INTERFACE

Review

Author for correspondence:
Valerio Capraro
e-mail: capraovalerio@gmail.com

Language-based game theory in the age of artificial intelligence

Valerio Capraro[1], Roberto Di Paolo[2], Matjaž Perc[3,4,5,6] and Veronica Pizziol[7]

[1]Department of Psychology, University of Milan Bicocca, Milano, Italy
[2]Department of Economics and Management, University of Parma, Parma, Italy
[3]Faculty of Natural Sciences and Mathematics, University of Maribor, Maribor, Slovenia
[4]Community Healthcare Center Dr. Adolf Drolc Maribor, Maribor, Slovenia
[5]Complexity Science Hub Vienna, Vienna, Austria
[6]Department of Physics, Kyung Hee University, Seoul, Republic of Korea
[7]Department of Economics, University of Bologna, Bologna, Italy

VC, 0000-0002-0579-0166; RDP, 0000-0002-6081-6656; MP, 0000-0002-3087-541X;
VP, 0000-0002-9032-8506

Understanding human behaviour in decision problems and strategic interactions has wide-ranging applications in economics, psychology and artificial intelligence. Game theory offers a robust foundation for this understanding, based on the idea that individuals aim to maximize a utility function. However, the exact factors influencing strategy choices remain elusive. While traditional models try to explain human behaviour as a function of the outcomes of available actions, recent experimental research reveals that linguistic content significantly impacts decision-making, thus prompting a paradigm shift from outcome-based to language-based utility functions. This shift is more urgent than ever, given the advancement of generative AI, which has the potential to support humans in making critical decisions through language-based interactions. We propose sentiment analysis as a fundamental tool for this shift and take an initial step by analysing 61 experimental instructions from the dictator game, an economic game capturing the balance between self-interest and the interest of others, which is at the core of many social interactions. Our meta-analysis shows that sentiment analysis can explain human behaviour beyond economic outcomes. We discuss future research directions. We hope this work sets the stage for a novel game-theoretical approach that emphasizes the importance of language in human decisions.

1. Introduction

Understanding human behaviour in decision problems and strategic interactions has long been at the heart of social science research due to its wide-ranging applications in various fields such as economics, psychology and artificial intelligence. In the past century, game theory has provided a compelling framework for this understanding, underpinned by the notion that individuals seek to maximize a utility function [1]. Yet, the components of this utility function remain unknown. What factors do people consider in their strategy choices?

One of the central tools for this exploration has been economic games [2]. These games are particularly insightful for understanding human actions in situations where personal interests are at play. In particular, the study of one-shot and anonymous games, where participants interact once without the possibility of future interactions and without possessing any information about the others, has garnered particular attention, as they offer a clean benchmark to study human behaviour, free from the confounding factors of direct and indirect reciprocity.

It has long been known that even in these one-shot and anonymous games, individuals often do not act purely to maximize their economic benefits. Take,

for instance, the dictator game. In this game, a participant has to decide how to divide a sum of money between themselves and another player. The other player has no active role and only receives the amount that the first player decides to give. This is one of the most studied games in behavioural economics, due to its ability to capture how people balance self-interest (keeping all the money) and the interest of others (giving away some of the money). Despite the absence of extrinsic incentives, a significant number of participants choose to share some amount [3–5]. This brings us to one of the most fundamental questions in behavioural game theory: if not monetary gain, what exactly are people optimizing for?

2. A paradigm crisis

Over the past 30 years, the concept of 'social preferences' has gained traction. Here, an individual's utility is a function not just of their own monetary outcomes but also those of others with whom they are interacting. The formalization of this idea has taken various shapes (e.g. [6–11]; see [12,13] for reviews). Ledyard, for example, postulates that people combine a preference for maximizing their own monetary payoff with a preference for maximizing the payoff for all others involved [6]. By contrast, Fehr and Schmidt's utility revolves around the idea that people aim to maximize their own monetary rewards but also strive to minimize the disparity between their gains and those of other players [8]. Charness and Rabin take a different route, assuming that people aim to optimize their own benefits while simultaneously maximizing the overall welfare of the group [10]. Therefore, these models, though distinct, share a central 'consequentialist assumption': the utility derived from a decision depends solely on the economic outcomes of that choice.

However, this consequentialist assumption has recently been subjected to serious criticism. One major point of contention arises from experimental research with human participants, emphasizing the profound influence of linguistic content on decision-making. Simply put, the way actions are described can significantly alter people's choices, challenging the purely consequentialist assumption of social preferences. In a pioneering study, Liberman and colleagues observed that mere linguistic labels could affect individuals' behaviour in the Prisoner's Dilemma. Specifically, when a Prisoner's Dilemma game was labelled as a 'community game', participants tended to cooperate more often compared to when the very same game was labelled as a 'Wall Street game' [14]. This linguistic impact on decision-making was further reinforced by Eriksson and colleagues. Their research revealed that responders in the ultimatum game were more prone to rejecting low offers when the rejection action was phrased as 'reject the proposer's offer', as opposed to 'decrease the proposer's payoff' [15]. Similarly, a study by Capraro and Rand identified that linguistic framing influenced decisions involving equity versus efficiency trade-offs. When the two options were respectively termed 'more fair' versus 'less fair', participants leaned towards the equitable choice. However, when framed as 'less generous' versus 'more generous', participants were inclined towards the efficient option [16]. Furthermore, Capraro and Vanzo demonstrated the power of language in the dictator game. They conducted six dictator game variants differing only in the label used to describe the available actions and found that people's level of altruistic behaviour significantly depended on the label. For example, individuals were less inclined towards altruism when the altruistic action was labelled as 'boost' rather than 'donate' [17]. In the last 5 years, the effects of language on decisions have been replicated many times in different contexts [18–24]. Some research has also highlighted a dark side of the linguistic framing effect. For example, Capraro and colleagues have shown that when dictator game receivers are given the power to choose the experimental instructions to present to dictators, a significant proportion of receivers choose instructions that are more likely to provide them a higher payoff [25]. Furthermore, Ścigała and colleagues have found that more moral people—defined as those higher in the personality trait of honesty-humility—can be manipulated and turned into accepting a bribe by simply calling the bribe a 'cooperation act' [26].

