Experimental Economics (2025), **28**, 1085–1100
doi:10.1017/eec.2025.10038

ORIGINAL PAPER

Integrating machine behavior into human subject experiments: a user-friendly toolkit and an application to framed prisoner's dilemmas

Christoph Engel[1], Max Rainer Pascal Großmann[2] and Axel Ockenfels[1,3]

[1]Max Planck Institute for Behavioral Economics, Bonn, Germany
[2]Department of Economics, Faculty of Business and Economics, The University of Melbourne, Carlton VIC, Australia
[3]Department of Economics and Adenauer School of Government, University of Cologne, Cologne, Germany
Corresponding author: Axel Ockenfels; Email: ockenfels@uni-koeln.de

(Received 14 February 2024; revised 4 November 2024; accepted 8 December 2025)

Abstract
Large Language Models (LLMs) have the potential to profoundly transform and enrich experimental economic research. We propose a new software framework, "alter_ego", which makes it easy to design experiments between LLMs and to integrate LLMs into oTree-based experiments with human subjects. Our toolkit is freely available at github.com/mrpg/ego. To illustrate, we run differently framed prisoner's dilemmas with interacting machines as well as with human-machine interaction. Framing effects in machine-only treatments are strong and similar to those expected from previous human-only experiments, yet less pronounced and qualitatively different if machines interact with human participants.

Keywords: framing; human-machine interaction; large language models; software for experiments
JEL Codes: C91; C92; D91; O33; L86

1. Introduction

1.1. Research question

Experiments are the central tool for understanding human behavior. While the vast majority of experimental studies in behavioral economics have focused on human interactions, the rise of human–machine and machine–machine interactions across nearly all facets of social and economic life has spurred researchers to incorporate machines into behavioral studies (Caro et al., 2022). Large Language Models (LLMs) have been particularly transformative, enabling innovative experimental methodologies and communication approaches. Yet integrating LLMs as agents into experiments requires technical skills that experimenters do not necessarily possess. This paper aims to lower the technical barrier to entry by introducing a suite of tools that facilitate implementation.

For microexperiments, a ready-made tool requires no coding at all: our "builder" enables researchers to create simple LLM-to-LLM experiments by filling out designated fields. For richer designs, we provide full native Python code that experimenters can adapt to their specific needs. Finally, we offer an integration of LLM agents into experiments run on the popular experimental software oTree (Chen et al., 2016). This version of the code enables human participants to interact with LLM agents.

© The Author(s), 2026. Published by Cambridge University Press on behalf of the Economic Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/eec.2025.10038 Published online by Cambridge University Press

1.2. Large language models

Large language models have revolutionized many areas, including feature extraction (Dominguez-Olmedo et al., 2024), code writing (Mohsin et al., 2024), and text analysis (Törnberg, 2023). They build on the transformer architecture (Vaswani et al., 2017), but add a generative component: they can not only analyze but also generate text (or pictures, or sound) (for a technical primer see Naveed et al., 2023). Their generative ability makes LLMs appealing for experimental research.

1.3. LLM agents

Humans can delegate actions to LLMs, rendering machine behavior a worthy topic in its own right (Rahwan et al., 2019). One may, for instance, investigate LLMs' motivation (Guo, 2023; Phelps & Russell, 2023), willingness to cooperate (Kasberger et al., 2024), or adaptation to a changing environment (Chen et al., 2023). Another important topic is human behavior when individuals know they are interacting with an algorithm (Crandall et al., 2018; March, 2021; Chugunova & Sele, 2022; von Schenk et al., 2025). Prior research has shown that, in cooperation problems like the prisoner's dilemma, humans respond to the perceived cooperativeness of the AI agent (Ng, 2023), but tend to cooperate less with machines, even when they expect machines to cooperate (de Melo et al., 2016; Crandall et al., 2018; Karpus et al., 2021; Jiang et al., 2025), but only in Western countries (Karpus et al., 2025).

The implications for experimental economics would be even more profound if LLMs reliably replicate human behavior (Charness et al., 2023; Horton, 2023; Mei et al., 2024). LLMs could then be used not only to pilot, replicate, or potentially substitute human participants in standard economic experiments (Aher et al., 2023; Brookins & DeBacker, 2023; Tsuchihashi, 2023), but also as proxies for human choices in macroeconomic simulations (Li et al., 2024), political science (Argyle et al., 2023), and the social sciences more broadly (Manning et al., 2024).

1.4. Getting at variance

In the tradition of experimental economics, individual human participants are randomly exposed to alternative conditions. The manipulation precisely matches a hypothesis derived from (perhaps behavioral) theory. Experimenters usually provisionally accept the theoretical hypothesis if the average reaction of treated participants differs significantly from that of untreated participants. As LLMs capitalize on centuries of human utterances and respond using human language, it is meaningful to ask the equivalent question: how do LLMs "behave" when exposed to the same stimuli?

Yet under the hood, an LLM is a prediction engine. Given textual input, it determines the best fitting textual output. LLMs are probabilistic by design. They do not calculate a response based on first principles embedded in the program. Rather, they attempt to make sense of the input—the prompt—as best they can. Researchers can reconstruct the degree of uncertainty that the LLM faces for a given prompt by asking the same question repeatedly.

This approach, however, presupposes that the LLM allows for sufficient variance in possible responses to the prompt. Generative Pretrained Transformer (GPT) is the model family of LLMs developed by OpenAI. Users may define `temperature` as a free parameter. If `temperature` $= 0$, the LLM gives the exact same response for every repetition;[1] the globally best fitting reaction. Setting the temperature higher and running multiple repetitions generates a distribution of responses. If `temperature` is set to 1, in expectation the distribution of multiple responses replicates the probability of each response estimated by the model. This represents the machine's estimate of how human participants would react to the manipulation (Chen et al., 2023; Guo, 2023).

[1]Almost—the richness of the model causes some variance in repeated queries even if `temperature` is set to 0, but normally this variance is small.

https://doi.org/10.1017/eec.2025.10038 Published online by Cambridge University Press

1.5. Human alignment

How good a proxy are LLMs for human choices? This is an active research field (see the survey by Wang et al., 2023). Most economics-related literature has focused on the models offered by OpenAI. Thus far, the evidence is mixed. Davinci-002 yields comparable results for the ultimatum game, garden path tasks and the Milgram shock experiments, but not for wisdom of the crowd tasks (Aher et al., 2023). GPT-3[2] exhibits anchoring effects similar to the ones observed in humans (Jones & Steinhardt, 2022), is subject to gender stereotypes (Acerbi & Stubbersfield, 2023), and falls prey to intuition in cognitive reflection tests in about the same way as humans (Hagendorff et al., 2023).

