

# THE PROBLEM OF SOCIAL ORDER: EGOISM OR AUTONOMY?

Arnout van de Rijt and Michael W. Macy

## ABSTRACT

*Individual rationality sometimes leads to collectively irrational outcomes, a fundamental problem in the social and life sciences that has attracted sustained attention from experimentalists in sociology, psychology, biology, and economics. But what is it about individual rationality that sometimes gets us into trouble? Is the problem the egoistic pursuit of individual self-interest? Or does the problem with individual rationality lie elsewhere? To find an answer, this chapter closely examines the theoretical and experimental literature on social dilemmas, to see how researchers identify the source of the problem. The review suggests that the prevailing theory wrongly points to egoism as the problem. Failing to do what is best for everyone can also happen among rational altruists, and sometimes egoism is needed to prevent it. The chapter concludes by pointing to what we believe is the fundamental problem – a tension not between individual self-interest and collective welfare, but between individual autonomy and collective interdependence.*

I put for a general inclination of all mankind a perpetual and restless desire of power after power, that ceaseth only in death. And the cause of this is not always that a man hopes for a more intensive delight than he has already attained to, or that he cannot be

---

**Altruism and Prosocial Behavior in Groups**  
**Advances in Group Processes, Volume 26, 25–51**  
**Copyright © 2009 by Emerald Group Publishing Limited**  
**All rights of reproduction in any form reserved**  
**ISSN: 0882-6145/doi:10.1108/S0882-6145(2009)0000026005**

content with a moderate power, but because he cannot assure the power and means to live well, which he hath present, without the acquisition of more.

– Thomas Hobbes, *Leviathan*, Part I, Chapter 11

The Hobbesian problem of order is one of the founding questions of the social sciences: how can rational individuals become trapped in collectively irrational outcomes and how might they escape? It is a question with compelling practical ramifications and profound theoretical importance, one that occupies a central place in several disciplines, including the study of collective action in sociology, public choice in political science, public goods economics, evolutionary game theory in biology and ethology, and psychological research on social dilemmas.

## THE BACK OF THE INVISIBLE HAND

Across the social and life sciences, the applications are different but the problem of order is posed in the same way. It is what Russell Hardin (1982, p. 6) calls “the back of the invisible hand,” a reference to Adam Smith’s (2008 [1789], p. 22) theory that “It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest.” Hardin’s point is that the invisible hand works only so long as self-interested actions have positive externalities for others, but when the externalities are negative, the pursuit of self-interest can have collectively disastrous consequences. “The narrow rationality of self-interest that can benefit us all in market exchange can also prevent us from succeeding in collective endeavors” (1982, p. 6). This tension between self-interest and collective interest frames the representation of the problem of order in three prominent lines of research – on collective action, the evolution of cooperation, and social dilemmas.

(1) *Collective action*. Prior to Olson, collective action was typically explained by pointing to the shared interest of the members of a group in obtaining the collective good. Olson pointed out the logical flaw in this line of reasoning. “But it is *not* in fact true that the idea that groups will act in their self-interest follows logically from the premise of rational and self-interested behavior. It does *not* follow, because all of the individuals in a group would gain if they achieved their group objective” (1965, p. 2). The problem is two-fold. The “free-rider problem” is that a member of the collective can benefit even if she does not contribute. The “efficacy problem” is that she may not benefit even if she does contribute. The logic of collective action thus leads each member of an interest group to the same mutually

ruinous conclusion: I may get little or no additional benefit from my own efforts should I choose to contribute, yet I will enjoy the benefits of others' efforts even if I fail to contribute. Olson illustrates the two problems. "The rational individual in the economic system does not curtail his spending to prevent inflation . . . because he knows, first, that his own efforts would not have a noticeable effect, and second, that he would get the benefits of any price stability that others achieved in any case. For the same two reasons, the rational individual in a sociopolitical context will not be willing to make any sacrifices to achieve the objectives he shares with others" (Olson, 1965, p. 166). Hence, commuters are generally unwilling to downsize their sport utility vehicles (SUVs), much less suffer the inconvenience of public transportation, and those who do are not acting rationally, even as all choke on congested freeways. In the absence of selective incentives, rational actors should not participate in collective action, unless the benefit of participation exceeds the cost, which becomes increasingly unlikely as the size of the group increases. But if rational self-interest leads to collectively irrational outcomes, how then is social order possible?

(2) *The evolution of cooperation.* The puzzle reappears in biology and ethology as the enigma of cooperation among organisms competing to survive and reproduce. "The logic of natural selection," to paraphrase Olson, is what Tennyson in his 1849 memorial to Arthur Henry Hallam, called "Nature, red in tooth and claw." Just as Olson led social scientists to reject collective interests as the explanation for collective action, evolutionists have mostly rejected "group selection" as the explanation for cooperation in nature: Although a cooperative species might have higher survival chances, mutant "free-riders" would enjoy this same benefit without the cost of contribution. They have proposed two alternative explanations, kin selection and conditional cooperation. In John Maynard Smith's (1982) pioneering work on evolutionarily stable strategies, game theory is used to explain how cooperation might nevertheless evolve when interaction is repeated under the threat of future retaliation. Alternatively, pro-social behavior toward genetically similar others may decrease an individual's chances of survival while increasing that of the gene that both individuals share. This happens, for example, if the two individuals are kin, in which case biologists speak of "kin selection." This idea was later popularized by Richard Dawkins (1976) as the theory of the "selfish gene."

