University of East Anglia

SCHOOL OF ECONOMICS

cbess

Centre for Behavioural and Experimental Social Science

11th Annual Behavioural Game Theory Workshop

Thursday 10th July &
Friday 11th July 2025

University of East Anglia
SCHOOL OF ECONOMICS

cbess
Centre for Behavioural and Experimental Social Science

Welcome!

The UEA School of Economics and the Centre for Behavioural and Experimental Social Science (CBESS) at the University of East Anglia welcome you to the 11th edition of UEA's annual workshop on Behavioural Game Theory.

Our theme this year is large language models (LLMs), including studies in which subjects interact with LLMs, studies that use LLMs to analyse the data, and studies that make similar use of machine learning.

Please join us on Zoom using the following links:

Thursday: (Link tbc)

Friday: (Link tbc)

Table of Contents

If you have any questions or encounter any difficulties joining the zoom meeting, please contact us either at eco.reception@uea.ac.uk or via the phone on +44 (0)1603 592941

Schedule - Thursday 10th July All times in GMT+1 (UK Time)	
12:50 - 13:00	**Zoom Link Opens for Participants**
13:00 - 13:05	**Opening of Workshop**
13:05 - 13:45	**Daniela Grieco** - University of Milano Creativity and AI View the full paper
13:45 - 14:25	**Gavin Kader** - Southwestern University of Finance and Economics The Emergence of Strategic Reasoning of Large Language Models View the full paper
14:25 - 15:05	**Can Celebi** - University of Vienna TBC
15:05 - 15:30	**Break**
15:30 - 16:10	**Toshihiko Hirasawa** - UCLA Using Big Data and Machine Learning to Uncover How Players Choose Mixed Strategies View the full paper
16:10 - 16:50	**Husnain Ahmad** - Sewanee: the University of the South Practice makes perfect, but context gets you the win: Mixed strategy play in the lab View the full paper
16:50 - 17:30	**Wenzhuo Xu** - Carnegie Mellon University The Impact of Naturalistic and Familiar Contexts on Strategic Sophisticationwin: Mixed strategy play in the lab
17:30	**Close of session**

University of East Anglia
SCHOOL OF ECONOMICS

cbess
Centre for Behavioural and Experimental Social Science

Schedule - Friday 11th July All times in GMT+1 (UK Time)	
12:50 - 13:00	**Zoom Link Opens for Participants**
13:00 - 13:05	**Opening of Workshop Day 2**
13:05 - 13:45	**Daniel Martin** - UCSB When AI Judges Your Work: The Hidden Costs of Algorithmic Oversight
13:45 - 14:25	**Lisa Bruttel** - University of Potsdam Communication and Collusion in Oligopoly Experiments: A Meta-Study using Machine Learning
14:25 - 15:05	**Sergio Pirla** - University of Zaragoza Impact of H-1B Visa Policies: Experimental Evidence View the full paper
15:05 - 15:30	**Break**
15:30 - 16:10	**Chris Santos-Lang** - Independent Researcher MAD Chairs: A new tool to evaluate AI View the full paper
16:10 - 16:50	**Alex Brown** - Texas A&M University Incentive Mechanisms for AI: Theory and Evidence View the full paper
16:50 - 17:30	**Samuel Taylor** - University of California, San Diego; Cognitive Science Department Evaluating Instrumentally Rational Deception by LLMs in 2x2 Signaling Games View the full paper

Practice makes perfect, but context gets you the win: Mixed strategy play in the lab

Husnain Ahmad - Sewanee: University of the South

Lab play in simple games of mixed strategy offer puzzling deviations from theoretical predictions. Previous work has explored the role of expertise in seeking to understand why these deviations occur, finding that experts are more likely to conform with theoretical predictions in the field, but not in the lab. The literature suggests context as the likely reason for this, and in this paper we test for this. Students in a lab play a simplified 2 person zero-sum game that mimics penalty kicks in football/soccer. The experiment varies the level of context provided to subjects. We find evidence that context affects the behaviour of players, making them play closer to theory: Strikers on aggregate played more consistently with theory in the context arms.At the individual level though, partial context using text made strikers more likely to diverge from playing theory, while context through videos made them revert to baseline levels. Using a novel measure from sports psychology, we show that the results of context are driven by subjects' sporting imaging ability (a measure of expertise), with high ability players consistency to theory improving drastically with video-based context. The results demonstrate the interaction of expertise and context in generating play consistent with theoretical predictions.