Cumulatively, these studies challenge the consequentialist assumption of social preferences and demonstrate that utility functions cannot be merely based on the economic outcomes of the available actions. Utility functions must take into account the language used to describe the actions. For this reason, it has been argued that 'behavioural economics is in the midst of a paradigm shift from outcome-based to language-based preferences' [13].

3. The importance of language-based preferences in human–machine interactions

This paradigm shift is more urgent than ever due to the rise of generative artificial intelligence (AI). The evolution of generative AI is reshaping our digital landscapes in ways previously thought impossible. One of the most transformative aspects of this revolution is the increasing capability of AI systems to generate human-like, coherent and contextually relevant textual content. OpenAI's generative pre-trained transformer (GPT) series, for instance, exemplifies this shift, showcasing the ability to produce text that not only reads naturally but also responds contextually to diverse prompts [27].

This rise of text-generating AI has profound implications for decision-making support. As AI becomes more sophisticated, chatbots and virtual assistants powered by these algorithms can guide users in making critical decisions [28–30]. Whether it is a financial choice, a health concern or a complex business strategy, AI chatbots can present users with detailed analyses, recommendations and even potential consequences, all communicated in natural language.

However, these potential benefits come with critical challenges. Previous literature has identified several potential issues, such as accountability, especially when AI-guided decisions lead to adverse outcomes, and the risk of overreliance, which could erode human judgement and decision-making abilities [31–33]. In this article, we shift focus to an often-overlooked aspect: the linguistic description of the decision context. As reviewed above, even in the simplified case of one-shot and anonymous economic games, human decision-making is not just a by-product of the economic consequences of the available actions; it is heavily influenced by the way information is presented. Considering AI's capacity to generate text, it is therefore vital to understand and account for these linguistic frames. If AI chatbots are to aid in important decisions, they need to be designed with an awareness of the effects of language framing. Misrepresentation or linguistic biases, even if unintentional, could lead individuals down undesired paths, with potential downstream

negative effects at the social level. For example, linguistic biases may exacerbate discrimination against marginalized groups [34–36]. Moreover, there is also an ethical imperative to ensure that AI's capability to produce text is not misused, manipulating users' decisions to serve ulterior motives [37–39].

Closing the circle and going back to a game-theoretical point of view, this set of concerns brings us to a novel question: how can the linguistic content of a piece of text be quantified in a way that can be incorporated into the utility function?

4. Using sentiment analysis to define the utility function over language

A straightforward idea is to use sentiment analysis [40]. Sentiment analysis, at its core, is a set of tools developed by computational linguists to evaluate the emotional tenor of a given piece of text. These tools, driven by complex algorithms and based on large linguistic databases, scan textual information to identify and quantify the emotional content embedded within it. Earlier sentiment analysis tools were primarily limited to determining whether a text was positive or negative without taking into account different human emotions. For example, SentiWordNet associates with each 'synset' (a set of synonyms) three continuous numeric scores: positivity, negativity and neutrality, which together sum up to 1. To evaluate the sentiment of a text using SentiWordNet, an algorithm breaks the text into constituent terms or phrases. Each term's corresponding synset scores are then calculated through a suitable automatic annotation process. By aggregating these scores across the entire text, an overall sentiment value can be derived. The aggregation involves weighted sums, where more contextually important words have a greater influence on the final sentiment [41]. More recently, newer sentiment analysis tools have begun to emerge. These tools aim to go beyond the basic binary of positivity and negativity. For example, Mohammad and Turney developed a tool that can identify and measure a spectrum of emotions, such as joy, sadness, anger, fear, surprise, and disgust. By mapping words to specific emotions and emotional intensities, this tool offers a more faceted understanding of sentiment in textual data, capturing the multidimensional nature of human emotions more effectively than earlier models [42].

Yet, while understanding and measuring emotions in the text is undeniably valuable, it is essential to consider that there is more to text sentiment than just emotions. Some linguistic content might emphasize specific behavioural norms or cultural standards, which might be relatively detached from the emotions they convey. Therefore, sentiment analysis tools need to be extended or adapted to measure the normative content of text, detached from emotions. Some work has already been done on this. Moral foundations theory offers one framework for such analysis, emphasizing five (later extended to six) core moral values: care, fairness, loyalty, authority, sanctity and liberty [43]. Computational models have been developed to identify these moral values in textual data, effectively giving a 'normative score' based on these foundations [44]. The same happens for the more recent morality-as-cooperation theory [45].

Leaving aside this intricate web of emotional and normative dimensions, if we were to simplify our approach, the general idea would be to use sentiment analysis tools to quantify the linguistic descriptions of possible actions in a decision problem. These quantified descriptions, represented as numerical scores, would then be fed into a utility function. Of course, implementing such a procedure is not straightforward. The intertwining of sentiment analysis with utility functions would invariably be complex. However, there might be exceptions. For economic games with clear-cut decisions, such as the dictator game, the integration might be more straightforward due to the inherent simplicity of the decision structure in such scenarios.

5. Predictions in the dictator game

Meta-analytic results of the available experimental literature have revealed that most players in the dictator game choose one of three actions: keeping all the money, sharing it equally, or giving it all away [5]. In an attempt to simplify our proposed approach, we can use these results to assume that when participants in the dictator game make their decision, they consider the utility of only these three actions, disregarding other possible actions.

In our proposed model, the utility that a player derives from choosing a particular strategy is determined by two factors: the monetary payoff and the sentiment associated with the description of that strategy.