(3) *Social dilemmas.* In social psychology, the problem of order is studied using laboratory experiments that present participants with a social dilemma. The term "social dilemma" was originally introduced by economist Gordon Tullock in a 1974 book of the same name. "In general," Tullock

writes, “conflict uses resources, hence it is socially inefficient, but entering into the conflict may be individually rational for one or both parties . . . The social dilemma, then, is that we would always be better off collectively if we could avoid playing this kind of negative sum game, but individuals may make gains by forcing such a game on the rest of us” (Tullock, 1974, p. 2). A negative sum game is one in which all players have an interest in avoiding an outcome that is costly for all, as in the Hobbesian war of all against all. Contemporary examples include global warming, the depletion of the world’s fisheries, the escalation of ethnic conflict, and nuclear proliferation.

Tullock’s definition is not very clearly stated and has largely disappeared from the literature. Psychologist Robyn Dawes tightened and extended Tullock’s definition in a way that avoids this limitation. Today, many experimental social psychologists accept Dawes’ specification (Komorita & Parks, 1994; Schroeder, 1995; Sell & Son, 1997; Smithson & Foddy, 1999). According to [5]Dawes (1980, p. 169), a social dilemma has “. . . two simple properties: a) each individual receives a higher payoff for a socially defecting choice (e.g., having additional children, using all the energy available, polluting his or her neighbors) than for a socially cooperative choice, no matter what the other individuals in society do, but b) all individuals are better off if all cooperate than if all defect.” Following Dawes, Ostrom and Walker (2003, p. 93) define a social dilemma as a game “in which the outcome that is achieved when subjects follow a strategy based on pure self-interest is inferior to an outcome that would be achieved if subjects chose strategies based on the interests of all group members.”

The common thread in Olson, Maynard-Smith, Tullock, Dawes, and Ostrom and Walker is the tension between the individual pursuit of self-interest and the attainment of the greater good. This chapter argues that the role of self-interest in the problem of order has been overstated. Failing to do what is best for everyone can also happen among rational altruists and sometimes egoism is needed to prevent it. Instead, it is proposed that the problem of order is one of *autonomy*. Autonomy means that there is no central authority that can constrain individual choices. The problem of order arises because *groups do not act, their members do*. Because of this distribution of control over the collective outcomes, the alignment of individual decisions with the outcome that is best for all is inherently problematic. Mutually destructive behavior can happen because of too much altruism as well as too much egoism. And whether egoist or altruist, autonomous individuals need to coordinate their actions to create a critical mass in a nascent social movement, to avoid congestion on an overcrowded highway, or to standardize on a superior technology for listening to music or

watching movies. In these situations, the fundamental problem is the interdependence of decision-makers whose choices have outcomes that depend in part on the autonomous choices of others.

The argument unfolds through a series of steps. To start with it is shown how the most common definition of social dilemmas carves away the coordination problem by imposing the requirement that the optimal choice for each actor does not depend on the choices of others. In doing so, the concept becomes much narrower than generally recognized – so narrow that famous examples of collective action in the sociological literature do not fit. Even the experimental designs in a surprisingly large number of laboratory studies of social dilemmas turn out not to pose a social dilemma according to the standard definition used in those very studies. Clearly, the authors did not intend to exclude their own experiments, but their use of a definition that assumes away the problem of coordination exemplifies a widespread misconceptualization in which the pursuit of self-interest is equated with individual rationality.

How the efforts to broaden the definition have not gone far enough are then shown. They have dropped the requirement that the best response does not depend on the strategies of others, but they have kept the requirement that the best response is an equilibrium. Using the example of Nuclear Chicken, it is found that the broader definition points incorrectly to bad luck as the source of the problem, instead of the possibility for miscalculation.

In conclusion, a turn is made from the narrowness of the definition of individual rationality to the narrowness of the definition of collective rationality, which excludes solutions to social dilemmas that rely on taking turns. Here again, the problem is one of coordination, not self-interest. Altruists depend on traffic lights no less than egoists. The problem of order can arise in a population of altruists as well as egoists, but it cannot arise in a population in which the Leviathan makes it possible to efficiently coordinate individual strategies. In short, the fundamental problem is not the egoism of rational individuals but their autonomy.

## **SOCIAL DILEMMAS: EGOISM PURE AND SIMPLE**

The narrowness of the prevailing definition of a social dilemma can be demonstrated using game theory. Fig. 1 depicts several  $2 \times 2$  normal-form one-shot games, meaning there are two players, each with two *strategies*, who interact once, with no expectation of ever doing so again. A strategy is simply

	$C_2$	$D_2$
$C_1$	$R_1, R_2$	$S_1, T_2$
$D_1$	$T_1, S_2$	$P_1, P_2$

Prisoner's Dilemma:  $T_i > R_i > P_i > S_i$   
 Stag Hunt:  $R_i > T_i > P_i > S_i$   
 Chicken:  $T_i > R_i > S_i > P_i$   
 Deadlock:  $T_i > P_i > R_i > S_i$   
 Red Queen:  $T_1 > R_1 > P_1 > S_1, R_2 > T_2 > P_2 > S_2$   
 Coordination:  $R_i = P_i > T_i = S_i$   
 $i = \{1, 2\}$

Fig. 1. The Normal Form Representation of  $2 \times 2$  Games.

a plan of action for the game, such as whether or not to “cooperate” with the partner or “defect.” The games in Fig. 1 are in *normal form* (represented as a matrix) to indicate that each player must choose a strategy without knowing what the partner has chosen.<sup>1</sup> More generally, a game consists of two or more players  $i$ , each with a set of two or more strategies (such as  $C_i$  or  $D_i$ ), and a utility function that assigns an individual *payoff*<sup>2</sup> (such as  $T_i, R_i, P_i$ , or  $S_i$ ) to each strategy *profile*.<sup>3</sup> The key idea is that the payoffs for a given strategy can depend on the strategies of other players. A Nash *equilibrium* is a profile of strategies that are *best responses*<sup>4</sup> to each other, which means that no player has an incentive to unilaterally change strategy. An outcome is Pareto efficient if any further improvement for some would come at some others’ expense. Hence, an outcome is Pareto deficient if some other outcome makes some better off and no one worse off. Pareto efficiency is a widely accepted standard for deciding if an outcome is collectively rational.