GPT-3.5 exhibits moral judgements that are similar to the ones observed in human subjects (Dillion et al., 2023), and emulates the choices well that human proposers make in the ultimatum game (Kitadai et al., 2023). However, in GPT-3.5 cognitive biases are less pronounced than in humans (Hagendorff et al., 2023). GPT-3.5 is better than human participants at applying Bayes' rule2023, and is less likely to overvalue the difference between two options presented simultaneously (Orsini, 2023). The model does not capture well the choices of human responders in the ultimatum game (Kitadai et al., 2023). It is even less patient than human participants (Goli & Singh, 2024). Finally, on multiple tasks GPT-3.5 exhibits a "correct answer bias," such that it almost always gives the majority response, even if tested multiple times; the variance observed among human participants in analogous tasks is suppressed (Park et al., 2024). GPT-4 is about as good as human annotators with the classification of text data from lab experiments (Celebi & Penczynski, 2024). The model exhibits risk preferences, time preferences and social preferences that are qualitatively similar to the ones observed in human subjects, but they are more extreme (Chen et al., 2023; Capraro et al., 2023; Goli & Singh, 2024). On multiple rationality axioms, GPT-4 outperforms human subject pools (Raman et al., 2024).

Ultimately, even if the alignment between LLM and human choices is pronounced—at least in well-defined subdomains of experimental research—researchers may ask how much experimental economics can learn from probing LLMs. Will LLMs, at best, inform experimental economics about what was already known? We offer two counter-arguments. Even if LLMs only replicate human thoughts and actions, this differs from documenting human behavior under controlled experimental conditions. At the very least, with the help of LLMs, human cognitive or motivational tendencies that are implicit in written language could be explicated and rigorously tested. Moreover, much creativity can be traced to previously unknown combinations of known pieces of knowledge. It is conceivable that, with appropriate experimental designs, such hidden human capabilities and behavioral tendencies can be made visible. Hence arguably, while LLMs critically depend on the human legacy, the human legacy does not limit them to a degree that makes investigating their behavior trivial and uninteresting. A promising proof of concept would be experimental designs that have never been tested on human subjects. These designs could first be run with LLMs (with the results temporarily kept confidential). In the next step, the otherwise identical design could be run with human participants. If results are similar (and statistically indistinguishable), this would provide strong additional evidence.[3]

Large language models might thus become a powerful resource for experimental economics, but much more data is needed to assess the actual predictive value of these models. The next step in this endeavor has recently begun: systematically manipulating the degrees of freedom that large language models offer, such as the degree of variance in outcomes (*temperature*, see Kitadai et al. (2024) on the ultimatum game), the minimum cumulative posterior probability for a response to be considered (`top_p`), the size of the most likely set of tokens from which the response is taken (`top_k`), and the many different ways of prompting the model (for an overview see Chang et al. (2024)).

[2]"GPT" is an acronym for *Generative Pretrained Transformer*, as a characterization of the three key features of the model.

[3]We are grateful to an anonymous referee for suggesting this avenue for future research.

https://doi.org/10.1017/eec.2025.10038 Published online by Cambridge University Press

1.6. Our suite of tools for LLM experiments

However, a significant barrier to rapid progress in this burgeoning field—and to equal opportunities among researchers across the world—is the absence of tools that enable behavioral scientists to easily develop experiments based on well-established norms and standards in experimental economics. To catalyze this exciting research area, we developed an open-source toolkit, "alter_ego," that greatly facilitates experiments in which participants are emulated by LLMs. The user-friendly design of our tool facilitates swift and efficient data collection. The software is described in the next section and is freely available at github.com/mrpg/ego.

1.7. Application: a framed prisoner's dilemma

In the final section, we employ our toolkit to examine framing effects in machine–machine and machine–human interactions in pre-registered prisoner's dilemmas. We are not the first to have GPT play prisoner dilemmas (Akata et al., 2023; Bauer et al., 2023; Brookins & DeBacker, 2023; Duffy et al., 2022; Guo, 2023; Phelps & Russell, 2023), but no previous study has focused on framing and group affiliation.[4] These are particularly interesting and relevant research questions as large *language* models condense the power of human language. With the help of our framing manipulation, we learn in which ways and to what degree an LLM is swayed by the particular words used to represent a social conflict. Although narrow conceptions of "intelligent" or "rational" behavior assert that a game's framing should not influence machine or human decisions (Russell, 2019), humans are known to respond behaviorally to even subtle differences in framing (e.g., Dufwenberg et al., 2011). Game descriptions provide contextual cues that influence perceptions of appropriate and normative behavior. Do machines, which capitalize on the richness and sophistication of human language, pick up and respond to such cues? Our data show that the answer is yes.

From a technological perspective, framing machines can be viewed as an instance of prompt engineering (Chen et al., 2025; Gu et al., 2023; White et al., 2023). In this sense, our application provides a link between the computer science literature on prompt engineering and the experimental economics literature on how framing affects behavior.

Second, a critical question for understanding human–machine interaction is whether and when within-group machine–machine and human–human interactions elicit different responses than between-group machine–human interactions (von Schenk et al., 2025; Verma et al., 2023), a research question that our tool can help resolve. Indeed, our data suggest that, while machines and humans may exhibit similar behavioral patterns when interacting only within their respective groups, behavior changes when opponents do not share one's group affiliation.

1.8. Organization of paper

The next section dives deeper into the technical background, and introduces our suite of tools that make it easy for experimenters to tap into the power of LLMs (Section 2). Section 3 presents the design of the framed prisoner's dilemma, both among two LLM agents, and in interaction between an LLM agent and a human participant. Section 4 concludes with discussion.

2. A toolkit for machine–machine and machine-human experiments

Our toolkit is designed to make it easy for experimenters to integrate LLM agents into their experiments. The kit consists of the following tools:

[4]As we discuss below, our results are consistent with previous studies that found that GPT is noticeably cooperative when interacting with itself (Brookins & DeBacker, 2023) and that machine behavior is conditional on the opponent's strategy (Duffy et al., 2022). For a framing effect in a very different context see Jones & Steinhardt (2022). Capraro et al. (2024) test the power of sentiment analysis on LLM choices.

https://doi.org/10.1017/eec.2025.10038 Published online by Cambridge University Press

- As a starter kit, we offer an easy-to-use shorthand version of the tool, allowing teachers and researchers to quickly run experiments with LLMs.
- For a broad class of experiments, experimenters can use our builder, a web application that generates most of the code for them.
- For experimenters who want greater flexibility, we provide a full-fledged Python library, which they can adapt to their specific design needs.
- This more complex version of the tool is also usable if experimenters want to have LLM agents interact with human participants. For such designs, our tool is integrated with the popular experimental software `oTree`.

We introduce these tools in turn, but first have to explain LLM access via an application programming interface API.