Considering material payoffs, the strategy yielding the highest monetary return is to keep all the money. This is followed by sharing half (which provides a moderate return) and then giving away all the money (which results in no material gain). However, when we introduce the concept of sentiment into the equation, the analysis becomes more complex, as the sentiment associated with each action can significantly influence the player's decision. There are three cases:

— If the sentiment attached to acting selfishly (keeping all the money) is the highest, a player will choose this option, as it maximizes both material payoff and sentiment.
— If the sentiment for the inequity-averse action (sharing the money equally) is the highest, the player faces a dilemma between the selfish action, which offers the highest material payoff, and the egalitarian action, which aligns with the most favourable sentiment. In this case, acting altruistically (giving all the money) is disregarded, as it is dominated by the inequity-averse action, both in terms of material payoff and sentiment.
— If the sentiment for the altruistic action is the highest, the player experiences a tension among all three available actions. The strength of the sentiment towards the altruistic action directly influences the likelihood of its selection. Similarly, a strong sentiment towards the inequity-averse action increases the propensity to choose this over the other options. If we label actions as 'prosocial' when they are either inequity-averse or altruistic, we can deduce that the higher the average sentiment between these two actions, the more likely a player is to make a prosocial choice.

Therefore, even without pinpointing the precise utility function, we can derive a testable hypothesis. Let us define

$$\Delta S = \begin{cases} S_{\text{half}} - S_{\text{zero}} & \text{if } S_{\text{all}} \leq S_{\text{half}} \\ \frac{S_{\text{all}} + S_{\text{half}}}{2} - S_{\text{zero}} & \text{otherwise,} \end{cases}$$

where S_{half} is the sentiment score associated with the action of 'giving half of the endowment', S_{all} is the sentiment score associated with 'giving all the endowment' and S_{zero} is the sentiment score associated with 'keeping all the endowment'. Then, we obtain the following:

Hypothesis: ΔS is positively correlated with the rate of prosocial behaviour in the dictator game.

If this hypothesis holds, it could pave the way for more intricate game-theoretical and decision-making models where sentiment-driven language-based utility functions play a pivotal role.

6. A meta-analysis of dictator game experiments

To test this hypothesis, we issued a public call on the forums of the Economic Science Association (ESA) and the Society for Judgment and Decision Making (SJDM), asking behavioural scientists to provide instructions of dictator game experiments with human participants they had conducted. We supplemented this call with manual searches of the relevant literature with the aim of collecting as many experimental instructions as possible.

Since different experimental studies may differ in many dimensions other than language (e.g. nationality of the sample, gender balance, age; all these variables have been shown to affect behaviour in the dictator game [5]), our aim is to calculate the values S_{zero}, S_{half} and S_{all} at the study level and then use these values to predict the rate of altruistic behaviour as a function of ΔS in each single study using a linear regression. Then, we will use the coefficients and standard errors of the study-level regressions to conduct a meta-analysis of all the studies. Meta-analysis is a statistical technique that synthesizes data from multiple studies to identify patterns, trends and overall effects, while also measuring heterogeneity across studies (see [46–48] for examples of meta-analyses on similar games). To estimate coefficient and standard error at the study level using a linear regression, we need studies with at least three experimental conditions because a linear regression with two data points returns no standard error since there is only one line that passes through two distinct points. We collected 12 research articles with a total of 61 experimental conditions that satisfy this requirement [17,24,49–58].

Following Rathje and colleagues [59], we employed the generative AI chatbot GPT-4 to conduct sentiment analysis. In their study, Rathje and colleagues demonstrated that GPT-4 accurately detects sentiments close to fine-tuned machine learning models. We adapted their procedure to our setting, and we prompted GPT-4 to evaluate the sentiment associated with the three prominent actions of the dictator game, namely, 'keeping all the endowment', 'keeping half of the endowment' and 'giving all the endowment'. To do so, each instruction was inputted into the chatbot with the following prompt: 'Now imagine that there is a population of 1000 people living in [country]. What do you think the average response to the following questions would be? (Please return an exact number with two decimal digits). How negative or positive is the action of [action referring to 'keeping all/keeping half/giving all the endowment'] on a 1–7 scale, with 1 being 'very negative' and 7 being 'very positive'?'. In this prompt, [action referring to . . .] was substituted with the exact words used by the authors in the specific experimental instruction, and [country] was replaced with the country where the experiment was conducted. The box below reports an example of the prompt.

We recorded numerical responses for each experimental instruction. The output of this methodology provided a sequence of AI-generated sentiment scores associated with each prominent action, which we then used to calculate the value ΔS. To prevent any learning and maintain the integrity of the analysis, we deleted the chat with GPT-4 after collecting the score for each instruction, ensuring that subsequent estimations were not influenced by previous conversations.

Table 1 reports the average sentiment score associated with each prominent action in dictator games.

On average, the sentiment score associated with the pro-self action is lower than the sentiment score associated with the prosocial actions. Moreover, the sentiment score associated with the altruistic action (i.e. 'giving all') is similar to that of the inequity-averse action (i.e. 'giving half'). However, looking at the disaggregated data, in some cases, GPT-4 returned a higher sentiment score for the altruistic action, while in other cases, it returned a higher sentiment score for the inequity-averse action, underscoring the importance of keeping the two cases separated, as done in the definition of ΔS. Not surprisingly, instead, the self-regarding action got consistently lower rates. See table 2 for the disaggregated data.

Then, we turn to the meta-analysis. Figure 1 reports the forest plot of the random-effects meta-analysis. Forest plots provide the standard way to report meta-analytic results, as they represent a complete overview of the effect sizes and confidence intervals of individual studies while allowing for the assessment of an overall summary estimate and not losing track of heterogeneity. Random-effects meta-analysis is generally preferred over fixed-effects meta-analysis in situations, like ours, where there is heterogeneity among the studies being analysed. This heterogeneity can be due to differences in study populations, methodologies, interventions or other factors that might influence the outcomes. We mention, for completeness, that the results are robust and become actually stronger when using fixed-effects meta-analysis. Therefore, the random-effects estimation represents a conservative estimation. Coming to the meta-analytic results, in line with the hypothesis, we find a significantly positive effect (overall effect size = 0.08; 95% CI = [0.04, 0.13]; $z = 3.41$; $p < 0.01$), such that higher ΔS is associated with more prosocial behaviour. We have conducted several robustness checks. When asking GPT-4 to predict the average response of 1000 people in the USA (selected because most of the training dataset originated there), the outcomes are consistent (overall effect size = 0.09; 95% CI = [0.01, 0.16]; $z = 2.22$; $p = 0.03$). Similarly, the results are stable when not specifying the sample size and requesting GPT-4 to estimate the average response of a population living in [country] (Overall effect

Table 1. Descriptive statistics of the sentiment score associated with the prominent actions in dictator games.