Fig. 1 includes six different one-shot  $2 \times 2$  games that illustrate how rational individuals can become trapped in collectively irrational outcomes – Prisoner’s Dilemma, Stag Hunt, Red Queen, Chicken, Deadlock, and Coordination.<sup>5</sup>

In the Prisoner’s Dilemma,  $(D_1, D_2)$  is a dominant-strategy Nash equilibrium. No matter what the other player does, defecting yields a higher payoff than cooperating (since  $T_i > R_i$  if the other cooperates and  $P_i > S_i$  if the other defects). The name comes from the game’s original anecdote, in which two suspects are offered a lighter sentence in exchange for implicating their partner. Dawes’ definition of a social dilemma is equivalent to a one-shot game of Prisoner’s Dilemma with continuous instead of binary choice and many players instead of just two. Recall the two properties identified by Dawes: Requirement “a” (a higher payoff for defecting no matter what others do) translates as “defection is the dominant strategy for all players” (or  $T_i > R_i$  and  $P_i > S_i$  in Fig. 1) and “b” (everyone is better off if all cooperate than if all defect) translates as “these dominating strategies ‘converge on a deficient equilibrium’” (Dawes, 1991, p. 54). “An outcome is deficient,”

Dawes continues, “when that outcome is less preferred by all choosers than some other outcomes” (or  $R_i > P_i$  in Fig. 1, yielding  $T_i > R_i > P_i > S_i$ ).

Of the six games depicted in Fig. 1, Prisoner’s Dilemma is unique in having a dominant strategy equilibrium that is Pareto deficient when players seek to maximize their own payoff. Stag Hunt, Chicken, Coordination, and Red Queen do not have dominant strategies, and the dominant strategy equilibrium in Deadlock is not Pareto deficient when the players maximize self-interest. The existence of a deficient dominant-strategy equilibrium is important because, like the prisoners used to illustrate the dilemma, it means the players have no escape. Even if prisoners are allowed to discuss and coordinate their strategies before deciding whether to confess, the outcome is the same. Players’ beliefs about others’ intentions are entirely irrelevant. Hence, the game is an ideal type for showing how collective irrationality can arise solely through the aggregation of rational self-interest. This makes the Prisoner’s Dilemma game very useful as a clear and precise representation of the problem of order posed by the rational pursuit of individual self-interest. Nevertheless, social dilemma researchers have adopted this definition without giving sufficient attention to two other ways in which individual rationality can lead to a collectively irrational outcome: miscoordination and altruism.

### **TOO MANY COOKS, TOO FEW CHICKENS, BUT NO SOCIAL DILEMMA**

In Stag Hunt ( $R_i > T_i \geq P_i > S_i$ ), both mutual cooperation and mutual defection are Nash equilibria, but neither is a dominant strategy equilibrium. The name comes from Rousseau, who describes the dilemma faced by two hunters who must cooperate to catch a stag. Each must decide whether to abandon the effort and grab a hare, in which case, the partner goes hungry. Both prefer half a stag to all of a hare, but each worries that the partner may defect. The best response is to match the other player’s strategy, cooperating when the partner cooperates (since  $R_i > T_i$ ) and defecting when the partner defects (since  $P_i > S_i$ ). Mutual cooperation is the only Pareto-efficient outcome, but the players may miscoordinate. Rational self-interest can lead to a collectively irrational outcome based on the belief that the partner will defect, whether or not the belief is correct. Yet this is not a social dilemma according to the prevailing definition.

In Chicken ( $T_i > R_i > S_i > P_i$ ),  $(C_1, D_2)$  and  $(D_1, C_2)$  are Nash equilibria. The name comes from the infamous bragging-rights contest played by

adolescents who drive at one another to see who will swerve rather than risk collision. Each prefers public humiliation to collision but knows the partner does as well and is therefore tempted to take the gamble. Here, a mismatch is the best response, cooperating (i.e., swerving) when the other defects (since  $S_i > P_i$ ) and defecting when the other cooperates (since  $T_i > R_i$ ). Rational self-interest leads to a collectively irrational outcome based on the false belief that the partner is “chicken.” A familiar example is the Cuban Missile Crisis in which Kennedy and Khrushchev stood eyeball to eyeball for five days, waiting for the other to blink, while the world held its breath. Yet this game does not pose a social dilemma according to the prevailing definition.

In Deadlock ( $T_i > P_i > R_i > S_i$ ),  $(D_1, D_2)$  is a dominant strategy Nash equilibrium. As in the Prisoner’s Dilemma, self-interested players are better off defecting, no matter what the partner does (because  $T_i > R_i$  and  $P_i > S_i$ ), as indicated in the upper panel of Fig. 2. But unlike the Prisoner’s Dilemma, when everyone defects, the outcome is Pareto efficient. As a result, Deadlock has been entirely ignored in the literature on social dilemmas. In fact, Deadlock is one of the most important of the games in Fig. 1 in that it demonstrates in a clear and compelling way that self-interest is not necessary for individual rationality to lead to a collectively irrational outcome. The lower panel of Fig. 2 shows what happens when Deadlock is played by two altruists who care only for the partner’s payoffs and ignore their