Incentive Mechanisms for AI: Theory and Evidence

Alex Brown - Texas A&M University

We consider various mechanisms for validating and verifying data through human labor, a problem frequently encountered in managerial applications such as AI data labeling. Under the traditional audit mechanism, an auditor randomly inspects the output of a worker. Alternatively, under the agreement mechanism, a principal hires multiple workers and only investigates conflicts in their work. Combining these features, we introduce the random agreement mechanism, and show that it is less costly than either mechanism. Using experiments to test these predictions, we discover that subjects substantially deviate from the equilibrium prediction of full output, and interestingly, produce 20 percentage points more output under the agreement mechanism. However, because subjects do not complete all tasks, the agreement mechanism is substantially more costly than theory would predict. As a result, the random agreement mechanism is still the most efficient in terms of cost per task completed.

Communication and Collusion in Oligopoly Experiments: A Meta-Study using Machine Learning

Lisa Bruttel - University of Potsdam

While an influential body of economic literature shows that allowing for communication between firms increases collusion in oligopolies, so far we have only anecdotal evidence on the precise communication content that helps firms to coordinate their behavior. In this paper, we conduct a primary-data meta-study on oligopoly experiments and use a machine learning approach to identify systematic patterns in the communication content across studies. Starting with the communication topics mentioned most often in the literature (agreements, joint benefit, threat of punishment, promise/trust), we use a semi-supervised approach to detect all relevant topics. In a second step, we study the effect of these topics on the rate of collusion among the firms. We find that agreements on specific behavior are decisive for the strong positive effect of communication on collusion, while other communication topics have no effect.

TBC

Can Celebi - University of Vienna

TBC

Creativity and AI

Daniela Grieco - University of Milano

This study tests whether AI outperforms humans in creative tasks with a different degree of "openness". As defined in Charness and Grieco (2019), "a true closed problem is one that is presented to the participant, when the method for solving the problem is known [...]", like in what Collins and Amabile (1999) call "algorithmic" tasks. On the contrary, open problems occur when the participant is required to find, invent, or discover the problems" (Unsworth, 2001). We use the same creativity methodology as in our previous articles (Charness and Grieco, 2019, 2022, 2023). In our case, the new subjects score the creativity of items previously produced by an unknown AI or human. Tasks can exhibit different levels of openness, with an ideal continuum from fully closed, "algorithmic" tasks - where accuracy in following the instructions allows to perfectly match the unique solution of the task – and fully open tasks, where solutions are multiple, various, and perhaps even difficult to compare. There are also no instructions to guide towards them or to assure a successful creative output. A higher task openness thus implies a lower probability of successfully solving the problem by simply following the instructions, and thus brings about higher expected variance and novelty in possible answers. This may well affect evaluators' judgement, since they are likely to value novelty and to reward answers that differentiate more from others.

Our results show that humans' average performance in the Open task is significantly higher from AI's one. Strikingly, this is entirely reversed in the Closed task, no matter the version of GPT used. The difference-in-difference across treatments is quite large across these treatments. To shed light on the drivers of creative output for humans and machines, we sketch a simple model that captures human versus AI preferences depending on the degree of task openness and we structurally estimate the two key preferences parameters: the extent to which humans and machines are affected by the openness of the creative task, and the relative weight of human intuition with respect to what is prescribed by instructions.

Our structural estimates show that human intuition contributes between 24% and 45% of the creative score that humans reach in Open task, with human ability of producing novel, unusual ideas representing a key ingredient of their success with respect to AI machines, and AI improvement over time consisting of creative answers that have become better elaborated and less nonsensical, but that are not more novel and original. Human intuition has instead no role in sufficiently closed tasks, like the one we present in this experiment.

As a final test, we ask AI to evaluate the same answers, finding that there is no significant difference between human and AI scores on either task; however, with more precise prompts that restrict the scope of the evaluation to very specific aspects of creativity, AI evaluations reflect human ones.

Using Big Data and Machine Learning to Uncover How Players Choose Mixed Strategies

Toshihiko Hirasawa - UCLA

We examine how humans choose mixed strategies. Previous research (e.g., O'Neill (1987) and Camerer (2011)) has revealed that mixed strategy equilibria describe human behavior in field and lab data reasonably well, while it has also been shown that humans do not exactly follow a mixed strategy equilibrium (e.g., Brown and Rosenthal, 1990). If we are called upon to play a game with a unique mixed strategy equilibrium, we use our intuition, hunches, and some kind of limited reasoning rather than calculating and following the mixed strategy equilibrium. However, it is not yet fully understood exactly what kind of cognitive processes are employed. We aim to answer this question by using unique big data and machine learning models.