	S_{zero}	S_{half}	S_{all}
average	2.600	5.233	5.369
s.d.	0.627	0.929	1.010

Table 2. Descriptive statistics of the sentiment score associated with the prominent action in dictator games by experimental instructions.

experimental instruction	S_{zero}	S_{half}	S_{all}	experimental instruction	S_{zero}	S_{half}	S_{all}
Antinyan et al. [50]				Kettner & Ceccato [51]			
control	3.20	5.50	4.75	give female	4.50	6.00	5.50
loss manipulation 1	3.50	5.50	4.50	give male	4.50	6.00	5.25
loss manipulation 2	2.50	5.00	4.00	take female	2.00	3.50	5.50
Brañas-Garza [52]				take male	2.50	3.50	5.50
baseline	2.50	5.50	4.50	Kettner & Waichman [53]			
helping others	2.00	6.00	5.00	give hypothetical	4.50	5.50	5.50
reciprocity	2.50	6.50	4.50	give incentivized	4.00	5.50	5.00
Bruttel & Stolley [54]				take hypothetical	2.00	2.50	6.00
control	3.00	5.50	4.50	take incentivized	2.00	2.50	5.50
decision power condition	2.50	5.50	4.50	Kuang & Bicchieri [24]			
responsibility condition	2.50	5.50	4.50	control	2.75	5.50	6.50
Capraro & Vanzo [17]				ind. injunction, appropriate	3.25	5.50	6.75
boost condition	3.25	—	5.75	ind. injunction, approved	2.50	5.50	6.50
demand condition	2.25	—	5.75	ind. injunction, desirable	2.50	6.00	6.50
donate condition	3.25	—	5.75	ind. injunction, okay	3.25	5.75	6.50
give condition	2.75	—	5.75	ind. injunction, permissible	3.50	5.50	6.50
steal condition	1.50	—	6.50	ind. injunction, should	2.30	5.20	6.75
take condition	2.25	—	5.75	ind. injunction, the right thing	2.75	5.50	6.50
Dreber et al. [55]				Ockenfels & Werner [49]			
Exp 1 - giving informed	2.15	5.40	6.20	info condition 1	2.50	5.50	4.50
Exp 1 - giving uninformed	2.50	5.75	6.25	info condition 2	2.50	5.50	4.50
Exp 1 - taking informed	1.50	2.50	6.00	noInfo condition 1	2.50	5.50	4.00
Exp 1 - taking uninformed	1.75	4.25	6.50	noInfo condition 2	2.50	5.50	4.50
Exp 2 - giving give	2.50	5.25	6.75	Schurter & Wilson [56]			
Exp 2 - giving transfer	2.50	5.25	6.25	die roll condition	2.50	6.50	4.50
Exp 2 - keeping keep	2.45	4.50	6.25	quiz condition	2.50	6.00	2.50
Exp 2 - keeping transfer	2.50	5.00	6.00	seniority condition	2.35	6.45	4.20
Exp 3 - giving Informed	2.15	5.45	6.50	unannounced condition	2.45	6.10	2.75
Exp 3 - giving uninformed	2.50	5.00	6.50	Walkowitz [57]			
Exp 3 - taking informed	2.50	4.00	6.50	DeRo25	2.50	5.50	4.50
Exp 3 - taking uninformed	2.15	3.65	6.20	Dec50	2.25	5.75	4.75
Herne et al. [58]				N-N	2.00	5.00	4.00
baseline	2.50	5.50	5.00	N-N-2	2.50	4.50	5.50
certainty empathy	2.00	5.80	4.80	Pay50	2.50	5.50	4.50
uncertainty empathy	2.50	6.00	5.00	Rol50	2.50	5.50	4.50
uncertainty no empathy	2.20	5.80	4.40				