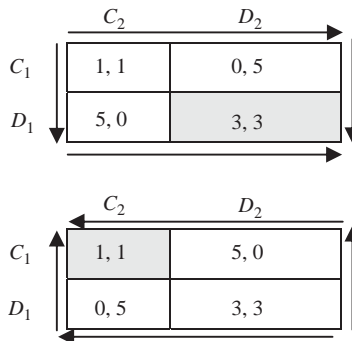


Fig. 2. The Game of Deadlock. In the upper panel, the game is played by egoists. The arrows indicate best replies and converge at the unique Nash equilibrium at  $(D, D)$ , which is Pareto efficient. In the lower panel, the game is played by altruists who optimize on the partner’s payoffs, which now appear in the designated location for each player. The arrows now converge at a dominant strategy Nash equilibrium at  $(R, R)$ , which is Pareto deficient.

own (Heckathorn, 1991, 1996). Each player then chooses “C,” leading to an outcome that neither prefers. Deadlock thus demonstrates that the problem of order can arise even among rational altruists. It is individual rationality that gets them into trouble, not egoism. In fact, a bit more egoism is the solution. An example is “too many cooks in the kitchen,” which can lead to the meal ending up on the floor and the hungry altruists thinking perhaps they should have called out for pizza.

The Coordination Game ( $R_i = P_i > T_i = S_i$ ) proves that collectively irrational outcomes can obtain even when there is no tension between individual and collective interests. In this game, individual and collective interests are perfectly aligned. If player 1 loses, so does player 2, and if player 1 wins, player 2 wins as well. Thus, it makes no difference whether the players are egoists seeking to win or altruists seeking to let the partner win – the outcome is the same. A familiar example is passing someone in a doorway. Whether egoist or altruist, each player is indifferent between left and right and wants only to choose whatever the partner chooses, whether for their own sake or the partner’s. Nevertheless, both egoists and altruists run the risk of guessing wrong and bumping together. The problem is that decision-making is decentralized. Whether to maximize the payoff to self or other, each wants to do what the other does, but when play is simultaneous, neither knows what the other will do.

The Coordination Game provides a compelling demonstration of the problem of order that can arise among interdependent but autonomous actors. Stag Hunt and Chicken show how miscoordination can occur among egoists, Deadlock shows how the problem can arise among altruists, but the Coordination Game shows how it makes no difference whether the players are egoists or altruists. In these four games, the fundamental problem is not the incompatibility between individual self-interest and collective welfare; it is the incompatibility between individual autonomy and collective interdependence.

The five games considered so far are pure types – the payoff inequalities are the same for both players (although their payoffs may differ). Asymmetrical payoffs also open up the possibility for hybrid games. For example, in the Red Queen dilemma illustrated in Fig. 3, Column is playing Stag Hunt but Row prefers rabbit to deer, giving that player a payoff inequality identical to Prisoner’s Dilemma. In biology, the classic example is an arms race between predator and prey (Vermeij, 1987), or a parasite and its host. Defensive adaptations neutralize the predator’s advantage to which the predator then adapts, thereby neutralizing the defensive adaptation, and so on. Similar arms races also appear in social life, including the use of

	$C_2$	$D_2$
$C_1$	3, 5	0, 3
$D_1$	5, 0	1, 1

Fig. 3. The Red Queen Dilemma. In this example, Row's payoffs resemble Prisoner's Dilemma while Column's payoffs are like the Stag Hunt game. The arrows indicate best replies and converge at the unique Nash equilibrium at  $(D, D)$ .

attack ads in a political campaign, the use of steroids in professional sports, the escalating volume of conversation in a crowded restaurant, not to mention actual arms races. Everyone ends up running faster and faster just to stay in place, like Alice on the Queen's treadmill, a problem that has come to be termed a "Red Queen competition" (Van Valen, 1973; Barnett & Hansen, 1996). Mutual defection is the unique equilibrium in the Red Queen dilemma, just as it is in Prisoner's Dilemma, even though the predator prefers mutual cooperation over mutual defection (given  $R_2 > P_2$ ), and even though everyone else prefers to cooperate if everyone else will (given  $R_1 > T_1$ ). Column knows that Row will defect no matter what Column does (since  $T_2 > R_2$  if Column plays  $C$  and  $>P_2 > S_2$  if Column plays  $D$ ). So Column is guaranteed to defect even though defection is not a dominant strategy, making  $(D, D)$  a unique Pareto-deficient Nash equilibrium. Hence, this game has the same guaranteed – and tragic – outcome as Prisoner's Dilemma when played by rational egoists. Yet Red Queen does not pose a social dilemma according to the orthodox definition.

To sum up, individual rationality can lead to collectively irrational outcomes through altruism (Deadlock), egoism (Prisoner's Dilemma, Red Queen, Stag Hunt, and Chicken), or both (Coordination Game). In three of the six games (Stag Hunt, Chicken, and Coordination) a collectively irrational outcome can also occur through miscoordination of players' beliefs about their partner's intentions. But only one of these games – the Prisoner's Dilemma – poses a social dilemma according to the standard definition.

## MANY PLAYERS, MANY PLAYS, BUT NO SOCIAL DILEMMA

Although Fig. 1 simplifies these six games by limiting them to just two players, each with just two choices, the logic scales up to larger groups and

to a continuum of more or less cooperative behaviors. The  $n$ -player, continuous-choice versions are often framed as the problem of contribution to collective action. For example, the  $n$ -player Chicken Game is known as the “Volunteer’s Dilemma” (Weesie, 1994). An example is the failure of neighbors to call the police, allowing Kitty Genovese to be brutally murdered (Latané & Rodin, 1969). If everyone assumes someone else will call the police, no one will.

