Diagnosing Social Dilemmas

Bryan Bruns

bryanbruns@bryanbruns.com

Sociologist, Independent Consultant, Greenbelt, Maryland, USA

Prepared for delivery at the Workshop on the Ostrom Workshop (WOW6) Conference

Indiana University Bloomington, June 19–21, 2019.

Abstract: Prisoner's Dilemma is only one of several possible social dilemmas where individual incentives can lead away from cooperation that would make everyone better off. Two-person two-move (2x2) games provide elementary models of the social dilemmas that have played a central role in thinking about problems of collective action. Diagnosing which kind of social dilemma may be present is important since different incentive structures pose different challenges for collective action and may require different solutions. This paper uses the Robinson-Goforth topology of payoff swaps in 2x2 games to analyze the diversity of social dilemmas; identify key questions that can distinguish between different problems of collective action even in the presence of limited information about outcomes; and discuss implications for diagnosis and potential solutions. A diagnostic flow chart provides key questions for distinguishing between social dilemmas.

Introduction

Social dilemmas pose conflicts between individual incentives and cooperation that would be mutually beneficial (Dawes 1980; Kollock 1998a; Lange et al. 2014). Different social situations pose different opportunities and risks for collective action, such as building trust, deterring defection, or contesting for advantage. Payoff matrices for two-person two-move games offer elementary models of such situations, named after stories that exemplify the issues involved, including Stag Hunts, Prisoner's Dilemma, and Chicken, (Luce and Raiffa 1957; Rapoport, Guyer, and Gordon 1976).

As a simple example of potential cooperation, farmers in northeast Thailand sometimes build small weirs to irrigate their crops (Bruns 1991). Two farmers on either side of a small stream can often do better by working together. However, the costs and benefits of a joint effort may differ. The possible incentive structures for collective action are not limited to those in Prisoner's Dilemma (Taylor and Ward 1982). A joint effort might be necessary and best for both, while one person's efforts would be wasted, posing a Stag Hunt-type problem of how to assure coordination on the best outcome (Rousseau 2004; Hume 2003; Sen 1967; Runge 1986; Skyrms 2004). The effort might be better if both share the work, but each could be tempted to free ride, trying to get the other to do most of the work while the shirker reaps more of the benefits (Olson 1971), a situation which can be modeled as a Prisoner's Dilemma. Or, either might be capable of building a weir, after which the other could irrigate with little or no extra effort, but without a weir neither might get a crop. One who is willing to be more manipulative, aggressive, deceptive, or just less inclined to effort may be able to get the other to do most or all the work, creating an incentive structure like the game of Chicken, also discussed as Hawk-Dove and Snowdrift (Kümmerli et al. 2007). The challenges to cooperation in situations like these can be modeled by matrices showing payoffs of combined choices, as in Stag Hunt, Prisoners' Dilemma, and Chicken games respectively. These situations, and their multi-person analogues such as the Tragedy of the Commons, have been the focus of research on collective action to understand why cooperation might fail even if it could make everyone better off and to examine how cooperation might develop (Olson 1971; Hardin 1968; Axelrod 1984; Elinor Ostrom 1990, 2007; Nowak and Highfield 2011).

Analysis of social dilemmas has concentrated on conflicts between individual incentives and collective benefits, particularly the temptation to defect from cooperation or to avoid risk and thereby fail to cooperate in ways that would make both better off (Dawes 1980; Dawes and Messick 2000; Kollock 1998a; Lange et al. 2014). This paper applies the Robinson-Goforth topology of payoff swaps in 2x2 games to analyze the characteristics of two-person social dilemmas and show that:

1) assurance problems, as in Stag Hunt games, can be more prevalent than the tragic incentives of Prisoner's Dilemmas;
2) uncertainty or instability in payoffs makes Prisoners' Dilemmas additionally problematic, including asymmetric variants of Prisoner's Dilemma;
3) even with limited information on payoffs, social dilemmas can be identified in terms of the presence or absence of win-win outcomes, dominant strategies, and the result of avoiding the worst payoff;

Methods

Links between social dilemmas. The topology of 2x2 games shows how elementary models of social situations are linked by changes that switch the ranking of outcomes (Robinson and Goforth 2005). Such changes link the most commonly known social dilemmas: switching the ranking of the payoffs for the top two outcomes turns Prisoners' Dilemma into a Stag Hunt; switching the lowest two payoffs turns Prisoners' Dilemma into Chicken; switching the top two payoffs turns Chicken into a win-win (no-conflict) game of Concord (Bruns 2018).

