Measures of individual and groupwise ex-post and ex-ante responsibility in extensive-form games with unquantifiable uncertainty

Jobst Heitzig and Sarah Hiller

FutureLab on Game Theory and Networks of Interacting Agents
Potsdam Institute for Climate Impact Research
PO Box 60 12 03, 14412 Potsdam, Germany
and
Institute for Mathematics, Freie Universität Berlin
Arnimallee 6, 14195 Berlin, Germany

Extended abstract for Formal Ethics 2019

Major real-world problems such as climate change typically involve many heterogeneous agents with conflicting interests, complex dynamic interactions, and possibly diverging information states containing various forms of uncertainty.

The dominant formal methodology for studying these problems is game theory and economics, where perceived failure of collective action is often explained by collective inefficiency of strategic equilibria arising from the supposed selfishness and individual rationality of the agents. Correspondingly, these analyses often (though not always) lead to rather pessimistic predictions (Barrett, 1994; Heitzig and Kornek, 2018).

However, empirical evidence shows that real-world agents are also guided by moral principles (Fehr and Gachter, 2000; Wang et al., 2017; Milinski et al., 2008). In the public debate on climate change particularly the concept of responsibility is featured prominently (Billett, 2010; Lorenzoni et al., 2007). While game-theoretic analyses tend to either ignore this or at best treat it as a form of signalling that might influence equilibrium selection, the moral philosophy literature takes the concept much more seriously. Building on foundational work on different forms of responsibility, focusing on thought experiments and paradigmatic situations such as the trolley problem, some authors have started to apply these concepts to more complex problems including climate change (Gardiner, 2004).

The literature and the public debate suggest there are different forms of responsibility, all of which are relevant in the real world. Vincent (2011), for example, distinguishes six different basic forms of the use of the term 'responsibility', such as *being a responsible person* or *being blameworthy*. We focus here on those aspects of responsibility that are most closely related to specific actions and their possible consequences. For these, we start by making a basic distinction between *ex-ante (forward-looking)* 'responsibility to' take certain actions, which is related to the concepts of 'obligation' and 'duty', and *ex-post (backward-looking)* 'responsibility for' certain consequences or outcomes, which is more related to the idea of 'blame'. For the latter we further distinguish between *factual* ex-post responsibility for actually realized consequences of actions taken and *counterfactual* ex-post responsibility for potential

consequences of actions taken that did however not actually materialize. This last concept is related to the discussion about 'moral luck' (Nagel, 1979; Andre, 1983). The literature further suggests differences between individual vs. collective responsibility and the responsibility assigned by an 'ethical observer' vs. that perceived by the agents themselves — a distinction which is relevant for understanding the psychological 'diffusion of responsibility' effect (Darley and Latané, 1968). Existing studies also indicate that there may be quantitative degrees of responsibility (Chockler and Halpern, 2004). Still, most of these variants are not sufficiently formalized to be clearly distinguishable, generally applicable, or even measurable. Naïve ad-hoc measures of responsibility such as the direct identification of 'historical responsibility' for climate change with the amount of past cumulative greenhouse-gas emissions (Botzen et al., 2008) are not readily generalizable and typically do not fulfill basic consistency requirements, e.g., that degrees of responsibility should rather depend on the magnitude of effects than on the magnitude of causes.

In this contribution, we are guided by the hypothesis that a mathematical theory of responsibility with quantitative measures is needed, and will explore an approach that bases these measures on probabilities and ethical evaluations of the uncertain consequences of possible and actual actions of agents. Doing so, the resulting theory shall work well in combination with established concepts and methods from game theory, control theory and Bayesian decision theory. A major challenge in this is the proper treatment of different forms of uncertainties, in particular regarding the behaviour of other agents. Assume it were regarded ethically correct if agents based their actions on beliefs about other agents' behaviour, e.g., encoded via subjective probability distributions as in Bayesian decision theory; then this would result in an underestimation of responsibility in situations where agents can hope others will solve a problem instead of themselves. So a proper theory needs to distinguish probabilistic uncertainty with known probabilities from uncertainty due to free will, and from other forms, such as when only probability intervals are given, as in the reports of the IPCC (Mastrandrea et al., 2011).