The games in Fig. 1 can also be extended from one-shot interactions to on-going relationships. One of the most widely observed versions of the Prisoner’s Dilemma arises in an on-going social interaction in which actors care about how their choices may affect what others do going forward – what Axelrod (1984) calls “the shadow of the future.” In game theory, on-going interactions are modeled as “supergames” composed of indefinitely repeated play of a “stage game.” Even if each individual is better off to defect in the stage game, that may not be true in the supergame if the probability that the supergame will continue is large enough and players care sufficiently about future payoffs.<sup>6</sup> The “shadow of the future” can then enforce an equilibrium in which every player cooperates. No player will deviate from the equilibrium strategy if doing so will trigger other players to revert to the Pareto-deficient dominant strategy of the stage game. When everyone cooperates, each earns a payoff that is higher than what they can guarantee themselves in the stage game by defecting. This violates Dawes’ requirement that “each individual receives a higher payoff for a socially defecting choice.” Thus, the prevailing definition limits social dilemmas not only to Prisoner’s Dilemma situations but also to just those Prisoner’s Dilemmas that are played in the here and now, as if there were no future.

The exclusion of on-going interactions has limited the horizons of social dilemma research and isolated it from the broader field of exchange theory, as noted in an important paper by Yamagishi and Cook (1993). Although “any system of generalized exchange involves the incentive structure characteristic of a social dilemma,” they point out that “neither students of generalized exchange nor social dilemmas have paid much attention to the links between those research traditions” (1993, p. 236). This unfortunate lacuna can be attributed to the prevailing theory of a social dilemma as a situation in which “if all group members cooperate, all gain,” but “for each individual, it is more beneficial not to cooperate” (Yamagishi & Cook, 1993, p. 236). With on-going exchanges, it can be more beneficial to follow a strategy of conditional cooperation. Hence, the requirement that “it is more beneficial not to cooperate” applies only to exchange that takes place without regard to the effects on future exchange. Yet systems of generalized

exchange do not dissolve after a single exchange. They involve on-going interactions, in which exchanges are repeated and expected to continue indefinitely. The “shadow of the future” in an on-going system of generalized exchange violates the requirement that defection must be the dominant strategy for all players. It is thus understandable that studies of generalized exchange have remained isolated from research on social dilemmas: Generalized exchange does not pose a social dilemma, according to the prevailing definition.

### **SOCIAL DILEMMA EXPERIMENTS DO NOT POSE A SOCIAL DILEMMA**

A review of the experimental literature on social dilemmas indicates that researchers are not generally aware of the narrowness of the standard definition. In the typical social dilemma experiment, two or more participants contribute an “endowment” to a “common pool” that is then multiplied by a factor that is larger than the total number of participants and divided equally by group members. Hence, rational egoists prefer to “free ride,” but when everyone does this, everyone loses. Many experimental designs also involve repeated play, in which participants are not told when the game will end. The failure to tell participants when the game will end is not an oversight. It is done deliberately, in order to create an on-going interaction.

The problem is that an on-going interaction does not pose a social dilemma, according to the definition of social dilemma used in these studies. For example, Komorita and Lapworth use the specification proposed by Dawes (1980): “A social dilemma is a situation in which each person has a dominating strategy and in which the choice of the dominating strategy results in a deficient equilibrium” (1982, p. 693). In their experiment, Komorita and Lapworth (1982, pp. 699–700) had participants play what they claimed was a social dilemma game. Yet clearly it was not, by their own criterion. The game consisted of “45 trials, but none knew when the experiment would be terminated.” Because the game was ongoing with no definite end-point, the situation was not one “in which each person has a dominating strategy and in which the choice of the dominating strategy results in a deficient equilibrium.”

This is not an isolated example. Parks, Henager, and Scamahorn (Parks, Henager, & Scamahorn, 1996, p. 135) also assume, following Dawes, that a social dilemma arises when defection is a strategy “that is

clearly dominant, in that performing it will guarantee a maximum personal outcome, regardless of what others do.” They then have participants play a Prisoner’s Dilemma with repeated trials, but “subjects did not know the exact number” (1996, p. 138). The on-going interaction, in turn, means that defection is not a dominant strategy. Here again, their experimental design does not pose a social dilemma, according to the definition used in their study.

This discrepancy is common place. Fleishman (1988) specifies a social dilemma as “a situation in which, as in the Prisoner’s Dilemma, it is always more profitable for a person, regardless of what others do, not to cooperate (i.e., free-riding is the dominant strategy)” (1988, p. 163). Then comes the experimental design in which free-riding is not the dominant strategy: “Subjects were not told the exact number of decision trials in the series” (1988, p. 168).

In yet another example, Sell and Son cite Dawes in noting that one of “the defining properties of a social dilemma” is that “the objective payoff to individuals for defecting (that is, taking from the common pool resource or not contributing to the public good) is greater than the payoff for contributing” (1997, p. 119). In their experiments, however, “the endpoint of the entire series of decisions was unknown” (1997, p. 128). Their social dilemma experiment did not create the conditions that the author’s definition implied were necessary to create a social dilemma.

Our point is not that these experiments contain a design flaw. Rather the opposite, we believe that the experiments are properly designed and that the studies are all important contributions to our understanding of social dilemmas. The problem is the authors’ assumption that a social dilemma requires a deficient dominant strategy equilibrium.

Not only are social dilemmas, as narrowly defined, difficult to find in the laboratory, they are even harder to find in the natural world. Situations that seem to fit the orthodox definition include neighbors who are all planning to move out of the neighborhood, a couple planning a divorce, a chance encounter on the highway, or an on-line purchase. Each of these cases could be modeled as a one-shot Prisoner’s Dilemma either because the players are strangers who do not expect to meet again (as on a highway or on-line) or because the players are terminating their relationship (by moving away or getting divorced). However, most social interactions are embedded in on-going relationships, in which the parties need to be concerned for how their strategies might affect future play. In these situations, the indefinitely repeated Prisoner’s Dilemma does not pose a social dilemma as conventionally defined.