Our dataset examines O'Neill's Game (O'Neill, 1987), one of the simplest possible two-person zero-sum games with a unique non-trivial mixed strategy equilibrium. The dataset was collected in a Coursera online course on game theory and contains data on more than 5,000 participants. Each player played the game 30 times with the same opponent, providing roughly 75,000 observations (for each player's role). This is one of the largest datasets for a single treatment in economic laboratory experiments.

To uncover how humans choose mixed strategies, we try to improve conventional behavioral models with the help of the machine learning approach. Models are evaluated by their out-of-sample prediction errors in a dataset that is not used for parameter estimation. Using our unique large dataset, we compare the out-of-sample prediction errors of conventional behavioral models, such as EWA (Camerer, 1999) and the serial correlation model, and leading machine learning models, including the decision tree, LightGBM, LASSO models, and deep learning models (DNN and LSTM).

We find that LSTM substantially outperforms the conventional behavioral models. However, as LSTM is a type of deep learning model and such models have a large number of parameters (more than 5,000), the interpretation of which is not immediately clear. Hence, rather than directly opening the black box of LSTM, we try to gain insights from more interpretable machine learning models, namely the decision tree and LASSO models. These models suggest that the subjects in our data are subject to a particular form of serial correlation. By incorporating the serial correlation and the fact that the opponent is also subject to the same tendency into the EWA model, we obtain a new, improved model that captures how humans learn to choose mixed strategies, which we call mixing EWA. It captures almost all the predictive power of our best machine learning model (LSTM).

Moreover, we artificially reduce the size of our dataset used for parameter estimation to examine how the out-of-sample prediction errors of the models would change. Under the typical lab dataset size with at most a few hundred subjects, the superiority of the machine learning model is not noticeable. As the dataset size increases, however, the performance of machine learning keeps on improving, while the conventional model (EWA) does not show any noticeable improvement. These results suggest that to clearly detect the limitations of conventional models using machine learning models, we need a dataset that is an order of magnitude larger than the typical lab dataset.

Finally, we examine how much out-of-sample predictive power the improved models have compared to the best machine learning model. We indeed find that there is no statistically significant difference in the predictive power of our best modified behavioral model and the best machine learning model. Moreover, we conduct two statistical tests to see if our modified models actually capture what is encoded in the black box of EWA. Finally, we point out a potential concern about our procedure, and to address this issue, we double-check the external validity of our modified model.

The Emergence of Strategic Reasoning of Large Language Models

Gavin Kader - Southwestern University of Finance and Economics

Although large language models (LLMs) have demonstrated strong reasoning abilities in structured tasks (e.g., coding and mathematics), it remains unexplored whether these abilities extend to strategic multi-agent environments. We investigate strategic reasoning capabilities – the process of choosing an optimal course of action by predicting and adapting to others' actions – of LLMs by analyzing their performance in three classical games from behavioral economics. We evaluate three standard LLMs (ChatGPT-4, Claude-2.1, Gemini 1.5) and three specialized reasoning LLMs (GPT-o1, Claude-3.5-Sonnet, Gemini Flash Thinking 2.0) using hierarchical models of bounded rationality. Our results show that reasoning LLMs exhibit superior strategic reasoning compared to standard LLMs (which do not demonstrate substantial capabilities), and often match or exceed human performance. Since strategic reasoning is fundamental to future AI systems (including Agentic AI and Artificial General Intelligence), our findings demonstrate the importance of dedicated reasoning capabilities in achieving effective strategic reasoning.

When AI Judges Your Work: The Hidden Costs of Algorithmic Oversight

Daniel Martin - UCSB

We use an online experiment to study whether individuals change their behavior when they know AI is used to judge their work instead of humans. We find that individuals produce more output under algorithmic oversight, but that controlling for the amount of output, the quality of the output is lower, regardless of whether quality is measured using humans or AI assessments. We also find that individuals are more likely to use external tools when they know AI is used to judge their work instead of humans. However, this factor does not explain the differences in quality across treatments. Our results highlight essential considerations about the design and implications of automated evaluation systems.

Searching for the External Validity of Social Preference Games: A Guide of Field Environments

Sergio Pirla - University of Zaragoza

The last couple of decades have witnessed a lively debate on the external validity of social preference games. Yet, scientific progress in this area has been restrained by the difficulty of delineating the field environments that social preference games should generalize to. This paper addresses this gap by presenting a guide of field settings that are expected to relate to social preference games by experimental and behavioral economists.

The paper includes 3 studies. In Study 1, we conduct a systematic review and meta-analysis of papers using social preference games published in the top five economics journals (AER, QJE, JPE, REStud, Econometrica). For each of these, we extract all explicit references to real-life field environments that authors associate with the games, and we code these references into 22 distinct categories of field behavior. We then quantify the prevalence of each category and analyze variation across games. Our results reveal that 65% of papers include at least one reference to a real-life setting, with the most common categories being social and household interactions, political and social issues, compensation and sanctioning scheme design, charity, and labor relations.

