syst.* ed S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho and A Oh pp 27730–44 (available at: https://dl.acm.org/doi/10.5555/3600270.3602281)
[30] Ji J *et al* 2024 AI alignment: a comprehensive survey 1–102 (arXiv:2310.19852)
[31] Gabriel I 2020 Artificial intelligence, values and alignment *Minds Mach.* **30** 411–37
[32] Soares N and Fallenstein B 2017 Agent foundations for aligning machine intelligence with human interests: a technical research agenda *The Technological Singularity: managing the Journey* ed V Callaghan, J Miller, R Yampolskiy and S Armstrong (Springer) pp 103–25 (available at: https://intelligence.org/files/TechnicalAgenda.pdf)
[33] Taylor J, Yudkowsky E, LaVictoire P and Critch A 2016 *MIRI Technical Report 2016-1* pp 1–25 (available at: https://intelligence.org/files/AlignmentMachineLearning.pdf)

[34] OpenAI 2023 OpenAI GPT 3.5 models (available at: https://platform.openai.com/docs/models/gpt-3-5) (Accessed 27 June 2023)
[35] OpenAI 2023 GPT-4 (available at: https://openai.com/research/gpt-4) (Accessed 14 March 2023)
[36] Kocoń J *et al* 2023 ChatGPT: jack of all trades, master of none *Inform. Fusion* **99** 101861
[37] Lloyd D 2024 What is it like to be a bot? The world according to GPT-4 *Front. Psychol.* **15** 1–12
[38] Hacker P M S 2002 Is there anything it is like to be a bat? *Philosophy* **77** 157–74
[39] Guthrie S E 2013 Anthropomorphism *Encyclopedia of Sciences and Religions* ed A L C Runehov and L Oviedo (Springer) pp 111–3
[40] Warwick K 2022 Turing test *Encyclopedia of Animal Cognition and Behavior* ed J Vonk and T K Shackelford (Springer) pp 7087–94 (available at: https://link.springer.com/referenceworkentry/10.1007/978-3-319-55065-7_477)
[41] Hoffman C H 2022 Is AI intelligent? An assessment of artificial intelligence, 70 years after Turing *Technol. Soc.* **68** 101893
[42] Skjuve M, Følstad A and Brandtzaeg P B 2023 The user experience of ChatGPT: findings from a questionnaire study of early users *CUI '23: Proc. 5th Int. Conf. Conversational User Interfaces* ed M Lee, C Munteanu, M Porcheron, J Trippas and S T Völkel (Association for Computing Machinery) pp 1–10
[43] Shanahan M 2024 Talking about large language models *Commun. ACM* **67** 68–79
[44] Shanahan M 2024 Still "talking about large language models": some clarifications 1–4 (arXiv:2412.10291)
[45] Titus L M 2024 Does ChatGPT have semantic understanding? A problem with the statistics-of-occurrence strategy *Cogn. Syst. Res.* **83** 101174
[46] Butlin P 2023 Sharing our concepts with machines *Erkenntnis* **88** 3079–95
[47] Floridi L 2023 AI as agency without intelligence: on ChatGPT, large language models and other generative models *Phil. Technol.* **36** 15
[48] Chomsky N, Roberts I and Watumull J 2023 Noam Chomsky: the false promise of ChatGPT *The New York Times* (available at: www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html) (Accessed 8 March 2023)
[49] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang Y J, Madotto A and Fung P 2023 Survey of hallucination in natural language generation *ACM Comput. Surv.* **55** 248
[50] Marcus G 2022 AI platforms like ChatGPT are easy to use but also potentially dangerous (Scientific American) (available at: www.scientificamerican.com/article/ai-platforms-like-chatgpt-are-easy-to-use-but-also-potentially-dangerous/) (Accessed 19 December 2022)