To approach this task, we first extend the existing tree-based data structure of extensive-form games by nodes representing unquantified uncertainty in the form of 'possibility sets'. Extending our work done in (Heitzig et al., 2018), we then select some paradigmatic example situations and thought experiments, translate them into the developed data structure, and analyse them regarding the abovementioned aspects of responsibility in order to relate these aspects to certain features of the formal representation. For these example situations, we also formulate candidates for the assessment of the various forms of responsibility the agents in these situations have individually and collectively, which then serve as boundary conditions for prospective general qualitative and quantitative measures of the various forms of responsibility. In addition, we postulate a small set of more general axioms representing desirable properties of such measures. Finally, we present first quantitative formulas that fulfill these boundary conditions and axioms and can thus serve as a proof of concept for our approach and as a starting point for further discussions and comparisons of different flavours of responsibility measures.

As an outlook, we will hint at natural connections of the above framework to possible-worlds semantics for alethic and deontic modal logics, in particular stit logics (Broersen, 2011), which may lead to a corresponding modal logic of responsibility.

References

Andre, Judith, "Nagel, Williams, and moral luck," *Analysis*, 1983, *43*, 202–207.

Barrett, Scott, "Self-enforcing international environmental agreements," *Oxford Economic Papers*, 1994.

Billett, Simon, "Dividing climate change: Global warming in the Indian mass media," *Climatic Change*, 2010, *99* (1), 1–16.

Botzen, W. J.W., J. M. Gowdy, and J. C.J.M. Van den Bergh, "Cumulative CO2emissions: Shifting international responsibilities for climate debt," *Climate Policy*, 2008, *8* (6), 569–576.

Broersen, Jan, "Modeling Attempt and Action Failure in Probabilistic stit Logic," in "Twenty-Second International Joint Conference on Artificial Intelligence" 2011, pp. 792–797.

Chockler, Hana and Joseph Y Halpern, "Responsibility and Blame : A Structural-Model Approach," *Journal of Artificial Intelligence Research*, 2004, *22*, 93–115.

Darley, John M and Bibb Latané, "Bystander Intervention in Emergencies: Diffusion of Responsibility," *Journal of Personality and Social Psychology*, 1968, *8* (4 PART 1), 377–383.

Fehr, By Ernst and Simon Gachter, "Cooperation and Punishment in Public Goods Experiments," *The American Economic Review*, 2000, *90* (4), 980–994.

Gardiner, Stephen M, "Ethics and Global Climate Change," *Ethics*, 2004, *114* (3), 555–600.

Heitzig, Jobst and Ulrike Kornek, "Bottom-up linking of carbon markets under far-sighted cap coordination and reversibility," *Nature Climate Change*, 2018.

_ , Wolfram Barfuss, and Jonathan F Donges, "A thought experiment on sustainable management of the earth system," *Sustainability*, 2018, *10* (6), 1–24.

Lorenzoni, Irene, Sophie Nicholson-Cole, and Lorraine Whitmarsh, "Barriers perceived to engaging with climate change among the UK public and their policy implications," *Global Environmental Change*, 2007, *17* (3-4), 445–459.

Mastrandrea, Michael D., Katharine J. Mach, Gian Kasper Plattner, Ottmar Edenhofer, Thomas F. Stocker, Christopher B. Field, Kristie L. Ebi, and Patrick R. Matschoss, "The IPCC AR5 guidance note on consistent treatment of uncertainties: A common approach across the working groups," *Climatic Change*, 2011, *108* (4), 675–691.

Milinski, M., R. D. Sommerfeld, H.-J. Krambeck, F. A. Reed, and J. Marotzke, "The collective-risk social dilemma and the prevention of simulated dangerous climate change," *Proceedings of the National Academy of Sciences*, 2008, *105* (7), 2291–2294.

Nagel, Thomas, "Moral Luck," in "Mortal Questions," Cambridge University Press, 1979.

Vincent, Nicola A, "A Structured Taxonomy of Responsibility Concepts," 2011.

Wang, Zhen, Marko Jusup, Rui-Wu Wang, Lei Shi, Yoh Iwasa, Yamir Moreno, and Jürgen Kurths, "Onymity promotes cooperation in social dilemma experiments," *Science Advances*, 2017, (March), 1–8.