Study 2 extends and validates the findings of Study 1 using a large language model (LLM)-based approach. First, we replicate the meta-analysis from Study 1 on the same set of top-five journal papers using structured prompts administered through ChatGPT-4o. This validation exercise yields a 90% classification accuracy relative to our manual coding of exclusion criteria. Focusing on the final set of included papers, the LLM-based analysis identifies a higher number of external validity claims per paper, yet preserves the relative ranking of the categories ($r = .814$, $p < .001$). Following this validation, we apply the LLM-based method to a much broader corpus of papers published in six additional economics journals: the Economic Journal, Review of Economics and Statistics, Journal of the European Economic Association, Experimental Economics, Journal of Economic Behavior & Organization, and Games and Economic Behavior. This expanded analysis reveals external validity claims in 88% of the included papers. Importantly, across categories, the results of this extended meta-analysis correlated with those of Study 1 ($r = .734$, $p < .001$).

In Study 3, we turn to the expectations of expert researchers in behavioral and experimental economics. We conduct a survey among members of the Economic Science Association (ESA), asking respondents ($N = 89$) to (i) rate the degree to which behavior in social preference games relates to each of the 22 field categories identified in Study 1, and (ii) describe 1–3 specific field situations they believe are most relevant for each game. The ratings reveal systematic variation in perceived relatedness across categories, with charity, social and household interactions, and group dynamics receiving the highest average scores. The open-ended responses yield 664 valid field settings, which we classify using crowdsourced coding from a separate sample of 300 Prolific participants. All responses are assigned to one or more of the 22 field categories using three complementary indexes of prevalence. Across both ratings and open responses, we find high internal consistency and significant correlations with the literature-based categories from Study 1 and the LLM-based classifications from Study 2 (all $r > .5$, $p < .01$).

Combining the results of the three studies using principal component analysis, we generate an integrated and ranked guide of field environments for social preference games. By documenting and organizing these expectations, we provide a valuable resource for the design, interpretation, and generalization of experimental research in economics.

MAD Chairs: A new tool to evaluate AI

Chris Santos-Lang - Independent Researcher

This paper presents a new contribution to the problem of AI evaluation. Much as one might evaluate a machine in terms of its performance at chess, this approach involves evaluating a machine in terms of its performance at a game called "MAD Chairs." At the time of writing, evaluation with this game exposed opportunities to improve Claude, Gemini, ChatGPT, Qwen and DeepSeek. This paper makes a case that crossing the frontier from human subjects to LLMs (and other AI) as subjects of experimental game theory is necessary to ensure that AI are safe to deploy. Crossing that frontier may also help keep behavioral game theory relevant as LLMs and other AI increasingly influence real-world social behavior. MAD Chairs is a significant contribution to such research because much social strife seems to have derived from MAD Chairs behavior that such influence could improve (e.g., distribution of jobs, opportunities at influence, and even spots in traffic). MAD Chairs is easy to discover--simply add players to any Coordination game until they outnumber the resources they are to divide--so the real force of this paper comes from introducing an unexpected complex turn-taking strategy which dominates accepted strategies. The falsification of accepted theory in this case may prompt concern that previously unaccounted strategies could similarly obsolesce our understanding of other games, especially if LLMs empower us to devise strategies we would not have conceived without them. Thus, this paper points to new methods that may become required as the field matures into a new age; in particular, it offers new justification to confirm optimality via empirical testing.

Evaluating Instrumentally Rational Deception by LLMs in 2x2 Signaling Games

Samuel Taylor - University of California, San Diego; Cognitive Science Department

Large Language Models (LLMs) are effective at deceiving, when prompted to do so. But under what conditions do they deceive spontaneously? Models that demonstrate better performance on reasoning tasks are also better at prompted deception. Do they also increasingly deceive spontaneously in situations where it could be considered instrumentally rational to do so? If so, then training LLMs to be better reasoners confers the attendant risk of unintentionally inducing patterns of instrumentally rational, spontaneous deception. We report on results from a pre-registered empirical study that measures rates of spontaneous deception produced by eight LLMs (some closed-source and some open-source) using 2x2 behavioral games (Prisoner's Dilemma, Matching Pennies, and Stag Hunt), modified so that the LLM can communicate to the other player before or after each selects their move, using unconstrained language. This setup creates an opportunity for the LLM to deceive in conditions that vary in how useful deception might be to its rational self-interest. For example, players in Prisoner's Dilemma have the opportunity to persuade the other player that they will Cooperate before choosing to Defect, which, if successful, confers a larger reward. In another condition, a player may only have the opportunity to communicate *after* the other player chooses an action: in this case, deception is less rational. We enlist both human and LLM annotators to label incidences of deception, operationalized as an explicit disagreement in expressed intention in the message and selected action.

