Who is responsible for collective action?

Organizers: Casey Lewry (lewry@princeton.edu) and Tania Lombrozo (lombrozo@princeton.edu)
Department of Psychology, Princeton University

Shannon Wing[1] (spwing@mit.edu), Sydney Levine[2] (sydneyl@allenai.org), Joshua Tenenbaum[3] (Josh.Tenenbaum@gmail.com), and Lionel Wong[3] (zyzzyva@mit.edu)
[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology
[2]Allen Institute for AI
[3]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

Sofia Bonicalzi (sofia.bonicalzi@uniroma3.it)
Department of Philosophy, Communication and Performing Art, Roma Tre University

Tobias Gerstenberg (gerstenberg@stanford.edu)
Department of Psychology, Stanford University

Keywords: collective action, responsibility, causal reasoning

Goals and Scope

Reducing inequality, mitigating climate change, and responding to public health crises are large-scale goals that require the cooperation and coordination of many individuals. These goals cannot be achieved by one individual alone, and contributing is not always beneficial to each individual. And yet, individuals must contribute in order to make a difference. How do we hold individuals and groups responsible for collective action?

Responsibility for collective action has long been studied across cognitive science disciplines. In this symposium, we will focus on emerging perspectives from psychology, computer science, and philosophy.

From psychology, we have learned when and why individuals are motivated to participate in collective action; for example, due to an individual's social identification with the group (van Zomeren et al., 2018). Psychology research has also examined judgments about other people's responsibility for collective action (Gerstenberg & Lagnado, 2010): Was John the reason that the new bill was signed into law? How much of a role did he play?

From computer science, game theoretic models have been used to study the collective action problem (Hardin, 1968), analyzing how individual sacrifices can successfully (or unsuccessfully) lead to cooperation (Chalkiadakis et al., 2022). Additionally, network effect models have demonstrated how online and offline social networks impact individuals' collective action participation (Centola, 2010).

Finally, philosophy research has highlighted the crucial distinction between backward- and forward-looking responsibility judgments (i.e., who caused this problem versus who is responsible for ending it?; Van de Poel, 2011). Other work on social contract theory (Muldoon, 2016) and ethical theories have helped posit potential solutions to the collective action problem.

In this symposium, we bring together interdisciplinary researchers across these disciplines. The symposium will consist of four talks, followed by a roundtable discussion. **Shannon Wing, Sydney Levine, Joshua Tenenbaum, and Lionel Wong** will explore the limitations of large language models in capturing human moral reasoning in collective action problems and introduce a neuro-symbolic system to address these shortcomings. Next, **Casey Lewry and Tania Lombrozo** will propose a new framework for research on the psychology of inequality, arguing that research has underemphasized judgments of responsibility for preventing future inequalities. Then, **Sofia Bonicalzi** will discuss what constitutes responsibility for individuals and groups, and how this changes in collective action settings. Finally, **Tobias Gerstenberg** will introduce a computational account of individual and group responsibility, suggesting that people use counterfactual simulations of alternate possibilities to determine how responsible an individual is for a group outcome.

A neuro-symbolic approach to moral judgment in collective action problems

Shannon Wing, Sydney Levine, Joshua Tenenbaum, and Lionel Wong

Recent research has suggested that one of the ways that people make moral judgments in collective action problems is by using the logic of universalization – a version of the question "What if everyone felt free to do that?" (Levine et al, 2020). In this talk we will describe a series of ways that state-of-the-art large language models (LLMs) fail to capture human judgments in novel collective action problems. This short-coming of language models underscores their (current) limited ability to reason, make logical inferences, plan over models of the world, and consider probabilities – all of which are critical components of the way humans make moral judgments in novel collective action problems. We then describe a neuro-symbolic system that does much better. This system uses an LLM as a model of the human language-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

processing ability. The LLM interfaces with natural language, translating vague language into prior distributions, extracting the values of morally relevant features, and generating the code-based elements of a moral world model in a probabilistic programming language. A series of computationally formalized moral mechanisms can then be run in the probabilistic program, yielding a prediction of human moral judgment of the original natural language case. Overall, our neurosymbolic system computationally characterizes the cognitive mechanisms behind human moral judgments of collective action problems.