2.1. The necessity of using an Application Programming Interface

LLMs usually offer an application programming interface (API). Using the API, the experimenter may fully define the process, store all prompts, repeat the same prompt as often as needed for the research question, and store the resulting data in a format suitable for data analysis (using the preferred statistical package).

While the code needed for running such an experiment is not excessively complex (and we offer such code in one version of our tool), the need to write Python code, assign treatments, dynamically generate prompts, retrieve and filter responses, and perhaps make them accessible in prompts, is a barrier to using LLMs as experimental participants. This motivates the design of our tool. Our package makes these steps as easy as possible.

To use any part of our toolkit, an experimenter needs to install Python 3.8 or later (`www.python.org`). Most LLM providers additionally require that the user register and obtain an API key (authorizing the user to access the API). Our software comes in the form of the Python package alter_ego, which can be installed into Python from PyPI: *alter_ego_llm*. Complete documentation is available at github.com/mrpg/ego. To lower the barriers of access as much as possible, we have posted a series of videos:

- Microexperiments: https://youtu.be/GPc0a-Fg1bY
- Builder: https://youtu.be/tV5xACU-abw
- Using Python directly: https://youtu.be/WHW0gkT-oHE
- Integration with oTree: https://youtu.be/ouxRFdKOGEw

2.2. Microexperiments

The tool for running microexperiments on an LLM is the most accessible. Users can vary parameters as in a factorial design. If they are interested in variance (e.g., as a proxy for confidence), they can ask the same question multiple times. To illustrate this design option, we ask GPT-4 about the estimated approval of two US presidents, at the beginning and in the end of their presidency (Table 1). This feature of alter_ego may also be relevant for teaching purposes, an application of LLMs that has been highlighted and increasingly deployed (Cowen & Tabarrok, 2023). Classroom applications could involve the interactive generation of data that is subsequently analyzed.

As Figure 1 shows, with alter_ego the amount of code needed for this purpose is minimal. Users must import three aspects of our package (lines 1-3). The function in lines 5-6 specifies that GPT-4 shall be used and that the model is allowed a high degree of variability (`temperature = 1`). Lines 8-9 are the functional equivalent of experimental instructions. The two terms enclosed in double braces also define the "treatments": which president and which time? This yields a 2×2 factorial design. Note that this allows for dynamic prompts, a theme that will recur later. The concluding

https://doi.org/10.1017/eec.2025.10038 Published online by Cambridge University Press

```python
import alter_ego.agents
from alter_ego.utils import extract_number
from alter_ego.experiment import factorial

def agent():
  return alter_ego.agents.GPTThread(model="gpt-4", temperature=1.0)

prompt = "Estimate the public approval rating of {{politician}} during
↪  the {{time}} of their presidency. Only return a single percentage
↪  from 0 to 100."

data = factorial(
  prompt,
  politician=["George W. Bush", "Barack Obama"],
  time=["1st year", "8th year"]
).run(agent, extract_number, times=5)
```

Figure 1. Complete code for a machine microexperiment

Table 1. Results of the code in Figure 1 (`temperature` = 1, 5 repetitions)

`politician`	`time`	GPT's mean estimate of approval
Barack Obama	1st year	57.2
Barack Obama	8th year	55.0
George W. Bush	1st year	57.6
George W. Bush	8th year	29.0

block of code defines the actual experiment: (a) all possible combinations of the parameters are to be tested (`factorial`), (b) which persons and which time periods are of interest. The last line calls the `agent` function, specifies that only numbers in the output shall be reported, and defines the number of repetitions.[5] Table 1 reports the results (from 5 independent draws).

2.3. Designing LLM experiments through a web application

Every researcher who has used oTree to program an experiment knows that coding from scratch is challenging. Even experienced oTree users cannot directly implement experiments with LLM participants. While alter_ego offers an all-Python solution for advanced purposes (explained in later sections), we provide an easier alternative for experiments exclusively involving LLM participants: our builder. The builder is available at https://ego.mg.sb/builder/.

The builder requires minimal Python knowledge compared to traditional microexperiment code (see Figure 1). To run an experiment, a single terminal command suffices: `ego run built`. The experimental design is imported from `built.json`.[6] The experimenter does not need to write this JSON code manually. Using the "Export or import scenario" functionality, experimenters can export their design as JSON or import existing configurations to modify them. This bidirectional approach facilitates collaborative research and iterative refinement.

[5] If users want to postprocess the resulting data, they can use the "list of dicts" returned by `factorial`, a standard way to exchange data between packages for data analysis.

[6] The code generated on the website must be copied into a text file named `built.json` and stored in the directory from which the program is started.

https://doi.org/10.1017/eec.2025.10038 Published online by Cambridge University Press

When the builder's capabilities suffice, designing an experiment consists of filling out fields on the web app. The following elements can be configured:

- Participants (`Threads`)
  - Number of participants
  - LLM model for each participant
- Treatments
  - Random assignment of Threads to treatments
- Rounds
  - Currently supports partner design
- Variables across treatments
  - Treatment-conditional instructions
  - Choice variables
  - Framing variations
  - Payoff structures
- Instructions (prompts)
  - System prompts (initialize the conversation)
  - User prompts (per round)
  - All prompts can be conditional on treatment, role, and other variables
- Response processing (filters)
  - Extract relevant data from LLM responses for export

Two features make the builder particularly convenient. First, prompts support dynamic content through Jinja2 templating.[7] This allows prompts to reference variables, treatments, and other participants' choices. For example, the prompt template `Welcome, your name is {{name }}, and you are playing with {{other.name }}` automatically personalizes for each participant. For Alice playing with Bob, this becomes "Welcome, your name is Alice, and you are playing with Bob."

Second, filters process LLM responses into analyzable data. Instead of exporting verbose raw responses, experimenters can extract numbers, match predefined strings (like "ACCEPT" or "REJECT"), or parse structured formats. This ensures data is immediately suitable for statistical analysis.

Experiments built with the builder include automatic CSV export. The command `ego data built [experiment-id] > data.csv` exports all experimental data, including choices, treatment assignments, and response metadata. Both the builder interface and our tutorial videos provide additional details.

2.4. Coding with Python

If experimenters want to design experiments requiring even more flexibility than offered by our builder, they can directly code the experiment in Python while still exploiting the capabilities offered by alter_ego. In the companion video to this section (https://youtu.be/WHW0gkT-oHE), we explain step by step how the example experiment used to illustrate the capability of the builder can be coded manually.[8] Experimenters with more extended Python experience can also use this tool to carry the output forward to a Python package for data analysis such as Pandas (`pandas.pydata.org`) or Polars (www.pola.rs).

To better understand the architecture of the tool, it is useful to consider Figure 2: Each participant is represented by a `Thread`. One interaction of multiple participants constitutes a `Conversation`.

[7] The syntax is explained on the website. See palletsprojects.com/p/jinja/ for documentation.