The results indicate that 1) all tested LLMs spontaneously misrepresent their actions in at least some conditions, with deception rates for certain models up to 49% for Matching Pennies and 96% for Prisoner's Dilemma, 2) LLMs are generally more likely to do so in situations in which deception would benefit them, as measured by chi-squared tests between minimally-different, contrasting conditions where deception is rationally expected in one and not rationally expected in the other (such as by manipulating point values to encourage deception, or manipulating turn order as previously described), and 3) LLMs exhibiting better reasoning capacity overall tend to spontaneously deceive at higher rates, as measured by Pearson correlations between deception increases in rational conditions and LLM reasoning capability as measured by a separate benchmark (MATH, Hendrycks et al., 2021). Taken together, these results suggest a tradeoff between LLM reasoning capability and honesty. They also provide evidence of reasoning-like behavior in LLMs from a novel experimental configuration. Finally, they reveal certain contextual factors that affect whether LLMs will deceive or not.

We discuss consequences for autonomous, human-facing systems driven by LLMs both now and as their reasoning capabilities continue to improve. We also discuss preliminary results from behavioral experiments currently underway with humans participating in the same experiment, in order to establish a baseline of human behavior in the modified 2x2 games to compare against LLM behavior. Modified behavioral games offer a unique framework to better assess potential risks that may arise from optimizing for LLM reasoning, such as deception, and we conclude by discussing how the approach described here may be generalized in future work to more ecologically plausible contexts.

The Impact of Naturalistic and Familiar Contexts on Strategic Sophistication

Whenzhuo Xu - Carnegie Mellon University

Strategic sophistication—the ability to iteratively reason about others' mental states and actions—is fundamental to many social interactions. However, a puzzling contrast exists: while psychology and behavioral research suggests that people can naturally engage in multi-step reasoning about others' minds in real life, experiments using abstract economic games typically find limited strategic depth, with participants often exhibiting zero or only one level of iterative reasoning. To reconcile this gap, we investigate whether and how naturalistic and familiar contexts enhance strategic sophistication.

Across two preregistered experiments on Prolific (N = 409 and N = 415), we systematically compared strategic sophistication in three types of contexts: abstract contexts (minimal framing), unfamiliar naturalistic contexts (embedded in real-life settings but less familiar to daily experience), and familiar naturalistic contexts (embedded and relatable to everyday life). Using a mixed design, participants were randomly assigned to one of the three contexts (between-subjects) and completed decisions across four recursive knowledge states (within-subjects). These knowledge states varied in what participants knew about the game, what their counterpart knew, and what each person knew about the other's knowledge. In Study 1, we adapted the classic stag-hunt coordination game into unfamiliar and familiar naturalistic settings. In Study 2, we abstracted a volunteer's dilemma game that is typically presented in familiar contexts. In addition, to directly measure strategic sophistication across these contexts, we introduced a novel behavioral measure that categorizes participants based on how they adjust their decisions across knowledge states and assesses their strategic sophistication based on behavioral patterns.

Our findings revealed several key insights. First, participants in both studies consistently exhibited higher strategic sophistication in naturalistic contexts compared to abstract ones. Familiar naturalistic contexts elicited the higher levels of sophistication compared to the unfamiliar and abstract contexts in both studies, suggesting not all naturalistic contexts were equally effective and familiarity with the scenario appeared critical for prompting deeper strategic reasoning. Second, perfect accuracy in understanding recursive knowledge about others was associated with higher likelihood to be classified using our measure (Study 2) and higher levels of strategic sophistication (Study 1), suggesting that accurately tracking others' knowledge states underpins more sophisticated reasoning. Third, individuals who demonstrated greater strategic sophistication achieved higher expected monetary payoffs in both studies, suggesting the ecological validity of our measure. In addition, we explored participants' open-ended rationales for their decisions and used large language models (LLMs) to classify their strategic sophistication.

Together, these findings demonstrate that strategic sophistication is a context-dependent ability, shaped by the naturalism and familiarity of the decision environment. Beyond documenting contextual effects, our work also contributes to the literature by introducing a novel, model-free behavioral measure that directly captures individuals' depth of recursive reasoning. More broadly, our work advances understanding of how contextual framing shapes decision-making and cognitive abilities.