Figure 1 is an enhanced visualization of the Robinson-Goforth topology where adjoining games are linked by payoff swaps (Goforth and Robinson 2009; Bruns 2015). Twelve strict symmetric ordinal games make up a diagonal axis, whose payoffs combine to form asymmetric games. Four "layers" differ by whether the best payoffs are in the same (win-win) outcome cell, as in Stag Hunt and Concord in the lower left layer; diagonally opposite cells, as in Prisoners' Dilemma and Chicken in the upper right layer; or the same row or column, as in Samaritan's Dilemma (*HaPd/PdHa*) and Cyclic Games in the upper left and lower right layers.

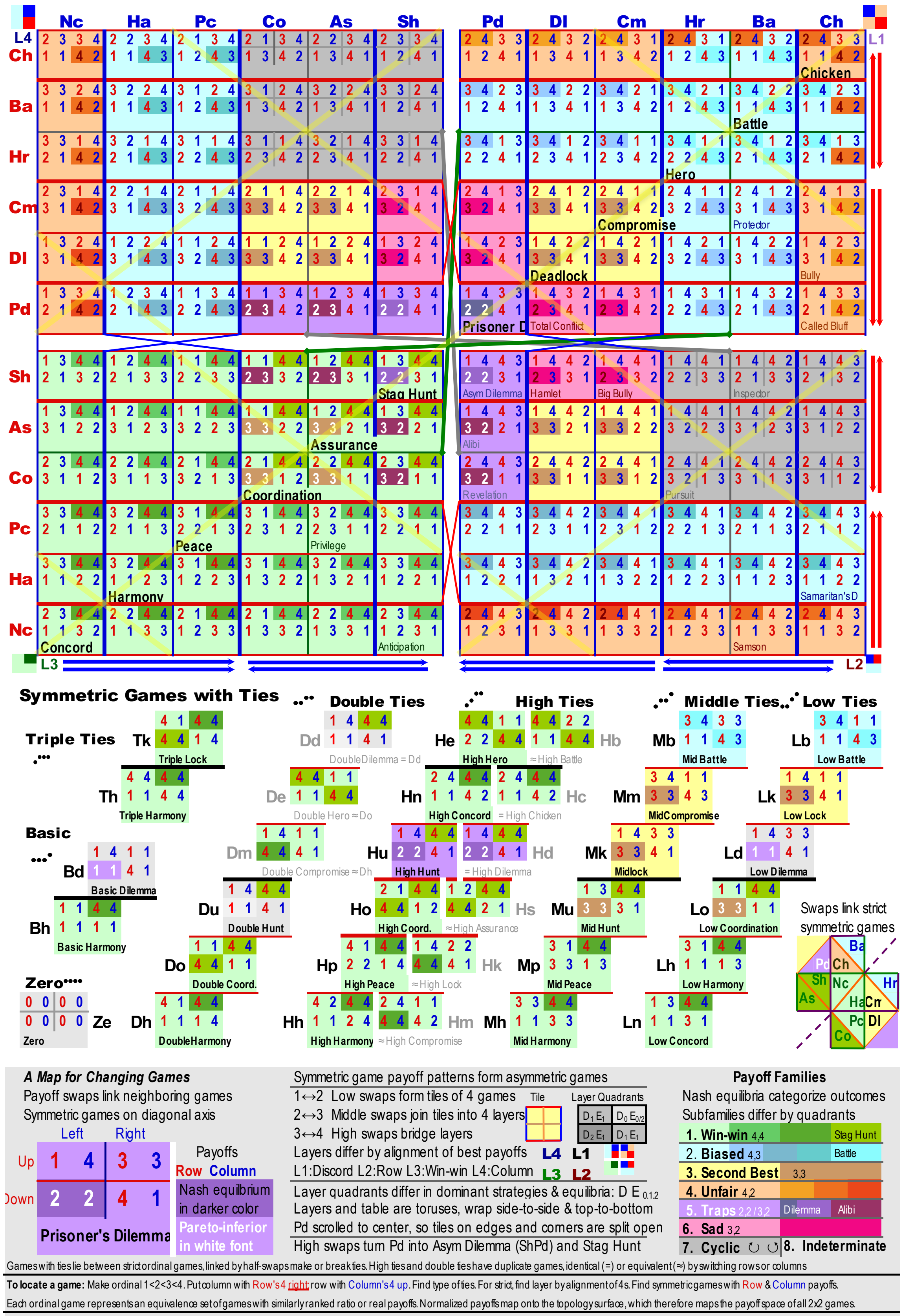

Figure 1 The Topology of 2x2 Games

Within each layer are four quadrants: in one quadrant both have dominant strategies leading to a single Nash equilibrium, a better move whatever the other does. In two adjoining quadrants, one or the other has a dominant strategy based on which the second actor has a move that is clearly best, again leading to a single equilibrium. In the other quadrant, there are (for strictly ordinal payoffs or pure strategies) either two equilibria, as in Stag Hunt and Chicken or else no equilibrium, as in cyclic games where one or the other would always prefer to move to a different outcome.

The topology was originally developed for strict ordinal games, where the four possible outcomes can be strictly ranked by preference, as illustrated by payoffs from one to four. However, the topology extends to include games with ties (indifference between outcomes) (Robinson, Goforth, and Cargill 2007). Payoffs can also be measured on interval (ratio) scales, as in a five- or seven-point Likert scale from best to worst or strong agreement to strong disagreement, as well as cardinal (real) scales, as with payoffs measured in terms of money. Such payoff values can also be normalized and mapped onto the topology (Goforth and Robinson 2012; Bruns 2010, 2015). Thus, the topology, and its visualizations reveal many key characteristics of the larger space of possible 2x2 games, including the incentive structures of social dilemmas.

The topology maps the space of possible 2x2 games. To the extent that payoffs are generated randomly, then payoff structures will tend to occur in the proportions shown in the topology (Simpson 2010). For actual situations, the frequency of different incentive structures is an empirical question, which may be related to resource characteristics, production functions for joint action, and other factors. However, unless or until more specific information is available about a situation, the proportions of games in the space of possible games, as well as the frequency with a random distribution of payoffs, offer a reasonable default expectation for discussing the possible prevalence of different situations.

