Event Description:

Research Topic, Current State of Research, and Questions:

Social reality is built on the capacity of human beings to engage in social behavior – complex forms of intentional, coordinated actions of multiple individuals. In certain cases, peoples' willingness to engage in such behavior can be readily explained by their interest in advancing their personal goals beyond what is achievable through individual actions. In other cases, their engagement in such behavior seems to go against their personal inter-ests, necessitating explanations that delve into motivations beyond those that can be re-duced to narrowly defined interests of individuals.

Social behavior is a topic that can be approached from many sides: social and political sciences, economics, social psychology, cognitive psychology, and philosophy. An interdis-ciplinary event bringing all these disciplines together would collect highly divergent meth-odologies. Such divergence would hamper tackling our research questions in depth. There-fore, we intend to maintain a relative methodological uniformity within the conference by using game theory as a common background. Game theory is an abstract theory that char-acterizes rational behavior in interdependent decision problems – situations where multi-ple agents interact with each other and the consequences of their interaction are deter-mined by their combined actions (Luce and Raiffa 1957). For more than 50 years, game theory has served as the primary conceptual framework for developing theories aiming to explain social behavior. It has been used to study social behavior in the aforementioned fields of research. Game-theoretic accounts converge on the idea that social behavior is sustained by shared expectations among interacting individuals regarding each other's ac-tions and their sufficiently aligned valuations of the consequences of combined actions, as paradigmatically set out in the groundbreaking work of Thomas Schelling (1960) and David Lewis (1969).

However, these accounts also disagree in their explanations of how these shared expectations and valuations emerge:

Social norm theories posit that social behavior stems from individuals' adherence to shared behavioral rules. There are many social norm theories and the differences between them are substantial (see, e.g., Schelling 1960, Lewis 1969, Sugden 1984, Axelrod 1986, Vanderschraaf 1995, Skyrms 1996, Binmore 2005, Bicchieri 2006, Alexander 2007, Gintis 2008). Virtually all of the theories define a social norm as a behavioral rule that prescribes each interacting agent a particular action or a sequence of actions in a certain type of social interaction. A behavioral rule becomes a social norm when there exists a substantial share of agents in the population who know that the rule exists for a certain type of social situation and are motivated to follow it in situations where they expect the other agents to do the same (Bicchieri 2006). In many social interactions, agents' incentives are mixed: they see some potential benefits from cooperating with others, but are also tempted to exploit others by not investing effort into joint actions. For this reason, deviations from social norms are often associated with expectations of sanctions that motivate the agents to comply (Axelrod 1986, Binmore 2005, Bicchieri 2006).

Despite these similarities, social norm theories differ in their explanations of how social norms emerge and what role the standard economic, or means-ends, rationality plays in agents' compliance with the norms (for an overview, Bicchieri, Muldoon and Santuoso 2023). In addition, they can be criticized for being too theoretically permissive and, as a result, suffering from empirical testability issues: with sufficiently severe punishments for non-compliance, game theoretic models of social norms can represent virtually any kind of behavior as agents' compliance with a norm.

Prosocial preference theorists argue that humans have a stable disposition to avoid behavior that is considered antisocial, and sanction those that engage in it. The simplest models assume that social agents prefer to avoid the outcomes of interactions that are considered unfair – excessively advantageous or disadvantageous to some of the interact-ing individuals (Fehr and Schmidt 1999, Bolton and Ockenfels 2000). More sophisticated prosocial preference models, known as reciprocity models, represent social agents as in-dividuals whose behavior is conditional on their beliefs about the other agents' actions: prosocial in situations where the observed or expected actions are deemed socially ac-ceptable, and antisocial or even punitive in situations where they are not (Rabin 1993, Falk and Fischbacher 2006).

Prosocial preference theories face the same empirical testability problems as the theories of social norms: with a sufficiently permissive treatment of the basic parameters of models, prosocial preference theory can account for virtually any observed social behav-ior. In addition, reciprocity models predict the same conditional preference for prosocial behavior as social norm theories, thus making both theories empirically indistinguishable.