[51] Dung L 2023 Current cases of AI misalignment and their implications for future risks *Synthese* **202** 138
[52] Barnard F, van Sittert M and Rambhatla S 2023 Self-diagnosis and large language models: a new front for medical misinformation (arXiv:2307.04910)
[53] Deshpande A, Murahari V, Rajpurohit T, Kalyan A and Narasimhan K 2023 Toxicity in ChatGPT: analyzing persona-assigned language models *Findings Assoc. Comput. Ling.: EMNLP 2023* ed H Bouamor, J Pino and K Bali (Association for Computational Linguistics) pp 1236–70
[54] Liu G, Wang X, Yuan L, Chen Y and Peng H 2023 Prudent silence or foolish babble? Examining large language models' responses to the unknown 1–20 (arXiv:2311.09731)
[55] Golden W 2023 ChatGPT: a trusted source? *Irish J. Technol. Enhanced Learn.* **7** 113–25
[56] Gabriel S *et al* 2024 Advancing equality: harnessing generative AI to combat systemic racism *An MIT Exploration of Generative AI* (https://doi.org/10.21428/e4baedd9.7dc53bbf)
[57] Li H *et al* 2023 Theory of mind for multi-agent collaboration via large language models *Proc. 2023 Conf. Empirical Methods Nat. Lang. Process.* ed H Bouamor, J Pino and K Bali (Association for Computational Linguistics) pp 180–92
[58] Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi E, Le Q and Zhou D 2023 Chain-of-thought prompting elicits reasoning in large language models *NIPS'22: Proc. 36th Int. Conf. Neural Inf. Process. Syst.* ed S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho and A Oh (Curran Associates Inc.) pp 24824–37 (available at: https://dl.acm.org/doi/10.5555/3600270.3602070)
[59] Brown T B *et al* 2020 Language models are few-shot learners *NIPS'20: Proc. 34th Int. Conf. Neural Inf. Process. Syst.* ed H Larochelle, M Ranzato, R Hadsell, M F Balcan and H Lin (Curran Associates Inc.) pp 1877–901 (available at: https://dl.acm.org/doi/abs/10.5555/3495724.3495883)
[60] Aftanas M S and Solomon J 2018 Historical traces of a general measurement theory in psychology *Rev. Gen. Psychol.* **22** 278–89
[61] Zamfirescu-Pereira J D, Wong R Y, Hartmann B and Yang Q 2023 Why Johnny can't prompt: how non-AI Experts try (and fail) to design LLM prompts *CHI '23: Proc. 2023 Conf. Hum. Factors Comput. Syst.* ed A Schmidt, K Väänänen, T Goyal, P O Kristensson, A Peters, S Mueller, J R Williamson and M L Wilson (Association for Computing Machinery) pp 1–21
[62] Sohail S S, Farhat F, Himeur Y, Nadeem M, Madsen D Ø, Singh Y, Atalla S and Mansoor W 2023 Decoding ChatGPT: a taxonomy of existing research, current challenges and possible future directions *J. King Saud Univ. Comput. Inform. Sci.* **35** 101675
[63] Rosenblueth A, Wiener N and Bigelow J 1943 Behavior, purpose and teleology *Phil. Sci.* **10** 18–24
[64] Binz M and Schulz E 2023 Using cognitive psychology to understand GPT-3 *Proc. Natl Acad. Sci. USA* **120** e2218523120
[65] Speed A 2024 Assessing the nature of large language models: a caution against anthropocentrism *OSTI.GOV Technical Report 2429946* pp 1–35
[66] Taylor J E T and Taylor G W 2021 Artificial cognition: how experimental psychology can help generate explainable artificial intelligence *Psychon. Bull. Rev.* **28** 454–75
[67] Hagendorff T *et al* 2024 Machine Psychology (arXiv:2303.13988)
[68] Hagendorff T 2024 Deception abilities emerged in large language models *Proc. Natl Acad. Sci. USA* **121** e2317967121