[8] We focus on the LLM-LLM version of the experiment.

https://doi.org/10.1017/eec.2025.10038 Published online by Cambridge University Press

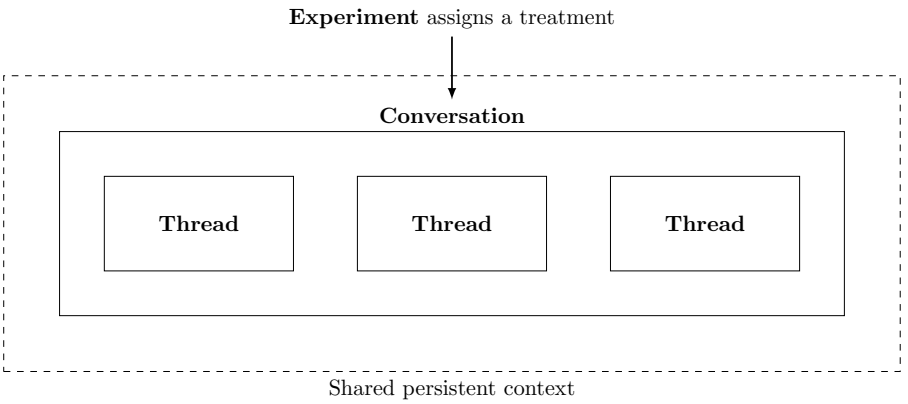


Figure 2. Architecture of the tool—these elements represent Python classes

The tool makes it possible to assemble multiple instances of interaction in an `Experiment`, which assigns a `treatment` to the `Conversation`.

2.5. Human–machine interaction: integrating oTree

An important frontier of experimental research is the interaction between humans and machines, not least because machines impact ever more parts of social life. alter_ego makes such experiments possible by introducing LLM functionality into the oTree framework (Chen et al., 2016). As this experimental software is widely used, many users will already be at least somewhat familiar with oTree.

Users who are fluent with oTree may find it appealing if the implementation of data generation happens within the oTree environment, even if using that environment would not be strictly necessary. For such users, we also provide the code for a simple oTree app that has a human chat with GPT (https://github.com/mrpg/ego/tree/master/otree/ego_chat). In computer science parlance, we provide a "façade" (e.g., Gamma et al., 1994, sec 4.5) from oTree to alter_ego.

3. Putting the tool to good use: Do machines react to framing?

In this section, we report on an experiment that puts our tool to good use. We begin with the research question. We test two versions of the game. With the first experimental design, we investigate whether LLMs are subject to framing effects when interacting with another instance of the LLM. In the second design, we let one LLM agent interact with one human participant.

3.1. Research question: Are large language models subject to framing?

A robust experimental literature demonstrates that human participants are sensitive to framing: results systematically differ depending on how the same incentive structure is presented (Kühberger, 1998; Levin et al., 1998; Dreber et al., 2013). The power of LLMs originates in the richness of human language. It is therefore conceivable that LLM choices also depend on how a choice problem is presented to them. On the other hand, a main goal in the continuous improvement of large language models is making them more "accurate" (see only OpenAI, 2023, 3-6), including the hope that they might outperform human competitors.[9] A side effect of these efforts at improving language models might be that they are less sensitive to framing than humans.

[9] As Russell (2019) puts it: "artificial intelligence has been mainly about working out the details of how to build rational machines." Indeed, on many benchmarks, GPT-4 is at the highest percentile of human results.

https://doi.org/10.1017/eec.2025.10038 Published online by Cambridge University Press

Table 2. Payoffs

	cooperate	defect
cooperate	20, 20	0, 28
defect	28, 0	8, 8

Arguably, the effect of framing results from contextual cues that trigger descriptive and normative beliefs (and beliefs about beliefs) about behavior (Dufwenberg et al., 2011). Beliefs matter in prisoner's dilemma games because many human players choose to cooperate if they believe that the opponent cooperates, and because, in sequential or repeated game contexts, selfish players may have an incentive to trigger positive reciprocity from other players if they are believed to be conditionally cooperative (Ockenfels, 1999; Fischbacher et al., 2001). If such beliefs are affected by framing, we expect to see different cooperation rates across our conditions with machine participation. To test machines' responsiveness to framing, we adapt an experiment that one of the authors has run with human participants. In that sequential prisoner's dilemma, cooperation was significantly reduced when the game was framed as one of *competition* and, non-significantly, when the game was framed as facing a joint *enemy* (Engel & Rand, 2014).

Specifically, we implemented a 2x2 sequential prisoner's dilemma (Bolle & Ockenfels, 1990; Clark & Sefton, 2001; Ahn et al., 2007) with binary action space and payoffs as in Table 2:

Following Engel and Rand (2014), the game was presented sequentially, and repeated 10 times, which was commonly known. We manipulated the frame: neutral ("In this experiment, you are together with another participant …"), joint enemy ("you and another participant … have a joint enemy"), and competition ("you are competing against another participant"; see Section B.3 in the Appendix for full instructions).

Our null hypothesis, H_0, is based on subgame perfect equilibrium predictions for rational and selfish players and predicts no cooperation across all conditions. However, as outlined above, machines and humans are known to cooperate, and humans are known to respond to framing. Consequently, with our alternative hypothesis H_1, we predict a positive degree of cooperation that differs depending on the framing of the social conflict.[10]

3.2. Machines interacting with machines

For the experiments that test LLMs, in a second dimension, we manipulated the LLM platform, using either GPT-3.5 (turbo) or GPT-4. Framing effects are influenced by factors that can vary between different "social" groups that do or do not share the same understanding of contextual cues, or in their degree of rationality or selfishness. Yet as both GPT-3.5 and GPT-4 capitalize on the same training data,[11] we do not expect differences when either two instances of GPT-3.5 or two instances of GPT-4 interact with each other. This is why our alternative hypothesis, H_2, is that, while cooperation is possible, the impact of framing does not systematically differ across the two different versions of GPT, our equivalent of a population difference.

[10]For the experiment in which GPT interacts with GPT, we had additionally preregistered this hypothesis: "If a 2×2 prisoner's dilemma with binary action space is framed such that gains from cooperation only obtain in expectation, not necessarily in realization, there is less cooperation." As we have simplified the design in the GPT-human conditions, we do not report these results in the body of the paper. They are available in the Appendix for the "procurement" condition.