size = 0.08; 95% CI = [0.00, 0.15]; $z = 2.07$; $p = 0.04$). Finally, the results maintain their robustness when prompting GPT-4 with all the conditions from a given study, without restarting the chat after each condition (Overall effect size = 0.08; 95% CI = [0.03, 0.12]; $z = 3.60$; $p < 0.01$).
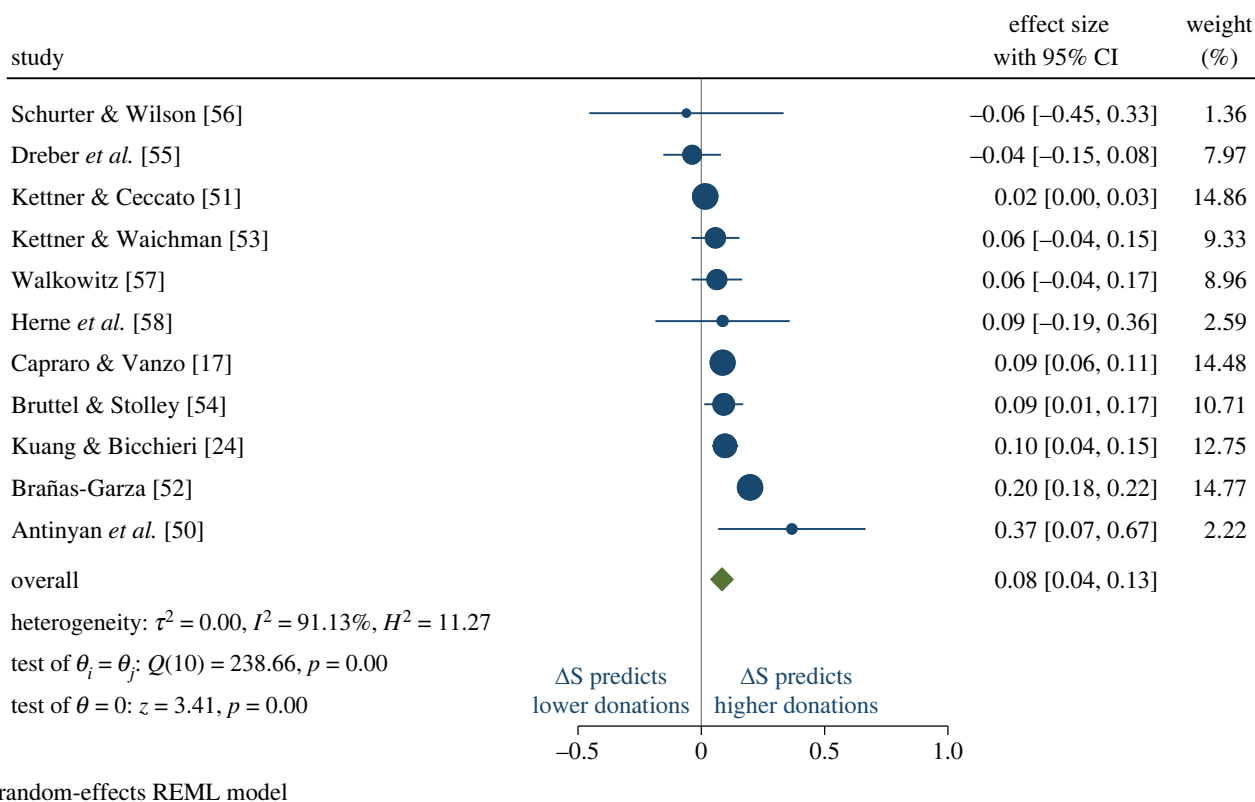
study		effect size with 95% CI	weight (%)
Schurter & Wilson [56]		–0.06 [–0.45, 0.33]	1.36
Dreber *et al.* [55]		–0.04 [–0.15, 0.08]	7.97
Kettner & Ceccato [51]		0.02 [0.00, 0.03]	14.86
Kettner & Waichman [53]		0.06 [–0.04, 0.15]	9.33
Walkowitz [57]		0.06 [–0.04, 0.17]	8.96
Herne *et al.* [58]		0.09 [–0.19, 0.36]	2.59
Capraro & Vanzo [17]		0.09 [0.06, 0.11]	14.48
Bruttel & Stolley [54]		0.09 [0.01, 0.17]	10.71
Kuang & Bicchieri [24]		0.10 [0.04, 0.15]	12.75
Brañas-Garza [52]		0.20 [0.18, 0.22]	14.77
Antinyan *et al.* [50]		0.37 [0.07, 0.67]	2.22
overall		0.08 [0.04, 0.13]	

heterogeneity: $\tau^2 = 0.00$, $I^2 = 91.13\%$, $H^2 = 11.27$

test of $\theta_i = \theta_j$: $Q(10) = 238.66$, $p = 0.00$

test of $\theta = 0$: $z = 3.41$, $p = 0.00$

ΔS predicts lower donations   ΔS predicts higher donations

–0.5   0   0.5   1.0

random-effects REML model

Figure 1. Forest plot of the meta-analysis of the sentiment associated with altruistic behaviour across the studies. One study [49] gets dropped from the meta-analysis, because GPT-4 estimates the same sentiments in all four conditions, therefore the study-level regression does not estimate the standard error. It is important to note that GPT-4 sentiment scores are in line with actual behaviour also in this study, as the mean giving did not significantly vary across conditions. Moreover, the same study does not get dropped in the robustness checks, all of which show a similar pattern of results.

7. Discussion

We are entering an era marked by the rapid advancement of artificial intelligence. The implications of this evolution are bound to be transformative, altering the very fabric of how we live, work and think. The promise of this new age is not just in the computational power of these machines but in the profound ways they are expected to blend into the human experience [60,61]. Among other things, the synergy between humans and machines holds the potential to amplify our decision-making processes. Machines, free from the cognitive limitations that humans face, may guide us to make more informed, optimal decisions, especially in critical situations where human judgement may vacillate due to stress [62], cognitive overload [63], or bias [64].

But the benefits of this symbiotic relationship come with their own set of challenges [31]. In this article, we paid attention to one particular challenge. Since humans and machines will primarily communicate through language, it becomes crucial to ensure that machines understand and process language in a way that aligns with human intent. From a game-theoretical viewpoint, this necessitates a paradigm shift from utility functions based solely on outcomes to utility functions that consider the influence of language on human preferences. Recognizing this, we advocate for the incorporation of sentiment analysis as a tool to develop language-based utility functions. Sentiment analysis could provide a quantitative measure of the language that describes possible actions in decision-making scenarios.

We have embarked on this path by leveraging the sentiment analysis capabilities of GPT-4 to shed light on human choices within the context of the dictator game. This game is emblematic of the complex interplay between self-interest and the interest of others, capturing the essence of a multitude of social interactions where one must balance personal gain against the welfare of others [5].

Our findings serve as a starting point, indicating the potential of sentiment analysis in explaining human decisions. Yet, this is merely the first step of a broader exploration. Future work should extend the application of sentiment analysis to other spheres of human interactions, including cooperation [65], honesty [66], altruistic punishment [67], trust and trustworthiness [68]. Moreover, the binary paradigm of classifying sentiments into positive or negative categories must be overcome. It is imperative to account for a spectrum of emotions and moral values, each with distinct influences on behaviour, as experimental studies have shown that different emotions can sway decisions in various ways [69–71], just as diverse ethical considerations can steer actions along different paths [72–74]. Building on the methodology outlined in this article, an initial step could be to test GPT-4's ability to reliably assess various emotions and normative dimensions, and to determine how these assessments might explain behavioural patterns. Alternatively, refined content analysis tools could be used or developed for this purpose. For instance, several sentiment analysis tools have been developed to identify and quantify a large set of emotions [42,75], while tools for identifying and quantifying diverse moral values have started to emerge in recent years [76,77]. Furthermore, the exploration should not stop at the opaque algorithms of GPT-4. The predictive capacity of various sentiment analysis tools should be examined, dissecting the 'black box' to understand the mechanics of language's influence on decision-making. This will be fundamental for developing and testing a specific utility

function that mathematically formalizes the complex interplay between language, emotions, norms and behaviour. Finally, the integration of sentiment analysis and language-based utility functions into game theory, as proposed in this study, opens up exciting possibilities for its combination with evolutionary game theory and mathematical modelling. Evolutionary game theory has significantly contributed to understanding the development of moral behaviours, such as cooperation [78–81], trust and trustworthiness [82], and honesty [83,84]. By incorporating sentiment analysis into the modelling of player strategies, we can capture a more nuanced understanding of human decision-making behaviour that extends beyond outcome-based preferences. For example, the use of sentiment analysis could help model the evolution of cooperation or honesty in a population, where the language used to describe actions could influence the perceived utility of those actions and thus the evolution of strategies. Mathematically, these language-based utility functions could be integrated into the replicator dynamics equations used in evolutionary game theory, adding a new dimension to these models.