The topology shows relationships between 2x2 games, including different types of social dilemmas. As Robinson and Goforth (2005) found, social dilemmas are neighbors in the topology. They form a compact connected region, including asymmetric dilemmas formed by combining payoffs from symmetric games. Prisoner's Dilemmas and Stag Hunts form a contiguous region of sixteen games, each of which has a Pareto-inferior equilibrium.

Results

Instability in social dilemmas. As identified by Robinson and Goforth, the Prisoner's Dilemmas and their neighbors form the most diverse region within the topology, where even changes in the lowest two payoffs can change the number of equilibria and the payoffs at equilibrium. Switching the lowest two payoffs in the asymmetric "Alibi Game" (*ShPd*) adjoining Prisoner's Dilemma, can turn it into a cyclic game with no equilibrium, or can yield a game (Called Bluff, *PdCh*) with a single highly unequal (4,2) equilibrium outcome. Switching the top two payoffs for one person can create a Stag Hunt with two equilibria, where both could get their best payoffs but may instead get stuck at second-worst. Switching middle payoffs yields poor (3,2) results with no possibility of both doing better. Swapping both middle payoffs forms a game where both get second-best (Deadlock, *Dl*). Thus, the Prisoner's' Dilemmas not only pose the problems of Pareto-inferior outcomes but are also highly sensitive to changes in payoffs, which can result in very different incentive structures.

Rival favorites and the uniqueness of Chicken. Chicken lies at the edge between Prisoner's Dilemmas and the set of games with rival equilibria. Two people may want to do something together, for example to watch a movie, have different favorites. This situation was encapsulated in the "Battle of the Sexes" story, also discussed as "Bach or Stravinsky" in an attempt to avoid gender stereotypes (Luce and Raiffa 1957; Osborne and Rubinstein 1994). Two versions of this problem are shown by the symmetric games that Rapoport (1967) analyzed based on whether the player moving from the second-worst outcome (and from a maximin strategy that avoids the worst outcome) would get their best outcome in Leader; or get second-best in Hero. Volunteer's Dilemma (Diekmann 1985), where everyone agrees something should be done, but wants others to incur the costs of action poses a similar challenge about unequal costs and benefits,. Volunteer's Dilemma has ties between middle payoffs and so lies between Chicken and Leader in the topology. Payoffs from the strict games with rival equilibria: Chicken, Leader (Battle), and Hero, combine to form nine rivalrous games.