The **team reasoning theory**, originally introduced by Robert Sugden (1993), suggests that certain structural properties of real-world social interactions my trigger a shift of in-teracting agents' mode of reasoning from individualistic reasoning to reasoning as mem-bers of a team. When agents reason individualistically, they identify the appropriate ac-tions in a social interaction by answering the question "what it is that *I* should do to pro-mote *my* interests?". By answering this question, each agent identifies actions that max-imize the satisfaction of their personal interests in light of constraints imposed by the ex-pected actions of other interacting agents. When agents reason as members of a team, they identify the appropriate actions by answering the question "what it is that *we* should do to promote *our* interests”. By answering this question, each agent identifies a combination of players actions that attains the team objective, and plays their part in implement-ing it if they expect the others to do the same (Radzvilas and Karpus 2021). Although the theory can claim substantial empirical support from game-theoretic experiments, a number of serious conceptual issues remain unresolved. Different versions of the team reasoning theory offer different answers to the question of what triggers a shift in agents’ mode of reasoning: some accounts suggest that it is triggered by psychological framing effects (Bacharach 1999, 2006), while others emphasize agents’ rational recognition that joint coordinated actions are mutually advantageous (Sugden 2011, 2015) or result in a unique and satisfactory resolution of agents’ conflict of interests (Karpus and Radzvilas 2018). There also exist different conceptions of team objectives. Some versions of the theory define the team objective as the attainment of Pareto optimal outcomes (Bacharach 1999, 2006). In other versions of the theory, the team objective is defined as the advancement or maximization of mutual advantage (Sugden 2011, 2015, Karpus and Radzvilas 2018).

Finally, the **virtual bargaining theory**, originally introduced by Jennifer M. Misyak and Nick Chater (2014), suggests that social agents resolve social coordination and cooperation problems by engaging in a hypothetical process of reasoning that allows them to identify the joint actions that they would agree to implement if they were bargaining – explicitly negotiating a joint action agreement that they would be motivated to implement rather than seek to satisfy their goals via individual actions. The theory uses the Nash bargaining solution (Nash 1950) – a formal characterization of a bargaining agreement satisfying a number of intuitively desirable properties – to predict the outcomes of interactions involv-ing such agents.

The theory is relatively new and faces a number of challenges. A key outstanding ques-tion for the theory of virtual bargaining is that of how this mode of reasoning can be gen-eralized to apply across a wide set of games. Since the existing approaches to solving ex-plicit bargaining decision problems (Nash 1950; Kalai and Smorodinsky 1975; Kalai 1977) rely on the existence of a non-agreement baseline – a unique outcome that obtains if play-ers fail to reach agreement – the theory needs to provide a general account (if such ac-count is possible) of what this baseline is when players bargain tacitly to solve decision problems that, in and of themselves, are not explicit bargaining decision problems to begin with (Karpus and Radzvilas 2021). In addition, it is unclear whether virtual bargaining the-ory should be viewed as an independent theoretical framework for analysing social inter-actions, or as a theory that explains how team reasoning agents identify mutually advan-tageous outcomes in games (see Misyak and Chater 2014).

All of the aforementioned accounts can claim substantial amounts of experimental results as supporting evidence. However, as mentioned, in many cases experimental evidence equally supports multiple or even all of the competing theories, thus creating a severe problem of underdetermination. So far, there has been no conclusive theoretical argument or an empirical test to select among the competing accounts.

This conceptually unsatisfactory state of affairs raises a number of important questions. Is there a methodology to select among the competing theories? Should these accounts be viewed as competing theories, or rather as theories that complement one another? Are there better unconsidered alternatives to existing theories? Is the game-theoretic frame-work – one that leads to a proliferation of empirically indistinguishable theories – truly the best approach towards explaining social behavior?

To facilitate the discussion of these issues and about the status and relationship of the various competing theories, we intend to invite leading scholars who have made substan-tial contributions to different versions of the aforementioned theories, as detailed in the list of prospective participants.