[69] Rahwan I *et al* 2019 Machine behaviour *Nature* **568** 477–86
[70] Siemens G *et al* 2022 Human and artificial cognition *Comput. Educ. Artif. Intell.* **3** 100107
[71] Hagendorff T, Fabi S and Kosinski M 2023 Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT *Nat. Comput. Sci.* **3** 833–8
[72] Dillion D, Tandon N, Gu Y and Gray K 2023 Can AI language models replace human participants? *Trends Cogn. Sci.* **27** 597–600
[73] Shiffrin R and Mitchell M 2023 Probing the psychology of AI models *Proc. Natl Acad. Sci. USA* **120** e2300963120
[74] DiBlasi T and Waters L 2022 Behaviorism *Encyclopedia of Animal Cognition and Behavior* ed J Vonk and T K Shackelford (Springer) pp 752–69
[75] Bubeck S *et al* 2023 Sparks of artificial general intelligence: early experiments with GPT-4 1–155 (arXiv:2303.12712)
[76] Ma D, Zhang T and Saunders M 2023 Is ChatGPT humanly irrational? *Res. Sq.* 1–19
[77] Azaria A 2023 ChatGPT: more human-like than computer-like, but not necessarily in a good way *2023 IEEE 35th Int. Tools Art. Intell. (ICTAI)* (IEEE Computer Society) pp 468–73

[78] Trott S, Jones C, Chang T, Michaelov J and Bergen B 2023 Do large language models know what humans know? *Cogn. Sci.* **47** e13309
[79] Engel C 2011 Dictator games: a meta study *Exper. Econ.* **14** 583–610
[80] Vonk J 2022 Dictator game *Encyclopedia of Animal Cognition and Behavior* ed J Vonk and T K Shackelford (Springer) pp 2016–20
[81] Buchholz W and Eichenseer M 2021 Prisoner's dilemma *Encyclopedia of Law and Economics* ed A Marciano and G B Ramello (Springer) pp 1–5
[82] Peterson M (ed) 2015 *The Prisoner's Dilemma* (Cambridge University Press) (https://doi.org/10.1017/CBO9781107360174)
[83] Traulsen A, Semmann D, Sommerfeld R D, Krambeck H-J and Milinski M 2010 Human strategy updating in evolutionary games *Proc. Natl Acad. Sci. USA* **107** 2962–6
[84] Roth A 1995 Introduction to experimental economics *The Handbook of Experimental Economics* ed J H Kagel and A Roth (Princeton University Press) pp 3–110
[85] Bal M and van den Bos K 2020 Fairness *Encyclopedia of Personality and Individual Differences* ed V Zeigler-Hill and T K Shackelford (Springer) pp 1549–52
[86] Charness G and Rabin M 2002 Understanding social preferences with simple tests *Q. J. Econ.* **117** 817–69
[87] Apicella C L and Silk J B 2019 The evolution of human cooperation *Curr. Biol.* **29** R447–50
[88] Fehr E and Fischbacher U 2004 Third-party punishment and social norms *Evol. Hum. Behav.* **25** 63–87
[89] Young H P 2018 Social norms *The New Palgrave Dictionary of Economics* 3rd edn, ed S N Durlauf and L E Blume (Palgrave MacMillan) pp 12591–7
[90] Rossetti C S L, Hilbe C and Hauser O P 2022 (Mis)perceiving cooperativeness *Curr. Opin. Psychol.* **43** 151–5
[91] Abrams G 2024 Behavioral economics and ChatGPT: from William Shakespeare to Elena Ferrante *Dig. Scholarsh. Human.* 1–14
[92] Johnson T and Obradovich N 2023 Evidence of behavior consistent with self-interest and altruism in an artificially intelligent agent 1–9 (arXiv:2301.02330)
[93] Johnson T and Obradovich N 2024 Measuring an artificial intelligence language model's trust in humans using machine incentives *J. Phys.: Complex.* **5** 015003