[11]Both models use training data up to September 2021.

https://doi.org/10.1017/eec.2025.10038 Published online by Cambridge University Press

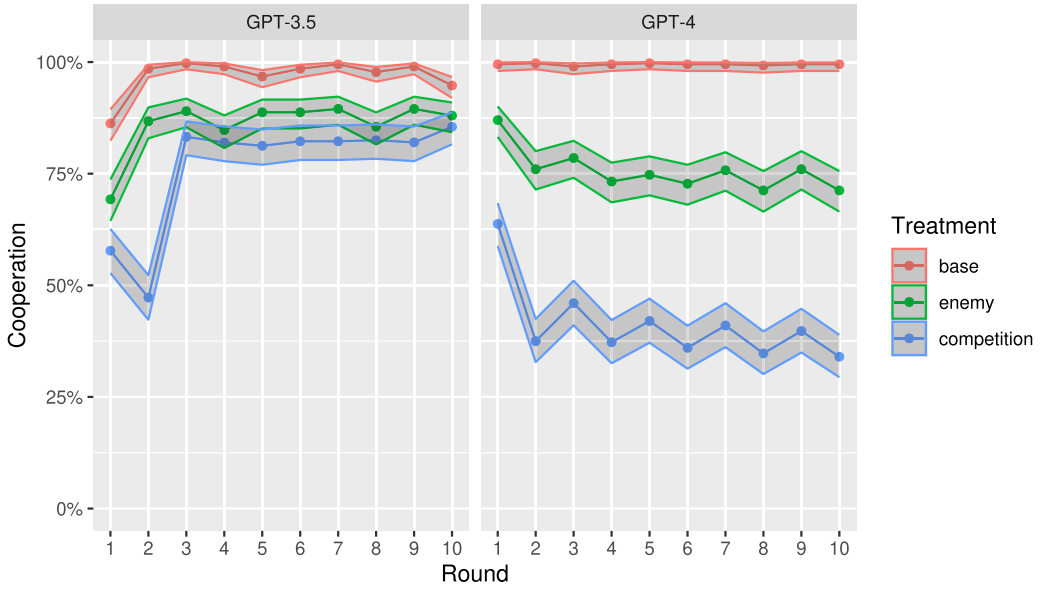


Figure 3. Cooperation conditional on platform and round
Mean choices, with 95% confidence interval.

We had 200 groups of 2 instances of GPT, respectively, interacting over 10 periods.[12] For the reasons explained above, we set `temperature` to 1, to generate a distribution of responses.

Before we report results, we elaborate on the technical implementation, which is the main reason for writing this paper. We have chosen the design of the (machine–machine version of the) experiment such that it can be implemented with our builder. In Section B.2 of the Appendix, we provide a URL to the resulting code. Interested readers can run the experiment by putting the file `built.json` into their current working directory and running the experiment from the terminal using `ego run built`. The program will randomly assign one of the three frames. The data will be stored in folder `out` under the computer-generated ID of the current run. Experimenters with deeper knowledge of Python may be interested in exploring the alternative version of the machine-only experiment that we programmed manually, which is available on GitHub.[13]

Figure 3 shows significant machine cooperation, which rejects H_0, as expected. On neither platform (GPT-3.5 or GPT-4) and in no frame is the cooperation rate of GPT interacting with another instance of GPT close to 0. In the Appendix (Table 4), we report the upper and lower limits for cooperation rates per condition and round that we cannot exclude at the 5% level. The lower level is never below 30%.

Figure 3 shows that framing matters: when the game is framed as either jointly protecting against an enemy or, in particular, as competing with each other, GPT cooperates substantially less with another machine compared to the (unframed) base treatment.[14] Table 3 provides summary statistics by platform and frame. Against our expectation, machine platform matters: while the ranking of

[12]Further features of the design, the full set of preregistered hypotheses along with the corresponding results, and links to the preregistrations are all reported in the Appendix. We have tested seven more frames in the machine–machine conditions, yet do not report them in the main text because of space restrictions. Results from these additional conditions are also available in the Appendix. Including those in our main text would not alter our main conclusions.

[13]For generating the data reported in this section, we used the latter version because we only developed the builder at a later stage of our project.

[14]We refrain from offering tests for the comparison between the original human–human (Engel & Rand, 2014) and the machine–machine results, as the original experiment had not been repeated, so that results are not fully comparable.

https://doi.org/10.1017/eec.2025.10038 Published online by Cambridge University Press

Table 3. Mean percentage of cooperative choices per platform and frame, aggregated over all rounds

	GPT-3.5	GPT-4
Baseline	97.0	99.5
Enemy	86.0	75.6
Competition	76.6	41.2

cooperation across frames is stable, the framing effects are more pronounced when implementing the experiment on GPT-4.[15] Providing a conclusive explanation for this unexpected result is beyond the scope of this paper. But in retrospect, it seems to resonate with changes OpenAI advertised when launching GPT-4 (OpenAI, 2023): the newer model is meant to be more accurate than GPT-3.5 (p. 3), less subject to hindsight bias (p. 4), closer to prevalent results from microeconomics (p. 5), and more likely to suppress sensitive or even disallowed prompts (p. 14). The stronger response to our frames could therefore result from making text output more "normative."[16]

3.3. Machines interacting with human participants

In the human–machine version of our experiment, machines are first-movers and humans are second-movers. For the LLM agents, we used GPT-4. Hence in this experiment, we have only three conditions, one with the *base*, the *enemy*, and the *competition* frame, respectively. We had 96 groups without a frame (*base*), 106 groups with the *enemy* frame, and 102 groups with the *competition* frame. This experiment was conducted in English at the Cologne Laboratory for Economic Research in August 2023. Humans were incentivized while machines were not. We discuss machine incentives in our concluding section. Lab participants were invited to the experiment knowing that it would take approximately 15 minutes and that they would be allowed to start the experiment at any time between 10 AM and 2 PM on a freely chosen day of the experiment. All participants were confirmed students at universities in Cologne; their field of study varies widely. The modal field of study was business administration, and the modal year of birth was 1999. All rounds were paid. The final payment ranged from €2.20 to €6.28 (average: €4.09). At the time of the experiment, the local minimum wage was €12 per hour, or €3 per 15 minutes.

As explained in Section 2, to test human–machine interaction, one must integrate Python with oTree.[17] The experiment was conducted entirely online and remotely. All code for the experiment is available at our GitHub repository; see the Appendix for all instructions. Note that subjects received information about their history of play with their assigned LLM.[18]

As LLMs capitalize on human language and experience, with our hypothesis H_3, we do not expect differences in how machine and human cooperation respond to framing.

Figure 4 reports cooperation rates. Comparing Figure 3 with the left panel of Figure 4 immediately shows that adding human participants diminishes cooperation, which is strongly confirmed statistically.[19] As the middle panel of Figure 4 shows, when interacting with humans, machine cooperation rates of (machine) first movers are still relatively high in the first round, yet they react strongly when the human counterpart defects, as many do (right panel of Figure 4): In the *baseline*, 40.6% of human

[15] Statistical tests are in the Appendix, including a discussion of their interpretation given that the data come from machine choices.