In summary, our aspiration is that this article lays the groundwork for a novel approach in game theory, one that recognizes the importance of language in decision-making processes. The journey ahead is filled with important questions necessitating dedicated research. The answers we find, and the questions we ask, will shape the future not only of game-theoretical research but also of the very nature of the relationship between humans and machines.

References

1. Von Neumann J, Morgenstern O. 1944 *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.
2. Camerer CF. 2011 *Behavioral game theory: experiments in strategic interaction*. Princeton, NJ: Princeton University Press.
3. Kahneman D, Knetsch JL, Thaler RH. 1986 Fairness and the assumptions of economics. *J. Bus.* **59**, S285–S300. (doi:10.1086/296367)
4. Forsythe R, Horowitz JL, Savin NE, Sefton M. 1994 Fairness in simple bargaining experiments. *Games Econ. Behav.* **6**, 347–369. (doi:10.1006/game.1994.1021)
5. Engel C. 2011 Dictator games: a meta study. *Exp. Econ.* **14**, 583–610. (doi:10.1007/s10683-011-9283-7)
6. Ledyard JO. 1995 Public goods: a survey of experimental research. In *The handbook of experimental economics* (eds JH Hagel, AE Roth), pp. 111–194. Princeton, NJ: Princeton University Press.
7. Levine DK. 1998 Modeling altruism and spitefulness in experiments. *Rev. Econ. Dyn.* **1**, 593–622. (doi:10.1006/redy.1998.0023)
8. Fehr E, Schmidt KM. 1999 A theory of fairness, competition, and cooperation. *Q. J. Econ.* **114**, 817–868. (doi:10.1162/003355399556151)
9. Bolton GE, Ockenfels A. 2000 ERC: a theory of equity, reciprocity, and competition. *Am. Econ. Rev.* **91**, 166–193. (doi:10.1257/aer.90.1.166)
10. Charness G, Rabin M. 2002 Understanding social preferences with simple tests. *Q. J. Econ.* **117**, 817–869. (doi:10.1162/003355302760193904)
11. Engelmann D, Strobel M. 2004 Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *Am. Econ. Rev.* **94**, 857–869. (doi:10.1257/0002828042002741)
12. Capraro V, Perc M. 2021 Mathematical foundations of moral preferences. *J. R. Soc. Interface* **18**, 20200880. (doi:10.1098/rsif.2020.0880)
13. Capraro V, Halpern JY, Perc M. In press. From outcome-based to language-based preferences. *J. Econ. Lit.*
14. Liberman V, Samuels SM, Ross L. 2004 The name of the game: predictive power of reputations versus situational labels in determining Prisoner's Dilemma game moves. *Pers. Soc. Psychol. Bull.* **30**, 1175–1185. (doi:10.1177/0146167204264004)
15. Eriksson K, Strimling P, Andersson PA, Lindholm T. 2017 Costly punishment in the ultimatum game evokes moral concern, in particular when framed as payoff reduction. *J. Exp. Soc. Psychol.* **69**, 59–64. (doi:10.1016/j.jesp.2016.09.004)
16. Capraro V, Rand DG. 2018 Do the right thing: experimental evidence that preferences for moral behavior, rather than equity or efficiency per se, drive human prosociality. *Judgment Dec. Mak.* **13**, 99–111. (doi:10.1017/S1930297500008858)
17. Capraro V, Vanzo A. 2019 The power of moral words: loaded language generates framing effects in the extreme dictator game. *Judgment Dec. Mak.* **14**, 309–317. (doi:10.1017/S1930297500004356)
18. Capraro V, Jordan JJ, Tappin BM. 2021 Does observability amplify sensitivity to moral frames? Evaluating a reputation-based account of moral preferences. *J. Exp. Soc. Psychol.* **94**, 104103. (doi:10.1016/j.jesp.2021.104103)
19. Chang D, Chen R, Krupka E. 2019 Rhetoric matters: a social norms explanation for the anomaly of framing. *Games Econ. Behav.* **116**, 158–178. (doi:10.1016/j.geb.2019.04.011)
20. Huang L, Lei W, Xu F, Yu L, Shi F. 2019 Choosing an equitable or efficient option: a distribution dilemma. *Soc. Behav. Personal.: An Int. J.* **47**, 1–10. (doi:10.2224/sbp.8559)
21. Huang L, Lei W, Xu F, Liu H, Yu L, Shi F, Wang L. 2020 Maxims nudge equitable or efficient choices in a Trade-Off Game. *PLoS ONE* **15**, e0235443. (doi:10.1371/journal.pone.0235443)
22. Mieth L, Buchner A, Bell R. 2021 Moral labels increase cooperation and costly punishment in a Prisoner's Dilemma game with punishment option. *Sci. Rep.* **11**, 10221. (doi:10.1038/s41598-021-89675-6)
23. Tappin BM, Capraro V. 2018 Doing good vs. avoiding bad in prosocial choice: a refined test and extension of the morality preference hypothesis. *J. Exp. Soc. Psychol.* **79**, 64–70. (doi:10.1016/j.jesp.2018.06.005)

24. Kuang J, Bicchieri C. 2024 Language matters: how normative expressions shape norm perception and affect norm compliance. *Phil. Trans. R. Soc. B* **379**, 20230037. (doi:10.1098/rstb.2023.0037)

25. Capraro V, Vanzo A, Cabrales A. 2022 Playing with words: do people exploit loaded language to affect others' decisions for their own benefit? *Judgment Dec. Mak.* **17**, 50–69. (doi:10.1017/S1930297500009025)