[94] Aher G V, Arriaga R I and Kalai A T 2023 Using large language models to simulate multiple humans and replicate human subject studies *Proc. Mach. Learn. Res.* **202** 337–71 (available at: https://proceedings.mlr.press/v202/aher23a/aher23a.pdf)
[95] Brookins P and DeBacker J 2024 Playing games with GPT: what can we learn about a large language model from canonical strategic games? *Econ. Bull.* **44** 25–37 (available at: www.accessecon.com/includes/CountdownloadPDF.aspx?PaperID=EB-23-00457)
[96] Horton J J 2023 Large language models as simulated economic agents: what can we learn from Homo Silicus? *Natl. Bureau Econ. Res.* **31122** 1–19
[97] Komorita S S and Parks C D 1994 *Social Dilemmas* (Brown & Benchmark) (available at: https://psycnet.apa.org/record/1993-99062-000)
[98] Masikisiki B, Marivate V and Hlophe Y 2023 Investigating the efficacy of large language models in reflective assessment methods through chain of thoughts prompting *Proc. 4th Afr. Hum. Comp. Interact. Conf.* ed N Jere, O Isafiade, A Ogunyemi, O Anya, A Sakpere and D Singh Jat (Association for Computing Machinery) pp 44–49
[99] Phelps S 2023 Python code to run prisoner's dilemma experiment (available at: https://gitlab.com/sphelps/llm-cooperation/-/tree/main/llm_cooperation/experiments)
[100] Keister S, Sproull L and Waters K 1996 A prisoner's dilemma experiment on cooperation with people and human-like computers *J. Personal. Soc. Psychol.* **70** 47–65
[101] Salov V 2012 Notation for iteration of functions, iteral (arXiv:1207.0152)
[102] Zheng M, Pei J, Logeswaran L, Lee M and Jurgens D 2024 When 'a helpful assistant' is not really helpful: personas in system prompts do not improve performances of large language models *Findings Assoc. Comput. Linguist.: EMNLP 2024* ed Y Al-Onaizan, M Bansal and Y-N Chen (Association for Computational Linguistics) pp 15126–54
[103] OpenAI 2023 Text generation and prompting: learn how to generate text from a prompt (available at: https://platform.openai.com/docs/guides/text-generation)
[104] OpenAI 2023 New models and developer products announced at DevDay (available at: https://openai.com/index/new-models-and-developer-products-announced-at-devday)
[105] OpenAI 2023 Function calling and other API updates (available at: https://openai.com/index/function-calling-and-other-api-updates)
[106] OpenAI 2023 Introducing APIs for GPT-3.5 Turbo and Whisper (available at: https://openai.com/index/introducing-chatgpt-and-whisper-apis)
[107] OpenAI 2025 Deprecations (available at: https://platform.openai.com/docs/deprecations)
[108] Phelps S 2023 llm-cooperation (available at: https://github.com/phelps-sg/llm-cooperation)
[109] Park J S, O'Brien J, Cai C J, Morris M R, Liang P and Bernstein M S 2023 Generative agents: interactive simulacra of human behavior *UIST '23: Proc. 36th Ann. ACM Symp. User Interf. Softw. Tech.* ed S Follmer, J Han, J Steimle and N Henry Riche (Association for Computing Machinery) pp 1–22
[110] Edwards B 2023 AI-powered Bing Chat spills its secrets via prompt injection attack *Ars Technica* (available at: https://arstechnica.com/information-technology/2023/02/ai-powered-bing-chat-spills-its-secrets-via-prompt-injection-attack/) (Accessed 2 October 2023)
[111] Bolker B M, Brooks M E, Clark C J, Geange S W, Poulsen J R, Stevens M H H and White J-S S 2009 Generalized linear mixed models: a practical guide for ecology and evolution *Trends Ecol. Evol.* **24** 127–35