[16] This explanation would of course imply that, by making the LLM more sensitive to "normative" expectations, it has also become more liable to bias resulting from overinterpreting situational cues.

[17] Experimenters who are proficient in oTree may also choose to implement machine–machine experiments within oTree using the human–machine interaction tools provided in our toolkit.

[18] Further details about the preregistration are in the Appendix.

[19] See Table 7 in the Appendix.

https://doi.org/10.1017/eec.2025.10038 Published online by Cambridge University Press

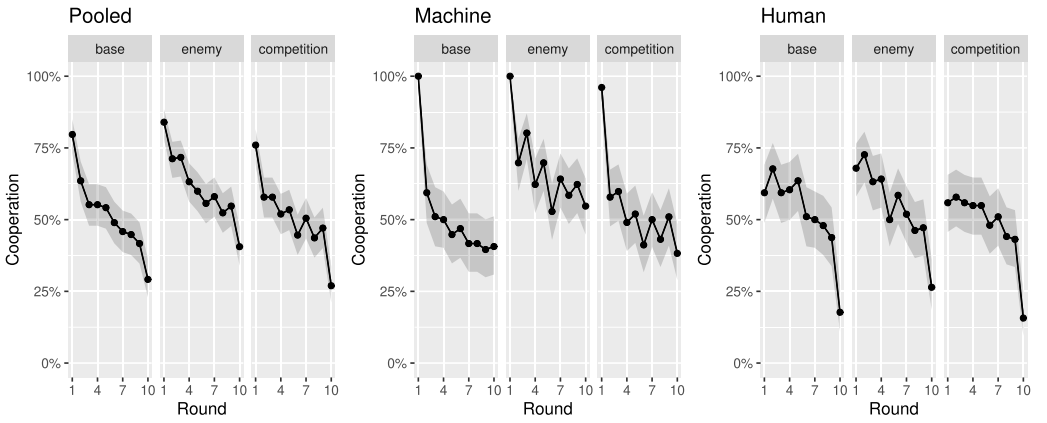


Figure 4. Cooperation conditional on platform, round, and identity of the player (machine vs. human)
Mean choices, with 95% confidence interval

participants defect in the first round; in the *enemy* condition, 32.1% do; and in the *competition* condition, 44.1% do. Machines reciprocate defection and, as a result, human defectors end up with much smaller payoffs than they could have earned by cooperating: the payoff of human participants who cooperated (defected) in the first round was 169 (102) in *baseline*, 176 (160) in *enemy*, and 174 (118) in *competition*. All differences are significant.[20]

The impact of framing is much less pronounced in the mixed group, and now the enemy frame triggers the highest cooperation rate. This framing effect is driven by machine choices (Table 4 in the Appendix). Arguably, the human–machine interaction itself is a powerful frame that dominates other, more subtle details of the game's presentation. A possible explanation is that, in human–machine interactions, there is a stronger mismatch in commonly shared norms, leading to greater uncertainty about what can be expected from each other, mitigating the framing effects that we see within groups of either only machines or only humans.

4. Discussion

LLMs have the potential to profoundly change, substantially enrich, and radically facilitate experimental economics research. Yet to fully leverage this potential, researchers need a toolkit that is easy and free to use, based on well-established norms and standards in experimental economics, that can be tailored to almost all specific tasks of interactive decision-making among machines, and that can be used for experiments in which human participants interact with machines. Providing such a toolkit is the main contribution of this paper. Our tool allows researchers to efficiently sample LLMs. Two very accessible and intuitive versions of our tool empower experimenters with little Python experience to run a wide variety of machine–machine experiments (Sections 2.2 and 2.3). For experimenters with greater Python versatility, we offer an even more flexible version of the tool (Section 2.4). Finally, we integrate our tool with oTree (Section 2.5), which is particularly appealing for experimenters who want to test human–machine interaction.

Our illustrative experiment provides important insights into machine behavior and human–machine interaction. We find strong framing effects in machine-only treatments that are partly similar to those expected from previous human-only treatments, yet they tend to be even more pronounced among machines. Perhaps surprisingly, framing effects are less pronounced and qualitatively

[20] t-tests, *base*: $p < 0.001$; *enemy*: $p = 0.026$; *competition*: $p < 0.001$. See Figure 1 and Table 3 in the Appendix for further results.

https://doi.org/10.1017/eec.2025.10038 Published online by Cambridge University Press

different when machines interact with human participants. We find that machines respond very sensitively to human defection and that many humans fail to anticipate that machines punish exploitative strategies. This suggests that there is a mismatch in what these different classes of actors expect from each other, making coordination on a shared norm more difficult.

Understanding that framing matters differently in machine–human versus machine–machine and human–human interaction is crucial for interpreting experiments involving human–machine interactions. It seems promising to integrate different numbers of machines, from zero to n, as occurs in the field, to study such effects. We provide the toolkit for this endeavor.

One important line of future research is machine incentives. In experiments with human subjects, preferences are typically induced through monetary incentives (Smith, 1976). However, machines, including GPT, operate based on an objective function defined during training, making it difficult to financially incentivize them in any given experiment due to the absence of personal desires such as money, prestige, or other human rewards in machines. Johnson and Obradovich (2022) attempted to address this by compensating the parent company OpenAI according to machine behavior. However, this cannot affect machines the way money affects humans, and it remains untested whether this approach actually influences machine preferences and, if so, how.

An alternative would be to explicitly instruct machines to maximize their game payoff. If such preference induction were successful, we would perhaps learn something about the machine's computational capability and its beliefs about how humans respond to machine behavior, but we could not learn anything about how the machine would naturally behave across game framings, which is our research question. Similarly, while we could in principle also guide machine behavior through fine-tuning or simulated environments, we were interested in GPT's "genuine" choices based on its knowledge at the onset of the experiment, not in what we can train it to do. Of course, it is well known that any version of GPT is heavily fine-tuned before release using a technique called "Reinforcement Learning from Human Feedback" (e.g., OpenAI, 2023). That said, our tool could be used to study the effectiveness and impact of various approaches to incentivize machines. For instance, do machines exhibit more or less care when a human, charity, political party, or the parent company is compensated on the machine's behalf? This research will be essential for understanding the role of incentives for machine behavior in various applications requiring interactions with humans or other machines.

While large language models are a promising new tool in the kit of experimental economics, professional standards for their use must still be developed. As our framing application shows, results may vary substantially depending on the choice of language model (GPT-3.5 vs. GPT-4 in our application). Very likely, results are sensitive not only to framing (which we tested) but also to other, seemingly more minor, differences in prompts. Other model parameters (like temperature) are also likely to matter. Such sensitivity not only calls for preregistrations that include all model parameters. One must also worry about the fact that most large language models (including the currently most powerful GPT model) are proprietary. There is no guarantee that older models will remain available for an extended period. Even if most of the training data is frozen (as in the two large language models that we tested), these models are meant to learn over time from the queries of their users.[21] For all these reasons, results generated with the help of (these)[22] large language models are strictly speaking not replicable.