Chicken is usually discussed as a social dilemma, with a "cooperative" (3,3) outcome from which each is tempted to defect. However, it also has two rival equilibria, where one gets their best outcome and the other second-worst. As in the story of two cars racing towards each other, stubborn pursuit of the best outcome by not swerving could lead to both getting the worst result. The rival equilibria are not Pareto-inferior to the cooperative outcome. So, in addition to a social dilemma of conflicts between individual and collective benefits, Chicken poses additional challenges for collective action of rivalry between inequitable equilibria and a common desire to avoid the worst outcome. Of the rivalrous "battles" where alternative equilibria offer unequal payoffs, only Chicken has a potential outcome where both can get second-best, but would both be tempted to defect. Thus, Chicken can be seen as an unusual outlier in these rivalrous games, as well as being unique as a social dilemma in having rival equilibria.

Prevalence of social dilemmas and other problems. Most game theory research focuses on social dilemmas, particularly Prisoner's Dilemma, Chicken, and stag hunts (assurance problems). In terms of the space of possible games, as well as the likely frequency of different kinds of situations if payoffs occur randomly, then within the social dilemmas, Stag Hunts (9) are slightly more likely to occur than Prisoners' Dilemmas (7). However, other games besides these social dilemmas would be far more likely to occur, as shown in Table 1. Symmetric games are only one twelfth of the total while the vast majority of games are asymmetric. Most games do not have an equilibrium with equal payoffs. Games with equal payoffs at equilibrium are composed of the layer of win-win games, the family of second-best games, and three of the Prisoners' Dilemmas with equal payoffs at equilibrium, totaling slightly over a third of the games. Cyclic games have no equilibrium in pure strategies and make up another eighth of the total. Most games, a bit less than two-thirds, either yield unequal payoffs at equilibrium or lack an equilibrium. Thus, the preoccupation of research with a few symmetric social dilemmas may offer a misleading guide to what kinds of situations are most likely.

Table 1 Proportions of possible games

	Percentage	Fraction
Social Dilemmas	12%	17/144
Symmetric	8.3%	12/144
Equilibrium with equal payoffs	35.5%	51/144
• **Win-win (4,4)**	25%	36/144
• **Second best (3,3)**	8.3%	12/144
• **PD (2,2)**	2%	3/144
Cyclic-No equilibrium in pure strategies	12.5%	18/144
Equilibrium with unequal payoffs	52%	75/144

Discussion of social dilemmas typically assumes equality between players, as in symmetric games, and equality in outcomes, as with cooperative outcomes and Pareto-inferior equilibria in Prisoners' Dilemma and Stag Hunt, as well as in the cooperative outcome in Chicken. In terms of default expectations for randomly distributed payoffs, unequal payoffs at equilibrium, as in Chicken, would be much more likely to occur than pareto-inferior equilibria. Thus, in terms of a default distribution of games with randomly generated payoffs, inequality in outcomes could be a more prevalent problem than failure to achieve mutual gains through cooperation.

Diagnosing social dilemmas with incomplete information on payoffs. Some changes in payoffs may leave the outcomes and the basic structure and challenge for collective action unchanged or relatively similar. These show up as neighboring games in the visualization of the topology. This fits with Robinson and Goforth's assumption that swaps in the lowest two payoffs are likely to be the least significant changes. Looking at social dilemmas as a region of games connected by payoff swaps helps show how social dilemmas might be identified even if some outcome payoffs are uncertain or variable. For Stag Hunts, the key characteristic is whether there are two possible equilibria, one of which allows both to get their best outcome, while the other payoffs can occur in a variety of configurations. This occurs in the square block of nine games.

For the tragic incentive structure of Prisoner's Dilemma, the key characteristic is the presence of a dominant strategy leading to a Pareto-inferior Nash equilibrium, which occurs in the L-shaped set of seven games. The region with stag hunts and prisoners' dilemmas contains sixteen games with Pareto-inferior equilibria. Thus, social dilemmas occur even when players face different incentive structures in asymmetric games or get unequal payoffs.