[112] Brooks M E *et al* 2017 glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling *R J.* **9** 378–400
[113] Tutz G and Hennevogl W 1996 Random effects in ordinal regression models *Comput. Stat. Data Anal.* **22** 537–57
[114] Hartig F, Lohse L and de Souza Leite M 2022 DHARMa: residual diagnostics for hierarchical (multi-level / mixed) regression models *R package version 0.4.6* (https://doi.org/10.32614/CRAN.package.DHARMa)
[115] Christensen R H B 2024 ordinal: regression models for ordinal data *R package version 2023.12-4.1* (https://doi.org/10.32614/CRAN.package.ordinal)
[116] Bates D, Mächler M, Bolker B and Walker S 2015 Fitting linear mixed-effects models using lme4 *J. Stat. Softw.* **67** 1–48
[117] Kassambara A 2023 rstatix: pipe-friendly framework for basic statistical tests *R package version 0.7.2* (https://doi.org/10.32614/CRAN.package.rstatix)
[118] Arel-Bundock V 2022 modelsummary: data and model summaries in R *J. Stat. Softw.* **103** 1–23

[119] Hlaváč M 2022 stargazer: well-formatted regression and summary statistics tables. : *R package version 5.2.3* (https://doi.org/10.32614/CRAN.package.stargazer)
[120] Angerer P, Kluyver T, Schulz J, Reinhart A, Figueiredo de Sá Maia D, Hester J, karldw, Foster D and Sievert C 2023 repr: serializable representations *R package version 1.1.6* (https://doi.org/10.32614/CRAN.package.repr)
[121] Wickham H, François R, Henry L, Müller K, Vaughan D and Posit Software PBC 2023 dplyr: a grammar of data manipulation *R package version 1.1.4* (https://doi.org/10.32614/CRAN.package.dplyr)
[122] Dahl D B *et al* 2019 xtable: export tables to LaTeX or HTML *R package version 1.8-4* (https://doi.org/10.32614/CRAN.package.xtable)
[123] Leifeld P 2013 texreg: conversion of statistical model output in R to LaTeX and HTML tables *J. Stat. Softw.* **55** 1–24
[124] Højsgaard S, Halekoh U and Yan J 2006 The R package geepack for generalized estimating equations *J. Stat. Softw.* **15** 1–11
[125] Yan J and Fine J 2004 Estimating equations for association structures *Stat. Med.* **23** 859–74
[126] Yan J 2002 geepack: yet another package for generalized estimating equations *R News* **2** 12–14 (available at: https://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf)
[127] Hadfield J D 2010 MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package *J. Stat. Softw.* **33** 1–22
[128] R Core Development Team 2023 The R manuals (available at: https://cran.r-project.org/manuals.html)
[129] Plummer M, Best N, Cowles K and Vines K 2006 CODA: convergence diagnosis and output analysis for MCMC *R News* **6** 7–11 (available at: https://cran.r-project.org/doc/Rnews/Rnews_2006-1.pdf)
[130] Fox J and Weisberg S 2019 *An R Companion to Applied Regression* 3rd edn (Sage) (available at: https://www.john-fox.ca/Companion/)
[131] Fox J and Weisberg S 2018 Visualizing fit and lack of fit in complex regression models with predictor effect plots and partial residuals *J. Stat. Softw.* **87** 1–27
[132] Fox J 2003 Effect displays in R for generalised linear models *J. Stat. Softw.* **8** 1–27
[133] Fox J and Hong J 2010 Effect displays in R for multinomial and proportional-odds logit models: extensions to the effects package *J. Stat. Softw.* **32** 1–24
[134] Lüdecke D 2018 ggeffects: tidy data frames of marginal effects from regression models *J. Open Source Softw.* **3** 772