Our personal interest in the conference is still another one, namely to bring in the perspective on the proposed topic that is elaborated within our Reinhart-Koselleck Project on **Reflexive Decision and Game Theory** (for details see https://www.philosophie.uni-konstanz.de/forschung/drittmittelprojekte/reinhart-koselleck-projekt/). There we put for-ward and elaborate a novel equilibrium concept called dependency equilibria. This notion results from applying a reflexive approach to decision and game theory – where "reflexive" means that interacting agents are represented as taking into account their present and future decision-relevant mental set-ups and their causal consequences. This notion of a dependency equilibrium has also the potential of throwing new light on phenomena of social cohesion. So, we think that we can provide an unconsidered alternative to existing theories on social behavior and want to learn how it fares in relation to those theories.

Reference List

1. Alexander, J. (2007). *The Structural Evolution of Morality*, Cambridge: Cambridge University Press.
2. Axelrod, R. (1986). An Evolutionary Approach to Norms. *American Political Science Review*, 80(4): 1095–1111.
3. Bacharach, M. (1999). Interactive Team Reasoning: A Contribution to the Theory of Co-operation. *Research in Economics* 53: 117–147.
4. Bacharach, M. (2006). *Beyond Individual Choice: Teams and Frames in Game Theory*. Princeton: Princeton University Press.
5. Bicchieri, C. (2006). *The Grammar of Society: The Nature and Dynamics of Social Norms*, Cambridge: Cambridge University Press
6. Binmore, K. (2005). *Natural Justice.* Oxford University Press.
7. Bolton, Gary E. and Axel Ockenfels, 2000, "ERC: A Theory of Equity, Reciprocity, and Competition", *American Economic Review*, 90(1): 166–193.
8. D'Arms, J. (1996). Sex, Fairness, and the Theory of Games. *Journal of Philosophy*, 93 (12): 615–627.
9. Fehr, E. and Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics*, 114(3): 817–868.

10. Falk, A. and Fischbacher, U. (2006). A Theory of Reciprocity. *Games and Economic Behavior*, 54(2): 293–315.
11. Gintis, H. (2008). *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton and Oxford: Princeton University Press.
12. Kalai, E. (1977). Proportional solutions to bargaining situations: interpersonal utility comparisons. *Econometrica* 45: 1623–1630.
13. Kalai, E. and Smorodinsky, M. (1975) Other solutions to Nash's bargaining problem. *Econometrica* 43: 513–518.
14. Karpus, J. and Radzvilas, M. (2018). Team reasoning and a measure of mutual advantage in games. *Economics and Philosophy* 34: 1–30.
15. Karpus, J. and Radzvilas, M. (2021). `Game Theory and Rational Reasoning'. In Heilmann, C. and Reiss, J. (eds.). *The Routledge Handbook of the Philosophy of Economics*. New York and London: Routledge, pp. 99-112.
16. Lewis, D. (1969). *Convention: A Philosophical Study*, Cambridge, MA: Harvard University Press. doi:10.1002/9780470693711
17. Luce, R. D. and Raiffa, H. (1957). *Games and Decisions: Introduction and Critical Survey*, New York: Dover Publications, Inc.
18. Misyak, J., Chater, N. (2014). Virtual Bargaining: A Theory of Social Decision-Making. *Philosophical Transactions of the Royal Society*, *B* 369: 1–9.
19. Nash, J. (1950). The bargaining problem. *Econometrica* 18: 155–162.
20. Nash, J. (1951). Non-cooperative games. *Annals of Mathematics* 54: 286–295.
21. Rabin, M. (1993). Incorporating Fairness into Game Theory and Economics. *American Economic Review*, 83(5): 1281–1302.
22. Schelling, T. C. (1960). *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.
23. Skyrms, B. (1996). *Evolution of the Social Contract*. Cambridge: Cambridge University Press.
24. Sugden, R. (1986) [2004]. *The Economics of Rights, Co-operation and Welfare*, second edition. Basingstoke: Palgrave Macmillan, 2004.
25. Sugden, R. (1993). Thinking as a team: towards an explanation of nonselfish behavior. *Social Philosophy and Policy* 10: 69–89.
26. Sugden, R. (2011). Mutual advantage, conventions and team reasoning. *International Review of Economics* 58: 9–20.
27. Sugden, R. (2015). Team reasoning and intentional cooperation for mutual benefit. *Journal of Social Ontology* 1: 143–166.
28. Vanderschraaf, P. (1995). Convention as Correlated Equilibrium. *Erkenntnis*, 42(1): 65–87.