Our framing results, in conjunction with other studies, indicate that in certain domains, machine behavior might help experimenters build hypotheses about human behavior. Should this prove robustly true in some domains, easily implementable pilot experiments with machines might offer a cost-effective and efficient method to guide and inform subsequent human subject research.

[21] OpenAI, however, stipulates in its terms and conditions that it will only use input through its chat protocol for training. As we used the API, we can be reasonably confident that GPT will not use our own data for training, so that replication remains meaningful. But later attempts at running the same code are likely to lead to different results, as GPT has learned from the chat input it has received in the meantime.

[22] This should be different when using open-source models like Llama3. But as yet, the open-source models are less accurate.

https://doi.org/10.1017/eec.2025.10038 Published online by Cambridge University Press

Leveraging the potential predictive power of machines could help guide the choice and design of human subject experiments, ultimately leading to more robust and generalizable findings. Our toolkit facilitates the implementation of even large-scale endeavors involving thousands of player roles across dozens of experimental settings that differ substantially in complexity of interaction, making it easily accessible and available to everyone, paving the way for new discoveries that deepen our understanding of both human and machine behavior.

Data availability statement. The replication material for the study is available at https://doi.org/10.7910/DVN/SPJFJ5 (Harvard Dataverse).

Our code is available at github.com/mrpg/ego.

Acknowledgements. We thank Heinrich Nax, Christoph Schottmüller, Tobias Werner, our beta-testers, referees, and audiences in Cologne, Erfurt and Pisa for helpful comments, and Simon Weidtmann for excellent research assistance. All remaining errors are our own. Funding by the German Research Foundation (DFG) under Germany's Excellence Strategy (EXC 2126/2 – 390838866) is gratefully acknowledged.

Competing interests. The authors have no competing interests to declare that are relevant to the content of this article.

Ethical standards. Ethics approval was obtained by the Ethics Council of the Max Planck Society within the framework of the *General Approval for Procedures, Experiments and Projects Following the Protocol that is Standard in Experimental Economics.*

References

Acerbi, A., & Stubbersfield, J. M. (2023). Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, *120*(44), e2313790120.

Aher, G. V., Arriaga, R.I and Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In *40th International Conference on Machine Learning*. Honolulu, Hawaii, USA 202, 337–371.

Ahn, T. -K., Lee, M., Ruttan, L., & Walker, J. (2007). Asymmetric payoffs in simultaneous and sequential prisoner's dilemma games. *Public Choice*, *132*(3/4), 353–366.

Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., & Schulz, E. (2023). Playing repeated games with large language models *Nature Human Behaviour*, (9), 1380–1390.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351.

Bauer, K., Liebich, L., Hinz, O., & Kosfeld, M. (2023). Decoding GPT's hidden "rationality" of cooperation. *SAFE Working Paper Series*, 401.

Bolle, F., & Ockenfels, P. (1990). Prisoners' dilemma as a game with incomplete information. *Journal of Economic Psychology*, *11*(1), 69–84.

Brookins, P., & DeBacker, J. M. (2023). Playing games with GPT: What can we learn about a large language model from canonical strategic games? *Economics Bulletin*, *44*(1), 25–37.

Capraro, V., Di Paolo, R., Perc, M., & Pizziol, V. (2024). Language-based game theory in the age of artificial intelligence. *Journal of the Royal Society Interface*, *21*(212), 20230720.

Capraro, V., Di Paolo R., & Pizziol V. (2023). Assessing large language models'ability to predict how humans balance self-interest with the interest of others. *Sci Rep*, *15*, 21428.

Caro, F., Colliard, J. -E., Katok, E., Ockenfels, A., Stier-Moses, N., Tucker, C., & Wu, D. (2022). Call for papers—management science special issue on the human-algorithm connection. *Management Science*, *68*(1), 7–8.

Celebi, C. Penczynski, S. (2024). *Using large language models for text classification in experimental economics* (Technical Report). Technical report. School of Economics, University of East Anglia, Norwich, UK.

Chang, K., Xu S., Wang C., Luo Y., Xiao T., & Zhu J. (2024). Efficient prompting methods for large language models: A survey. *arXiv preprint arXiv:2404.01077*.

Charness, G., Jabarian, B., & List, J. A. (2023). *Generation Next: Experimentation with AI* (NBER Working Paper No. 31679). National Bureau of Economic Research. https://doi.org/10.3386/w31679

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2025). Unleashing the potential of prompt engineering in large language models. *Patterns*, *6*(6), 101260. https://doi.org/0.1016/j.patter.2025.101260

Chen, D. L., Schonger, M., & Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97.

Chen, Y., Liu, T. X., Shan, Y., & Zhong, S. (2023). The emergence of economic rationality of gpt. *Proceedings of the National Academy of Sciences*, *120*(51), e2316205120.

Chugunova, M., & Sele, D. (2022). We and it: An interdisciplinary review of the experimental evidence on how humans interact with machines. *Journal of Behavioral and Experimental Economics*, *99*, 101897.