## LIEBRAND BROADENS THE CONCEPT

The narrowness of the standard definition has led some researchers to broaden the definition, beginning with Liebrand (1983). “Dawes’ requirement of a dominating strategy for each person,” Liebrand concluded, “does not appear to be crucial for considering a situation a social dilemma” (1983, p. 124). He expanded the set of conditions to include on-going interactions as well as one-time coordination problems. According to Liebrand, “a social dilemma is defined as a situation in which (1) there is a strategy that yields the person the best payoff in at least one configuration of strategy choices and that has a negative impact on the interests of the other persons involved and (2) the choice of that particular strategy by all persons results in a deficient outcome” (1983, p. 124). This specification of the necessary conditions eliminates the requirement that the deficient outcome is a dominant-strategy Nash equilibrium, which extends the research domain to include on-going social interactions as well as one-time coordination problems. In particular, Liebrand’s specification applies to repeated play of the Prisoner’s Dilemma, as well as two additional social dilemmas – Chicken ( $T_i > R_i > S_i > P_i$ ) and Stag Hunt ( $R_i > T_i > P_i > S_i$ ), as well as their  $n$ -player variants. In the one-shot version of these games, defection “yields the person the best payoff” when everyone else cooperates (in Chicken, paying  $T_i$ ) and when everyone else defects (in Stag Hunt, paying  $P_i$ ), which “has a negative impact on the interests of the other persons” (who receive  $S_i$  instead of  $P_i$  in Chicken and who receive  $P_i$  instead of  $T_i$  in Stag Hunt). When everyone defects, this “results in a deficient outcome” (namely  $P_i$ ). Liebrand’s theory also includes these games, as well as Prisoner’s Dilemma, when they are played repeatedly for an indefinite period.

Liebrand also avoids another limitation. Tullock and Dawes limit choices to a single continuum of cooperation and defection. In contrast, Liebrand allows for any number of qualitatively distinct strategies. For example, the players could cooperate, defect, or exit (Stanley, Ashlock, & Tesfatsion, 1994).

Liebrand’s specification broadens Dawes’ but also introduces a new restriction of its own. His definition is limited to symmetric social dilemmas. “The choice of that particular strategy by all persons” refers to games in which the same strategy is available to everyone, a condition that rules out one of the most important examples of a social dilemma, the trust game. In the trust game, a trustor chooses to withhold or place trust and, if placed, a trustee chooses to honor or abuse trust. A rational trustee will abuse trust.

Anticipating this, a rational trustor will withhold trust. Yet both strictly prefer honored trust to this Pareto-deficient dominant strategy equilibrium. Further, even in those games where all players can play all other players' strategies, Liebrand's specification includes only those games in which the deficient outcome consists of an identical choice "by all persons," which excludes situations that are problematic because of failures to act in concert (e.g., failing to pull at the same time).

## THE DEFICIENT EQUILIBRIUM

All of the specifications considered so far were proposed by psychologists (Dawes, Liebrand) or economists (Tullock). Sociologists have favored a different approach. This approach broadens the concept to include many problematic situations in which miscoordination and not egoism is the problem. In a survey of the field for the *Annual Review of Sociology*, Kollock (1998) noted that "all social dilemmas are marked by at least one deficient equilibrium. It is deficient in that there is at least one other outcome in which everyone is better off" (1998, p. 184). Similarly, Raub and Buskens (2004) characterize a social dilemma as a situation "in which individual rationality (equilibrium behavior) and collective rationality (Pareto optimality) fall apart" (p. 10, translated from the original German<sup>7</sup>). Likewise, Weesie points to "situations with Pareto-inferior solutions" (1994, p. 559).<sup>8</sup>

This "deficient equilibrium" definition is becoming the standard in sociology (e.g., Simpson, 2004). It broadens the concept of social dilemma to include the on-going Prisoner's Dilemma game, Red Queen, Stag Hunt, Chicken, Deadlock, and the  $n$ -player variants of each.<sup>9</sup>

Although much broader than Dawes', the deficient equilibrium definition also excludes situations in which a Pareto deficient outcome can obtain among rational players, even though it is not an equilibrium. All equilibrium strategies are individually rational, and any deviation from the equilibrium is not rational. However, not all rational strategies are equilibrium strategies. For example, it is rational to defect in Chicken, but if both players do so, the outcome is not an equilibrium since mutual defection allocates a payoff to each player that is worse than the payoff for unilateral cooperation. In game theory, a strategy is considered rational if it can be rationally justified, and there is a precise test for determining if this is the case. A strategy is *rationalizable* (Bernheim, 1984; Pearce, 1984) if it survives

a process of iterated deletion of strategies that are not a best response. First, strategies are deleted if they are not a best response to some combination of strategies of the other players that have not yet been deleted (whether or not these strategies are themselves a best response). Then strategies are deleted that are not a best response to any of the remaining strategies, until no further strategies can be deleted. The remaining strategies are rationalizable, meaning that they can be rationally justified, even though they may be based on incorrect beliefs.

For example, in the Red Queen game (Fig. 3), a start can be made with Column deleting any of Column's strategies that are not a best response to any of Row's strategies. Looking at the arrows in Fig. 3, it is seen that  $C$  is Column's best response to Row playing  $C$  (since  $R_c = 5 > T_c = 3$ ) and  $D$  is Column's best response to Row playing  $D$  (since  $P_c = 1 > S_c = 0$ ). Thus, one either  $C$  or  $D$  cannot yet be deleted for Column. Turning to Row, it is seen that  $D$  is Row's best response to Column playing  $C$  (since  $T_r = 5 > R_r = 3$ ) and also to Column playing  $D$  (since  $P_r = 1 > S_r = 0$ ). Hence, strategy  $C$  from Row's strategy set can be deleted, since it is not a best response to either of Column's strategies. Turning back to Column, it is observed that, having deleted  $C$  from Row's strategy set,  $C$  is no longer Column's best response to any of Row's remaining strategies. Hence, it is now also possible to delete strategy  $C$  from Column's strategy set. Each player is now left with only one strategy,  $D$ , which is then the best response to the remaining strategy of the partner. Therefore, no further strategies can be deleted, and  $D$  remains as the only rationalizable strategy for either player, and  $(D, D)$  as the only rationalizable strategy profile of the Red Queen Dilemma.