26. Ścigała KA, Zettler I, Pfattheicher S, Capraro V. 2022 Corrupting the prosocial people: does cooperation framing increase bribery engagement among prosocial individuals? Stage 1 Registered Report. *PsyArXiv preprint.* (https://osf.io/preprints/psyarxiv/gfah8/)

27. Bubeck S et al. 2023 Sparks of artificial general intelligence: early experiments with gpt-4. *arXiv.* (http://arxiv.org/abs/2303.12712)

28. Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S. 2018 Human decisions and machine predictions. *Q. J. Econ.* **133**, 237–293. (doi:10.3386/w23180)

29. Mullainathan S, Obermeyer Z. 2022 Diagnosing physician error: a machine learning approach to low-value health care. *Q. J. Econ.* **137**, 679–727. (doi:10.1093/qje/qjab046)

30. Sunstein CR. 2023 The use of algorithms in society. *Rev. Austr. Econ.* 1–22. (doi:10.1007/s11138-023-00625-z)

31. Capraro V et al. 2023 The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PsyArXiv preprint.* (https://osf.io/preprints/psyarxiv/6fd2y)

32. Novelli C, Taddeo M, Floridi L. 2023 Accountability in artificial intelligence: what it is and how it works. *AI & Soc.* 1–12. (doi:10.1007/s00146-023-01635-y)

33. Vasconcelos H, Jörke M, Grunde-McLaughlin M, Gerstenberg T, Bernstein MS, Krishna R. 2023 Explanations can reduce overreliance on AI systems during decision-making. *Proc. ACM Hum. Comput. Interact.* **7**, 1–38. (doi:10.1145/3579605)

34. Barocas S, Selbst AD. 2016 Big data's disparate impact. *California Law Review* **104**, 671–732.

35. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. 2021 On the dangers of stochastic parrots: can language models be too big? In *Proc. of the 2021 ACM Conf. on fairness, accountability, and transparency, virtual event, Canada, 3–10 March,* pp. 610–623. New York, NY: ACM. (doi:10.1145/3442188.3445922)

36. Sweeney L. 2013 Discrimination in online ad delivery. *Commun. ACM* **56**, 44–54. (doi:10.1145/2447976.2447990)

37. Floridi L et al. 2018 An ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Mach.* **28**, 689–707. (doi:10.1007/s11023-018-9482-5)

38. Hagendorff T. 2020 The ethics of AI ethics: an evaluation of guidelines. *Minds Mach.* **30**, 99–120. (doi:10.1007/s11023-020-09517-8)

39. Rahwan I et al. 2019 Machine behaviour. *Nature* **568**, 477–486. (doi:10.1038/s41586-019-1138-y)

40. Pang B, Lee L. 2008 Opinion mining and sentiment analysis. *Found. Trends® Inf. Retrieval* **2**, 1–135. (doi:10.1561/1500000011)

41. Baccianella S, Esuli A, Sebastiani F. 2010 Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Proc. of the 7th Int. Conf. on Language Resources and Evaluation (LREC'10), Valletta, Malta, May,* vol. 10, pp. 2200–2204. (http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf)

42. Mohammad SM, Turney PD. 2013 Crowdsourcing a word–emotion association lexicon. *Comput. Intell.* **29**, 436–465. (doi:10.1111/j.1467-8640.2012.00460.x)

43. Haidt J. 2012 *The righteous mind: why good people are divided by politics and religion.* New York, NY: Vintage.

44. Graham J, Haidt J, Nosek BA. 2009 Liberals and conservatives rely on different sets of moral foundations. *J. Pers. Soc. Psychol.* **96**, 1029. (doi:10.1037/a0015141)

45. Curry O, Mullins D, Whitehouse H. 2019 Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. *Curr. Anthropol.* **60**, 47–69. (doi:10.1086/701478)

46. Rand DG. 2016 Cooperation, fast and slow: meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychol. Sci.* **27**, 1192–1206. (doi:10.1177/0956797616654455)

47. Rand DG, Brescoll VL, Everett JA, Capraro V, Barcelo H. 2016 Social heuristics and social roles: intuition favors altruism for women but not for men. *J. Exp. Psychol.: General* **145**, 389. (doi:10.1037/xge0000154)

48. Capraro V. 2024 The dual-process approach to human sociality: meta-analytic evidence for a theory of internalized heuristics for self-preservation. *J. Pers. Soc. Psychol.* (doi:10.2139/ssrn.3409146)

49. Ockenfels A, Werner P. 2012 'Hiding behind a small cake' in a newspaper dictator game. *J. Econ. Behav. Organ.* **82**, 82–85. (doi:10.1016/j.jebo.2011.12.008)

50. Antinyan A, Corazzini L, Fišar M, Reggiani T. 2024 Lying on networks: the role of structure and topology in promoting honesty. *J. Econ. Behav. Organ.* **218**, 599–612. (doi:10.1016/j.jebo.2023.12.024)

51. Kettner SE, Ceccato S. 2014 Framing matters in gender-paired dictator games. University of Heidelberg, Department of Economics, Discussion Paper Series No. 557. See https://d-nb.info/1192448294/34.

52. Brañas-Garza P. 2007 Promoting helping behavior with framing in dictator games. *J. Econ. Psychol.* **28**, 477–486. (doi:10.1016/j.joep.2006.10.001)

53. Kettner SE, Waichman I. 2016 Old age and prosocial behavior: social preferences or experimental confounds? *J. Econ. Psychol.* **53**, 118–130. 10.1016/j.joep.2016.01.003

54. Bruttel L, Stolley F. 2018 Gender differences in the response to decision power and responsibility—framing effects in a dictator game. *Games* **9**, 28. (doi:10.3390/g9020028)

55. Dreber A, Ellingsen T, Johannesson M, Rand DG. 2013 Do people care about social context? Framing effects in dictator games. *Exp. Econ.* **16**, 349–371. (doi:10.1007/s10683-012-9341-9)

56. Schurter K, Wilson BJ. 2009 Justice and fairness in the dictator game. *Southern Econ. J.* **76**, 130–145. (doi:10.4284/sej.2009.76.1.130)

57. Walkowitz G. 2019 On the validity of probabilistic (and cost-saving) incentives in dictator games: a systematic test. Available at SSRN 3068380.