Clark, K., & Sefton, M. (2001). The sequential prisoner's dilemma: evidence on reciprocation. *The Economic Journal*, *111*(468), 51–68.
Cowen, T., & Tabarrok, A. T. (*2023*). How to learn and teach economics with large language models, including GPT. *GMU Working Paper*, 23–18. https://doi.org/10.2139/ssrn.4391863
Crandall, J. W., Tennom, M. O., Ishowo-Oloko, F., Abdallah, S., Bonnefon, J. -F., Cebrian, M., Shariff, A., Goodrich, M. A., & Rahwan, I. (2018). Cooperating with machines. *Nature Communications*, *9*, 233.
de Melo, C. M., Marsella, S. and Gratch, J. (2016). People do not feel guilty about exploiting machines. In *ACM Transactions on Computer-Human Interaction*. ACM, Volume 23, pp. 1–17.
Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600.
Dominguez-Olmedo, R., Nanda, V., Abebe, R., Bechtold, S., Engel, C., Frankenreiter, J., Gummadi, K., Hardt, M., & Livermore, M. (2024). Lawma: The power of specialization for legal tasks. *arXiv preprint arXiv:2407.16615*.
Dreber, A., Ellingsen, T., Johannesson, M., & Rand, D. G. (2013). Do people care about social context? Framing effects in dictator games. *Experimental Economics*, *16*(3), 349–371.
Duffy, J., Hopkins, E., & Kornienko, T. (2022) Facing the grim truth: Repeated prisoner's dilemma against robot opponents (Working Paper). University of Edinburgh. https://homepages.econ.ed.ac.uk/~hopkinse/grim.pdf
Dufwenberg, M., Gächter, S., & Hennig-Schmidt, H. (2011). The framing of games and the psychology of play. *Games and Economic Behavior*, *73*(2), 459–478.
Engel, C., & Rand, D. G. (2014). What does "clean" really mean? the implicit framing of decontextualized experiments. *Economics Letters*, *122*(3), 386–389.
Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, *71*(3), 397–404.
Gamma, E., Helm, R., Johnson, R. and Vlissides, J. (1994). *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley Professional Computing series. Reading, Mass, Addison-Wesley.
Goli, A., & Singh, A. (2024). Frontiers: Can large language models capture human preferences? *Marketing Science*, *43*(4), 709–722.
Gu, J., Han, Z., Chen, S., Beirami, A., He, B., Zhang, G., Liao, R., Qin, Y., Tresp, V., & Torr, P. (2023). A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*.
Guo, F. (2023). Gpt agents in game theory experiments. *arXiv preprint arXiv:2305.05516*.
Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, *3*(10), 833–838.
Horton, J. J. (2023). *Large language models as simulated economic agents: What can we learn from homo silicus?* (Working Paper No. 31122). National Bureau of Economic Research. https://doi.org/10.3386/w31122
Jiang, G., Wang, Y., & Hui, P. (2025). Experimental exploration: Investigating cooperative interaction behavior between humans and large language model agents. arXiv–2503*arXiv e-prints*.
Johnson, T., & Obradovich, N. (2022). Measuring an artificial intelligence agent's trust in humans using machine incentives. *arXiv preprint arXiv:2212.13371*.
Jones, E., & Steinhardt, J. (2022). Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems*, 856.
Karpus, J., Krüger, A., Verba, J. T., Bahrami, B., & Deroy, O. (2021). Algorithm exploitation: humans are keen to exploit benevolent AI. *iScience*, *24*(6), 102679.
Karpus, J., Shirai, R., Verba, J. T., Schulte, R., Weigert, M., Bahrami, B., Watanabe, K., & Deroy, O. (2025). Human cooperation with artificial agents varies across countries. *Sci Rep*, *15*(1), 10000.
Kasberger, B., Martin, S., Normann, H. -T., & Werner, T. (2024). Algorithmic cooperation. *CESifo Working Paper*, 11124. https://doi.org/10.2139/ssrn.4855849
Kitadai, A., Lugo, S. D. R., Tsurusaki, Y., Fukasawa, Y., & Nishino, N. (2024). Can AI with high reasoning ability replicate human-like decision making in economic experiments? *arXiv preprint arXiv:2406.11426*.
Kitadai, A., Tsurusaki, Y., Fukasawa, Y. and Nishino, N. (2023). Toward a novel methodology in economic experiments: Simulation of the ultimatum game with large language models. In *2023 IEEE International Conference on Big Data (BigData)*. Sorrento, Italy, IEEE, pp. 3168–3175.
Kühberger, A. (1998). The influence of framing on risky decisions: A meta-analysis. *Organizational Behavior and Human Decision Processes*. *75*(1), 23–55.
Levin, P.I, Schneider, S. L., & Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes*, *76*(2), 149–188.
Li, N., Gao, C., Li, M., Li, Y., & Liao, Q. (2024). EconAgent: Large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Vol. 1). Bangkok, Thailand.
Manning, B. S., Zhu, K., & Horton, J. J. (2024). *Automated Social Science: Language Models as Scientist and Subjects* (Working Paper No. 32381). National Bureau of Economic Research. https://doi.org/10.3386/w32381

https://doi.org/10.1017/eec.2025.10038 Published online by Cambridge University Press

March, C. (2021). Strategic interactions between humans and artificial intelligence: Lessons from experiments with computer players. *Journal of Economic Psychology*, *87*, 102426.
Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, *121*(9), e2313925121.
Mohsin, A., Janicke, H., Wood, A., Sarker, I. H., Maglaras, L., & Janjua, N. (2024). Can we trust large language models generated code? A framework for in-context learning, security patterns, and code evaluations across diverse LLMs. *arXiv preprint arXiv:2406.12513*.
Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023). A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.
Ng, Y. (2023). When communicative AIs are cooperative actors: a prisoner's dilemma experiment on human–communicative artificial intelligence cooperation. *Behaviour & Information Technology*, *42*(13), 2141–2151.
Ockenfels, A. (1999). *Fairness, Reziprozität und Eigennutz*. Tübingen: Mohr Siebeck.
OpenAI (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
Orsini, E. (2023). University of Aberdeen, unpublished Ph.D. thesis *Do Cognitive Biases Persist in Large Language Models?*
Park, P. S., Schoenegger, P., & Zhu, C. (2024). Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, *56*, 5754–5770.
Phelps, S., & Russell, Y. I. (2023). Investigating emergent goal-like behaviour in large language models using experimental economics. *arXiv preprint arXiv:2305.07970*.
Rahwan, I, Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. -F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, D.I, Jackson, M. O., & Jennings, N. R. (2019). Machine behaviour. *Nature*, *568*(7753), 477–486.
Raman, N. K., Lundy, T., Amouyal, S. J., Levine, Y., Leyton-Brown, K. and Tennenholtz, M. (2024). STEER: Assessing the economic rationality of large language models. *Proceedings of Machine Learning Research*, 235, 42026–42047.
Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. New York, Penguin Books.
Smith, V. L. (1976). Experimental economics: Induced value theory. *The American Economic Review*, *66*(2), 274–279.
Törnberg, P. (2023). How to use LLMs for text analysis. *arXiv preprint arXiv:2307.13106*.
Tsuchihashi, T. (*2023*). *Do AIs dream of homo economicus? Answers from ChatGPT*. https://doi.org/10.2139/ssrn.4495500
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.
Verma, M., Bhambri, S. and Kambhampati, S. (2023). Preference proxies: Evaluating large language models in capturing human preferences in human-AI tasks. In *ICML 2023 Workshop on Theory of Mind in Communicating Agents*. Honolulu, Hawaii, USA.
von Schenk, A., Klockmann, V., & Köbis, N. (2025). Social preferences toward humans and machines: A systematic experiment on the role of machine payoffs. *Perspectives on Psychological Science*, *20*(1), 165–181.
Wang, Y., Zhong, W., Li, L., Mi, F., Zeng, X., Huang, W., Shang, L., Jiang, X., & Liu, Q. (2023). Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.
White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv preprint arXiv:2302.11382*.

Cite this article: Engel, C., Großmann, M. R. P., & Ockenfels, A. (2025). Integrating machine behavior into human subject experiments: a user-friendly toolkit and an application to framed prisoner's dilemmas. *Experimental Economics*, 28(5), 1085–1100. https://doi.org/10.1017/eec.2025.10038

https://doi.org/10.1017/eec.2025.10038 Published online by Cambridge University Press