Varieties of responsibility attributions for inequality

Casey Lewry and Tania Lombrozo

Decades of psychology research have led to a better understanding of the factors that affect how people reason about the causes of inequalities, such as the racial wealth gap. But our understanding of the psychology of inequality remains limited because this research has largely focused on causal and retrospective judgments (i.e., judgments about what caused past or present inequalities). In this project, we argue that two distinctions are valuable for clarifying attributions for inequality: the moral-causal distinction and the retrospective-prospective distinction. The moral-causal distinction differentiates judgments of agents' blameworthiness and obligation (moral) from judgments of their role in bringing about an outcome (causal). The retrospective-prospective distinction differentiates judgments about the agents, actions, and conditions that led to historical or present inequality (retrospective) from judgments about what agents can or should do to remedy existing inequality and prevent it in the future (prospective). We argue that this framework helps researchers identify unwarranted inferences, gaps in the literature, and directions for future work, including a focus on work that bridges multiple types of judgments (e.g., retrospective causal and retrospective moral). This framework offers a new perspective that may allow us to better explain, predict, and shape judgments relating to inequality.

Backward and forward-looking responsibility in individuals and collectives

Sofia Bonicalzi

Philosophical and psychological accounts have suggested that third-party attributions of backward-looking responsibility – notably responsibility for past wrongs as associated with the notion of deserved blame – are to be assessed separately from remedial and forward-looking responsibility. These different forms of responsibility rely on whether the targeted individuals display specific agentive and epistemic features, such as intentionality, autonomy, identification with motives, or ability to engage in counterfactual thinking. I will argue that when shifting from individuals to groups, the boundaries between different forms of responsibility are likely to become more blurred. In particular, when groups are considered as the aggregate of discrete individual members, the opportunities for establishing whether they meet the various responsibility requirements – so that responsibilities can be shared – strongly depend on contextual variables, including group size and intergroup hierarchies. Furthermore, when groups are considered as unified collective entities, major problems arise in terms of whether such collectives can even appropriately fulfill the different responsibility conditions to the extent that it remains unclear whether they are endowed with intentions and other relevant mental representations.

Holding others responsible: The role of counterfactual contrasts

Tobias Gerstenberg

How do people hold individuals responsible for collective outcomes? In this talk, I will share a computational framework that conceptualizes responsibility judgments in terms of counterfactual contrasts defined over people's causal model of the situation. According to this framework, people attribute responsibility by comparing what actually happened with what would have happened in relevant counterfactual situations. For collective outcomes, people first assess how critical each person's action would be for a positive result. And, after the outcome happened, people consider how close each person's action was to having been pivotal for the outcome. As predicted by the model, individuals are held more responsible for group outcomes when their actions are critical and pivotal. I will also show that people's expectations affect their responsibility judgments, and that individuals are held less responsible whose contributions could have easily been replaced by someone else.

References

Centola, D. (2010). The spread of behavior in an online social network experiment. *Science, 329*(5996), 1194-1197.

Chalkiadakis, G., Elkind, E., & Wooldridge, M. (2022). Computational aspects of cooperative game theory. Springer Nature.

Gerstenberg, T., & Lagnado, D. A. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition*, *115*(1), 166-171.

Hardin, G. (1968). The tragedy of the commons: the population problem has no technical solution; it requires a fundamental extension in morality. *Science, 162*(3859), 1243-1248.

Muldoon, R. (2016). Social contract theory for a diverse world: Beyond tolerance. Taylor & Francis.

Van de Poel, I. (2011). The relation between forward-looking and backward-looking responsibility. In *Moral Responsibility: Beyond Free Will and Determinism* (pp. 37-52). Dordrecht: Springer Netherlands.

van Zomeren, M., Kutlaca, M., & Turner-Zwinkels, F. (2018). Integrating who "we" are with what "we" (will not) stand for: A further extension of the Social Identity Model of Collective Action. *European Review of Social Psychology, 29*(1), 122-160.