For analyzing social situations and identifying social dilemmas, there is no need to assume symmetry in payoff structures or equality in outcomes. Figure 2 provides a flow chart with diagnostic question concerning whether there is a Pareto-inferior equilibrium and a win-win outcome, to distinguish assurance/stag hunt type problems from the tragic incentives of Prisoner's Dilemmas; or a second-best outcome from which both would like to defect but then would lead to the worst outcome for both, creating a Chicken-type situation. These key questions are sufficient to identify and distinguish social dilemmas, including those with asymmetric payoff structures and asymmetric Stag Hunts and Prisoner's Dilemmas where equilibrium payoffs are unequal. Figure 3 provides a more complete flow chart to diagnose the full set of possible payoff families for 2x2 games.

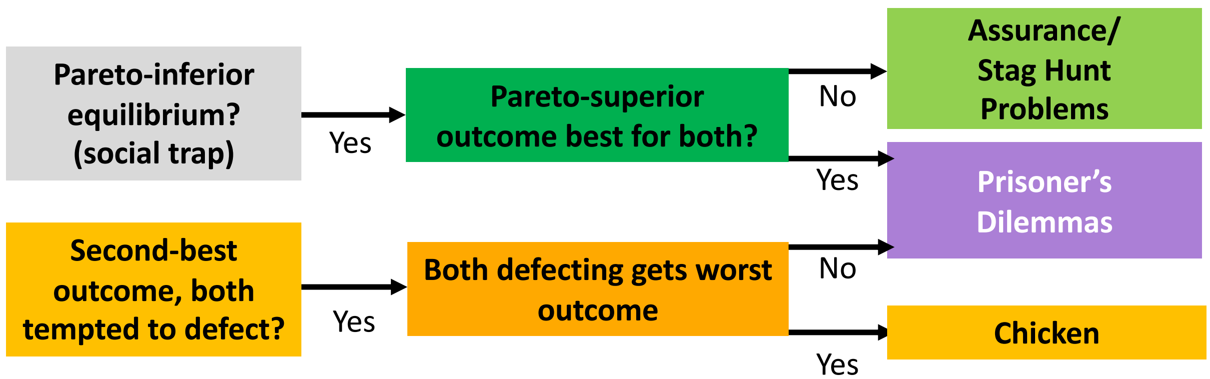

Figure 2 Diagnosing Social Dilemmas

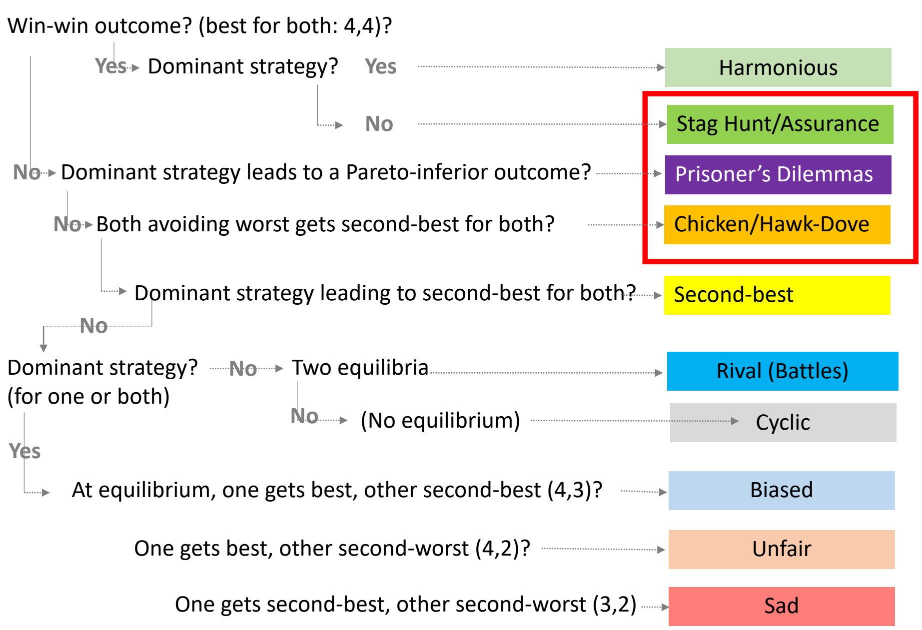

Figure 3 Diagnosing Social Situations

Discussion

Assurance may be a more prevalent problem than tragic incentives. Most game theory research on social dilemmas has focused on Prisoner's Dilemmas, often on the particular payoff structure used by Axelrod (1984; Robinson and Goforth 2005). However, rather than a few exemplary games or only ordinal payoff structures, social dilemmas can be better understood as a region of potential payoffs and incentive structures, best characterized by the presence of equilibrium outcomes with Pareto-inferior payoffs. Within this region, assurance problems, with two Nash equilibria may be more frequent than tragic incentive structures where a dominant strategy leads to a single equilibrium. While Prisoners' Dilemma has received far more attention from researchers, Stag Hunt situations may actually be more frequent, as conjectured by Kollock (1998a). As argued by Skyrms (2004, 2014), Stag Hunts may be more important in understanding collective action, cooperation, and social structure.