In this example, it is not possible for either player to make a mistake if they are rational, because there is only one rationalizable strategy. However, that is not always the case. Fig. 4 shows that games also exist that have no

	$C_2$	$D_2$	$E_2$
$C_1$	0, 7	2, 4	7, 0
$D_1$	4, 2	5, 5	4, 2
$E_1$	7, 0	2, 4	0, 7

Fig. 4. An Efficient Outcome Is Not Guaranteed. This  $2 \times 3$  game contains four rationalizable Pareto-deficient pure-strategy outcomes (shaded), but the unique Nash equilibrium (dark) is Pareto efficient.

Pareto-deficient Nash equilibria but which nevertheless cannot guarantee a Pareto-efficient outcome even when this outcome is the unique Nash equilibrium. In the normal form  $2 \times 3$  game depicted in Fig. 4,  $(D_1, D_2)$  is the unique Nash equilibrium, which is Pareto efficient. Any other response to  $D$  than  $D$  itself leads to an outcome that is Pareto dominated by  $(D_1, D_2)$ . Yet all three strategies for each player ( $C$ ,  $D$ , and  $E$ ), and thus all nine strategy profiles in Fig. 4, are rationalizable. More generally, individual rationality can lead to a collectively irrational outcome even if the deficient outcome is not a Nash equilibrium, so long as the outcome is obtainable through rational play, that is, the outcome corresponds to a profile of rationalizable strategies. If a deficient outcome is rationalizable, then a collectively rational outcome cannot be guaranteed through the exercise of individual rationality.

There are important theoretical implications of the inability to guarantee an efficient outcome even when the game does not contain a deficient equilibrium. The problem is one of coordination. To see this, consider first what happens when miscoordination is precluded by having the players move sequentially. If all moves are sequential, players know the subgame<sup>10</sup> selected by prior movers, and miscoordination on a deficient outcome is no longer rational. Sequential games are represented in extensive form, in which rational play is further constrained to strategies that are *subgame rationalizable* (Bernheim, 1984, p. 1022), analogous to Selten's subgame-perfect Nash equilibrium (Selten, 1965).<sup>11</sup> In more intuitive language, the players must take into account not only whether a strategy is a best response, but also what will be the best response to that strategy by the player who is moving next. Rationalizable strategies are best replies to some strategy profile that other rational players might have chosen, whether or not they actually did. For example, in sequential-play Stag Hunt, the only rationalizable strategy for Row (who moves first) is to cooperate, because Column is then better off cooperating as well (and better off defecting if Row were to defect, which would be worse for Row). In sequential-play Chicken, the only rationalizable strategy for Row is to defect, because Column is then better off cooperating (and better off defecting if Row were to cooperate). The only rationalizable strategy profile is therefore Row defecting and Column cooperating. In short, in sequential play, miscoordination on a deficient outcome is possible between rational players only if the outcome is an equilibrium.

This is not the case, however, if play is simultaneous. When the players have no way of knowing which strategy their partners will play, they become vulnerable to miscoordination on a Pareto-deficient outcome, even though

this outcome is not an equilibrium. In simultaneous games with the possibility for miscoordination, collective irrationality can result from mistaken beliefs. Players can stumble into a deficient outcome that is not an equilibrium through miscoordination of conjectures about one another's strategies. Each player selects a strategy that is a best response to what the other players *might* have selected, not what they *actually* played. In other words, individual rationality allows for mistakes, given uncertainty about the intentions of other players. A player who guesses wrong is nevertheless rational so long as the guess was about a strategic decision that could itself be rationally justified. For example, in sequential-play Stag Hunt, unilateral defection is not an equilibrium and is not subgame rationalizable. However, unilateral defection is rationalizable if play is simultaneous, even though this is not an equilibrium. That is because defection is the best response to anticipated defection, but the anticipated strategy might not be the one that is actually played. Similarly, mutual defection is not an equilibrium in sequential-play Chicken and is not possible if both players are individually rational. However, in simultaneous-play Chicken, every outcome is rationalizable, including mutual defection (which obtains if both players mistakenly believe that the partner will cooperate).

In simultaneous play, if rationalizable strategies converge on a deficient non-equilibrium outcome, all players may regret their decision. That is not the case when rationalizable strategies converge on a deficient equilibrium. A deficient Nash equilibrium is without remorse, in that no player would have any regret about his/her strategy, given the strategies of others, even if there might exist another outcome with a higher payoff for every player. As Kollock notes, a deficient equilibrium is tragic in Whitehead's sense, in trapping the players in "the remorseless working of things" (Whitehead, quoted in Kollock, p. 184). In contrast, a miscoordinated outcome can leave the players with remorse for having made the wrong guess. This, then, is the pivotal theoretical question posed by the deficient equilibrium definition of a social dilemma: Is it necessary that the players have no regret when a possible deficient outcome occurs?

To answer that question, it is useful to consider the game of Nuclear Chicken, which is simply a Chicken game with an extremely low value for  $P_i$  (e.g., the end of life on earth as we know it). There is no dominant strategy equilibrium in this game, and therefore it does not pose a social dilemma according to Dawes' definition. Mutual destruction is clearly Pareto deficient but it is not a Nash equilibrium because each player would be better off "swerving" (i.e., "blinking" or "backing down" from the brink) even if the other side did not.