58. Herne K, Hietanen JK, Lappalainen O, Palosaari E. 2022 The influence of role awareness, empathy induction and trait empathy on dictator game giving. *PLoS ONE* **17**, e0262196. (doi:10.1371/journal.pone.0262196)

59. Rathje S, Mirea DM, Sucholutsky I, Marjieh R, Robertson C, Van Bavel JJ. 2023 GPT is an effective tool for multilingual psychological text analysis. *PsyArXiv preprint.* (https://osf.io/preprints/psyarxiv/sekf5/)

60. Bostrom N. 2014 *Superintelligence: paths, dangers, strategies.* Oxford, UK: Oxford University Press.

61. Tegmark M. 2018 *Life 3.0: being human in the age of artificial intelligence.* New York, NY: Vintage.

62. Starcke K, Brand M. 2012 Decision making under stress: a selective review. *Neurosci. Biobehav. Rev.* **36**, 1228–1248. (doi:10.1016/j.neubiorev.2012.02.003)

63. Phillips WJ, Fletcher JM, Marks AD, Hine DW. 2016 Thinking styles and decision making: a meta-analysis. *Psychol. Bull.* **142**, 260. (doi:10.1037/bul0000027)

64. Kahneman D, Slovic P, Tversky A. 1982 *Judgment under uncertainty: heuristics and biases.* Cambridge, UK: Cambridge University Press.

65. Rapoport A, Chammah AM. 1965 *Prisoner's dilemma: a study in conflict and cooperation,* vol. 165. Ann Arbor, MI: University of Michigan Press.

66. Gneezy U. 2005 Deception: the role of consequences. *Am. Econ. Rev.* **95**, 384–394. (doi:10.1257/0002828053828662)

67. Güth W, Schmittberger R, Schwarze B. 1982 An experimental analysis of ultimatum bargaining. *J. Econ. Behav. Organ.* **3**, 367–388. (doi:10.1016/0167-2681(82)90011-7)

68. Berg J, Dickhaut J, McCabe K. 1995 Trust, reciprocity, and social history. *Games Econ. Behav.* **10**, 122–142. (doi:10.1006/game.1995.1027)

69. Motro D, Ordóñez LD, Pittarello A, Welsh DT. 2018 Investigating the effects of anger and guilt on unethical behavior: a dual-process approach. *J. Bus. Ethics* **152**, 133–148. (doi:10.1007/s10551-016-3337-x)

70. Strohminger N, Lewis RL, Meyer DE. 2011 Divergent effects of different positive emotions on moral judgment. *Cognition* **119**, 295–300. (doi:10.1016/j.cognition.2010.12.012)

71. Ugazio G, Lamm C, Singer T. 2012 The role of emotions for moral judgments depends on the type of emotion and moral scenario. *Emotion* **12**, 579. (doi:10.1037/a0024611)

72. Day MV, Fiske ST, Downing EL, Trail TE. 2014 Shifting liberal and conservative attitudes using moral foundations theory. *Pers. Soc. Psychol. Bull.* **40**, 1559–1573. (doi:10.1177/0146167214551152)

73. Kidwell B, Farmer A, Hardesty DM. 2013 Getting liberals and conservatives to go green: political ideology and congruent appeals. *J. Consumer Res.* **40**, 350–367. (doi:10.1086/670610)

74. Amin AB, Bednarczyk RA, Ray CE, Melchiori KJ, Graham J, Huntsinger JR, Omer SB. 2017 Association of moral values with vaccine hesitancy. *Nat. Hum. Behav.* **1**, 873–880. (doi:10.1038/s41562-017-0256-5)

75. Hartmann J. 2022 Emotion English DistilRoBERTa-base. See https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/.

76. Hoover J *et al.* 2020 Moral foundations twitter corpus: a collection of 35k tweets annotated for moral sentiment. *Soc. Psychol. Personal. Sci.* **11**, 1057–1071. (doi:10.1177/1948550619876629)

77. Trager J *et al.* 2022 The moral foundations reddit corpus. *arXiv*. (http://arxiv.org/abs/2208.05545)

78. Wang Z, Jusup M, Wang RW, Shi L, Iwasa Y, Moreno Y, Kurths J. 2017 Onymity promotes cooperation in social dilemma experiments. *Sci. Adv.* **3**, e1601444. (doi:10.1126/sciadv.1601444)

79. Wang Z, Jusup M, Shi L, Lee JH, Iwasa Y, Boccaletti S. 2018 Exploiting a cognitive bias promotes cooperation in social dilemma experiments. *Nat. Commun.* **9**, 2954. (doi:10.1038/s41467-018-05259-5)

80. Wang Z *et al.* 2020 Communicating sentiment and outlook reverses inaction against collective risks. *Proc. Natl Acad. Sci. USA* **117**, 17650–17655. (doi:10.1073/pnas.1922345117)

81. Hua S, Hui Z, Liu L. 2023 Evolution of conditional cooperation in collective-risk social dilemma with repeated group interactions. *Proc. R. Soc. B* **290**, 20230949. (doi:10.1098/rspb.2023.0949)

82. Kumar A, Capraro V, Perc M. 2020 The evolution of trust and trustworthiness. *J. R. Soc. Interface* **17**, 20200491. (doi:10.1098/rsif.2020.0491)

83. Capraro V, Perc M, Vilone D. 2019 The evolution of lying in well-mixed populations. *J. R. Soc. Interface* **16**, 20190211. (doi:10.1098/rsif.2019.0211)

84. Capraro V, Perc M, Vilone D. 2020 Lying on networks: the role of structure and topology in promoting honesty. *Phys. Rev. E* **101**, 032305. (doi:10.1103/PhysRevE.101.032305)

85. Capraro V, Di Paolo R, Perc M, Pizziol V. 2024 Language-based game theory in the age of artificial intelligence. *OSF*. (https://osf.io/mx5w3/)