Instability makes social dilemmas hard to identify and solve. In the topology, games where both get best or second best form a *sea of stability* where outcomes are relatively robust to changes in payoffs (darker green, blue, and yellow areas in Table 1) (Bruns 2015). However, in the social dilemma region if there is uncertainty or instability (noisy or trembling payoffs) even for the lowest two payoffs, then it may be hard to be sure whether a social dilemma is present, particularly for Prisoners' Dilemmas. Similarly, if payoffs are variable then shifts between social dilemmas and other situations and outcomes are likely. Potential changes in Prisoner's Dilemma may lead to other difficult situations, with poor payoffs, highly unequal payoffs, or no equilibrium, and thus different problems and opportunities for collective action. Stag Hunts are somewhat more robust to changes, and outcomes of payoff changes less diverse than Prisoners' Dilemmas, but changes in payoffs are still more prone to lead to neighboring games with more diverse equilibrium outcomes than games in the sea of stability.

Diagnosis. Incomplete or uncertain information about some specific outcomes does not necessarily prevent diagnosing social dilemmas, for which the presence of a Pareto-inferior outcome and an alternative where both could do better is a key characteristic. Furthermore, even with limited information, and uncertain or variable outcomes, the presence of a win-win equilibrium where both could get their best result helps distinguish assurance problems from the potentially tragic incentives for defection leading to a social trap in Prisoner's Dilemmas. Rather than simply saying there is a social dilemma, it is feasible and useful to identify and distinguish assurance problems where both could do best from the tragic incentives and inferior results of Prisoner's Dilemma-type situations.

For repeated interaction in the Prisoners' Dilemma situation, cooperation can transform expected payoffs into a Stag Hunt structure (Skyrms 2004). Players then face a problem of assuring coordination. Depending on the actual payoffs, coordination could happen through choosing moves so that both get at least second-best, or by taking turns getting an outcome that allows a higher cumulative payoff. If play is repeated and payoffs can be measured in a comparable way, then a key question is whether taking turns getting unequal payoffs could yield a better total result compared to the “cooperative” move (Goforth and Robinson 2012). Games beyond a “reconciliation line” where taking turns pays off better pose a somewhat different problem than those where repeated choice of the “cooperative” (second-best) outcome yields superior results. For repeated interaction, the potential to transform expected payoffs from a Prisoner’s Dilemma to a Stag Hunt structure offers an additional reason why assurance problems may be more common and more important in the development of social order. Furthermore, even in single-shot Prisoner’s Dilemma situations, participants may perceive and act as if they were playing an assurance game (Kollock 1998a, 1998b).

Conclusions

The topology of 2x2 games can be applied to identify social dilemmas not just in terms of a few exemplary games but as a region of similar social situations characterized by the presence of Pareto-inferior equilibria, in most of which the two players face different incentive structures and for many of which equilibrium payoffs are not only Pareto-inferior but also unequal. Even with incomplete information, analysts can identify social dilemmas and distinguish assurance problems from situations with tragic incentives. Social dilemmas, particularly Prisoner’s Dilemmas, are challenging not just due to the frustration of inferior outcomes, but also their sensitivity to changes in payoffs, making them harder to identify and to solve.

Assuring coordination is likely to be a more prevalent problem for governance than tragic temptations to defect from cooperation, as a default expectation based on the space of possible payoff structures and the likely frequency of games if payoffs occur randomly. Furthermore, problems of unequal opportunities and outcomes are likely to be much more prevalent than failure to achieve cooperation that is better for both. Diagnosis of social situations should pay attention to distinguishing between different challenges for cooperation, including the difference between assuring coordination and deterring selfish defection, as well as the potential problems of and remedies for inequality in opportunities and outcomes.

References

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Bruns, Bryan. 1991. “The Stream the Tiger Leaped: A Study of Intervention and Innovation in Small Scale Irrigation Development in Northeast Thailand.” Cornell University.