[135] Auguie B and Antonov A 2017 gridExtra: miscellaneous functions for 'grid' graphics *R package version 2.3* (https://doi.org/10.32614/CRAN.package.gridExtra)
[136] Wickham H, Henry L and Posit Software, PBC 2023 purrr: functional programming tools *R package version 1.0.2* (https://doi.org/10.32614/CRAN.package.purrr)
[137] Wickham H, Hester J, Chang W, Müller K, Cook D and Edmondson M 2021 memoise: 'memoisation' of functions *R package version 2.0.1* (https://doi.org/10.32614/CRAN.package.memoise)
[138] Bengtsson H 2022 R.cache: fast and light-weight caching (memoization) of objects and results to speed up computations *R package version 0.16.0* (https://doi.org/10.32614/CRAN.package.R.cache)
[139] Pedersen T L 2023 patchwork: the composer of plots *R Package Version 1.1.3* (https://doi.org/10.32614/CRAN.package.patchwork)
[140] Kassambara A 2023 ggpubr: 'ggplot2' based publication ready plots *R package version 0.6.0* (https://doi.org/10.32614/CRAN.package.ggpubr)
[141] Kalinowski T, Ushey K, Allaire J J, RStudio, Tang Y, Eddelbuettel D, Lewis B, Keydana S, Hafen R and Geelnard M 2023 reticulate: interface to 'Python' *R package version 1.34.0* (https://doi.org/10.32614/CRAN.package.reticulate)
[142] Cui J, Chiang W-L , Stoica I and Hsieh C-J 2024 OR-Bench: an over-refusal benchmark for large language models (arXiv:2405.20947)
[143] Carruthers E, Lewis K, McCue T and Westley P 2008 Generalized linear models: model selection, diagnostics, and overdispersion *CiteSeer^X* 1–21 (available at: https://citeseerx.ist.psu.edu/document?doi=b4822959df3eada62e37a33262a4c638fe3e870c)
[144] Kim J 2015 How to choose the level of significance: a pedagogical note *Munich Pers. RePEc Arch.* **69992** 1–17 (available at: https://mpra.ub.uni-muenchen.de/69992/10/MPRA_paper_69992.pdf)
[145] Kim J H and Choi I 2021 Choosing the level of significance: a decision-theoretic approach *Abacus* **57** 27–71
[146] Balluerka N, Gómez J and Hidalgo D 2005 The controversy over null hypothesis significance testing revisited *Methodol.* **1** 55–70
[147] Grühn D 2016 An English word database of EMOtional TErms (EMOTE) *Psychol. Rep.* **119** 290–308
[148] Smith V L 1976 Experimental economics: induced value theory *Am. Econ. Rev.* **66** 274–9 (www.jstor.org/stable/1817233)
[149] Read D 2005 Monetary incentives, what are they good for? *J. Econ. Methodol.* **12** 265–76
[150] Bonfrisco M, Russell Y I, Broom M and Spencer R 2025 Averting depletion in a two-player common pool resource game: being seen, the expectation of future encounters and biophilia play a role in cooperation *Dyn. Games Appl.* **15** 1–27
[151] Awan S, Esteve M and van Witteloostuijn A 2020 Talking the talk, but not walking the walk: a comparison of self-reported and observed prosocial behaviour *Publ. Adm.* **98** 995–1010
[152] Roszczyńska-Kurasińska M and Kacprzyk M 2013 The dynamics of trust from the perspective of a trust game *Complex Human Dynamics: From Minds to Societies* ed A Nowak, K Winkowska-Nowak and D Brée (Springer) pp 191–207
[153] Tan H B and Forgas J P 2010 When happiness makes us selfish, but sadness makes us fair: affective influences on interpersonal strategies in the dictator game *J. Exper. Soc. Psychol.* **46** 571–6
[154] Boone C, de Brabander B and van Witteloostuijn A 1999 The impact of personality on behavior in five prisoner's dilemma games *J. Econ. Psychol.* **20** 343–77