———. 2010. “Navigating the Topology of 2x2 Games: An Introductory Note on Payoff Families, Normalization, and Natural Order.” *Arxiv Preprint ArXiv:1010.4727*.

———. 2015. “Names for Games: Locating 2 × 2 Games.” *Games* 6 (4): 495–520. https://doi.org/10.3390/g6040495.

———. 2018. “From Tragedy to Win-Win Transforming Social Dilemmas in Commons.” In . Georgetown Law Center, Washington DC, USA.

Dawes, Robyn M. 1980. “Social Dilemmas.” *Annual Review of Psychology* 31 (1): 169–193.
Dawes, Robyn M., and David M. Messick. 2000. “Social Dilemmas.” *International Journal of Psychology* 35 (2): 111–116.
Diekmann, Andreas. 1985. “Volunteer’s Dilemma.” *Journal of Conflict Resolution*, 605–610.
Goforth, David, and David Robinson. 2009. “Dynamic Periodic Table of the 2x2 Games: User’s Reference and Manual.”
———. 2012. “Effective Choice in All the Symmetric 2x2 Games.” *Synthese*, 1–27.
Hardin, Garrett. 1968. “The Tragedy of the Commons.” *Science*, no. 162: 1243–48.
Hume, David. 2003. *A Treatise of Human Nature*. https://www.gutenberg.org/ebooks/4705.
Kollock, Peter. 1998a. “Social Dilemmas: The Anatomy of Cooperation.” *Annual Review of Sociology* 24 (1): 183–214.
———. 1998b. “Transforming Social Dilemmas: Group Identity and Cooperation.” *Modeling Rational and Moral Agents*, 186–210.
Kümmerli, Rolf, Caroline Colliard, Nicolas Fiechter, Blaise Petitpierre, Flavien Russier, and Laurent Keller. 2007. “Human Cooperation in Social Dilemmas: Comparing the Snowdrift Game with the Prisoner’s Dilemma.” *Proceedings of the Royal Society B: Biological Sciences* 274 (1628): 2965–2970.
Lange, Paul Van, Daniel P. Balliet, Craig D. Parks, and Mark van Vugt. 2014. *Social Dilemmas: Understanding Human Cooperation*. Oxford University Press.
Luce, Robert Duncan, and Howard Raiffa. 1957. *Games and Decisions: Introduction and Critical Survey*. Courier Dover Publications.
Nowak, Martin, and Roger Highfield. 2011. *SuperCooperators: Altruism, Evolution, and Why We Need Each Other to Succeed*. Free Press.
Olson, Mancur. 1971. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, MA: Harvard University Press.
Osborne, Martin J., and Ariel Rubinstein. 1994. *A Course in Game Theory*. MIT Press.
Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press.
———. 2007. “A Diagnostic Approach for Going beyond Panaceas.” *Proceedings of the National Academy of Sciences* 104 (39): 15181.
Rapoport, A. 1967. “Exploiter, Leader, Hero, and Martyr: The Four Archetypes of the 2 Times 2 Game.” *Behavioral Science* 12 (2): 81.
Rapoport, A., M. Guyer, and D. G Gordon. 1976. *The 2 x 2 Game*. Univ of Michigan Press.
Robinson, David, and David Goforth. 2005. *The Topology of the 2x2 Games: A New Periodic Table*. London: Routledge.
Robinson, David, David Goforth, and Matt Cargill. 2007. “Toward a Topological Treatment of the Non-Strictly Ordered 2x2 Games.” *Working Paper*.
Rousseau, Jean-Jacques. 2004. *A Discourse upon the Origin and the Foundation of the Inequality among Mankind*. http://www.gutenberg.org/ebooks/11136.
Runge, Carlisle Ford. 1986. “Common Property and Collective Action in Economic Development.” *World Development* 14 (5): 623–35.

Sen, A. K. 1967. "Isolation, Assurance and the Social Rate of Discount." *The Quarterly Journal of Economics* 81 (1): 112–124.

Simpson, J. 2010. "Simulating Strategic Rationality." Ph.D. Dissertation, Edmonton: University of Alberta.

Skyrms, Brian. 2004. *The Stag Hunt and the Evolution of Social Structure*. Cambridge Univ Pr.

———. 2014. *Social Dynamics*. Oxford University Press.

Taylor, M., and H. Ward. 1982. "Chickens, Whales, and Lumpy Goods: Alternative Models of Public Goods Provision." *Political Studies* 30 (3): 350–370.