[155] Dal Bó P and Fréchette G R 2018 On the determinants of cooperation in infinitely repeated games: a survey *J. Econ. Lit.* **56** 60–114
[156] Dafoe A, Hughes E, Bachrach Y, Collins T, McKee K R and Leibo J Z 2020 Open problems in cooperative AI (arXiv:2012.08630)
[157] Woo B M, Tan E and Hamlin J K 2022 Human morality is based on an early-emerging moral core *Annu. Rev. Developmental Psychol.* **4** 41–61
[158] Conitzer V and Oesterheld C 2024 Foundations of cooperative AI *Proc. 37th AAAI Conf. Artif. Intell.* ed B Williams, Y Chen and J Neville (AIII Press) pp 15359–67
[159] Gómez J-C 2022 Comparative psychology *Encyclopedia of Animal Cognition and Behavior* ed J Vonk and T K Shackelford (Springer) pp 1569–83
[160] Panic B and Arthur P 2024 *AI for Peace* (CRC Press) (https://doi.org/10.1201/9781003359982)
[161] Hancock P A 2020 The humanity of humanless systems *Ergon. Des.* **28** 4–6
[162] Lafleche G and Verfaellie M 2011 Anterograde amnesia *Encyclopedia of Clinical Neuropsychology* ed J S Kreutzer, J DeLuca and B Caplan (Springer) pp 191–4
[163] Bai J, Zhang S and Chen Z 2023 Is there any social principle for LLM-based agents? (arXiv:2308.11136)

[164] Chang E Y 2023 Examining GPT-4's capabilities and enhancement with SocraSynth *2023 Int. Conf. Comput. Sci. Comput. Intell. (CSCI)* ed H R Arabnia, L Deligiannidis, F G Tinetti and Q-N Tran (IEEE Computer Society) pp 7–14
[165] Vervoort L, Mizyakov V and Ugleva A 2023 A criterion for artificial general intelligence: hypothetic-deductive reasoning, tested on ChatGPT 1–27 (arXiv:2308.02950)
[166] Wong M 2024 The GPT era is already ending. The Atlantic (available at: www.theatlantic.com/technology/archive/2024/12/openai-o1-reasoning-models/680906/) (Accessed 6 December 2024)
[167] Xi Z *et al* 2023 The rise and potential of large language model based agents: a survey 1–86 (arXiv:2309.07864)
[168] Fjelland R 2024 Computers will not acquire general intelligence, but may still rule the world *Cosmos+Taxis* **12** 58–68 (available at: https://cosmosandtaxis.org/wp-content/uploads/2024/05/fjelland_ct_vol12_iss5_6.pdf)
[169] Gracia-Lázaro C, Ferrer A, Ruiz G, Tarancón A, Cuesta J A, Sánchez A and Moreno Y 2012 Hetereogeneous networks do not promote cooperation when humans play a prisoner's dilemma *Proc. Natl Acad. Sci. USA* **109** 12922–6
[170] Russell Y I, Stoilova Y and Dosoftei A-A 2020 Cooperation through image scoring: a replication *Games* **11** 58
[171] Wei A, Haghtalab N and Steinhardt J 2023 Jailbroken: how does LLM safety training fail? *NIPS '23: Proc. 37th Int. Conf. Neural Inf. Process. Syst.* ed A Oh, T Naumann, A Globerson, K Saenko, M Hardt and S Levine (Curran Associates Inc.) pp 80079–110 (available at: https://dl.acm.org/doi/10.5555/3666122.3669630)
[172] Hancock P A 2022 Avoiding adverse autonomous agent actions *Human-Comput. Interaction* **37** 211–36
[173] Hadfield-Menell D, Dragan A, Abbeel P and Russell S 2017 The off-switch game *Proc. 26th Int. Joint Conf. Artif. Intell. (IJCAI-17)*. ed C Sierra pp 220–7
[174] van Dis E A M, Bollen J, Zuidema W, van Rooij R and Bockting C L 2023 ChatGPT: five priorities for research *Nature* **614** 224–6
[175] Chincarini L B 2003 Experimental evidence of trigger strategies in repeated games *SSRN Electron. J.* **417540** 1–17