Nevertheless, Chicken is recognized as problematic by the “deficient equilibrium” criterion. The deficient Nash equilibrium in this game is a conjunction of mixed strategies. A mixed strategy is a probability distribution over each of the pure strategies of a game. For example, in Fig. 1, suppose player 1 cooperates in a Chicken game with probability  $(P_2 - S_2) / (P_2 - S_2 + R_2 - T_2)$ . Player 2 then earns  $(P_2 R_2 - T_2 S_2) / (P_2 - S_2 + R_2 - T_2)$  for cooperating with any probability, from zero (pure defection) to one (pure cooperation). Similarly, if player 2 plays  $C_1$  with probability  $(P_1 - S_1) / (P_1 - S_1 + R_1 - T_1)$ , then player 1 expects to earn  $(P_1 R_1 - T_1 S_1) / (P_1 - S_1 + R_1 - T_1)$ . When both play this particular mixed strategy, neither player can earn a higher expected payoff (nor a lower payoff) by unilaterally choosing another strategy (pure or mixed). The strategies therefore constitute a Nash equilibrium. Moreover, the mixed strategy is Pareto deficient because  $T_i > S_i > (P_i R_i - T_i S_i) / (P_i - S_i + R_i - T_i)$  for both players.

It is also a social dilemma according to the “deficient rationalizable outcome” criterion, because every Nash equilibrium is also rationalizable. However, Chicken also has another rationalizable outcome that is deficient *but is not an equilibrium*. In Nuclear Chicken, this outcome is mutual destruction. It is then asked which of these Pareto-deficient rationalizable outcomes – the mixed strategy Nash equilibrium or the non-equilibrium of mutual destruction – better captures our intuition about what makes this situation highly problematic for the continuation of the human race.

This question is more important than might be immediately apparent. The deeper question is whether the problem of social order is fundamentally about *incentives* that lead actors to deliberately engage in high-risk anti-social behaviors, or whether the problem is about *beliefs* that lead actors, even with the best of intentions, to make tragic mistakes when the beliefs turn out to be mistaken.

More precisely, attributing the problematic of Nuclear Chicken to a deficient mixed-strategy Nash equilibrium (MSNE) leads to the following diagnosis of the causes of catastrophe. When playing the mixed strategy, both players know there is some positive probability that both sides will defect. Both players chose to play a mixed strategy with their eyes wide open, both knew mutual destruction could happen, and both accepted this risk. If they end up blowing up the world, there was no miscalculation, it was just bad luck.

Mutual destruction in the Nuclear Chicken game might also come about through the rational play of pure strategies, but the reasoning process is very different from playing a mixed-strategy equilibrium. The Chicken game does not have a dominant strategy. The best response for each player depends on

the partner's strategy. If both sides believe that the partner will blink, then the best response for each player is to stay the course and refuse to swerve. However, neither side can know for sure what the other will do. Thus, mistaken beliefs about the partner's strategy are possible. Such mistakes do not mean the players are irrational. On the contrary, even if both players are expert game theorists, miscalculation remains possible. That is because the game has two pure-strategy equilibria in which one side blinks and the other does not. The problem is that the game does not provide the players with any way to coordinate on which of these equilibria to play. Thus, both sides may mistakenly believe the other will blink, leading to mutual destruction. And should nuclear catastrophe obtain in this way, the players would deeply regret the miscalculation that ended this planet's brief experiment with individually rational life-forms.

Such a miscalculation more closely corresponds to our intuition about what it is that makes Nuclear Chicken problematic. The possibility for miscalculation in Nuclear Chicken illustrates why a theory about the conditions that pose social dilemmas should include games where individual rationality leads players to deficient outcomes that are out of equilibrium, as well as games where the equilibrium is itself deficient. Individual rationality cannot guarantee a collectively rational outcome when individuals must make decisions whose consequences depend on the decisions of others, and they must make these decisions without always knowing what other people are going to do. In those situations, things can go badly wrong.

## THE COLLECTIVE RATIONALITY OF TAKING TURNS

Although an MSNE carries a positive probability of mutual annihilation in Nuclear Chicken, there is another way of implementing probabilistic strategies that can greatly expand the possibilities for Pareto improvements. It is easy to see which outcome in Chicken is Pareto deficient, but what is the best alternative that both players can hope for? The answer appears obvious – mutual cooperation. However, in games like Chicken in which the best reply is to do something different, it is possible for both players to do better than mutual cooperation by playing probabilistically – not the mixed strategy equilibrium but what Aumann (1974) has termed “a correlated strategy.” An everyday example of a correlated strategy in repeated games is a rule to “take turns,” or in a single encounter, a rule to flip a coin to “see who goes first.” Another familiar example is a traffic light at a busy intersection. Computer

scientists and traffic engineers have also used correlated strategies to address the problem of “selfish routing” on crowded networks and highways (Roughgarden & Tardos, 2002; Helbing, Schoenhof, Stark, & Holyst, 2006).<sup>12</sup> More generally, a correlated strategy is a probability distribution over all possible pure-strategy profiles (e.g., the cells in a payoff matrix) that prescribes to each player what pure strategy to play.

To illustrate, consider a symmetrical Chicken game between Row and Column, with payoffs (9, 5, 3, 0) for ( $T_i$ ,  $R_i$ ,  $S_i$ ,  $P_i$ ), respectively. Mutual cooperation has a payoff of 5. However, there is a correlated strategy that assigns equal probabilities to ( $D$ ,  $C$ ) and ( $C$ ,  $D$ ), with an expected payoff to each player of  $(9+3)/2 = 6$ , which strictly dominates mutual cooperation. The coin toss that decides who cooperates has given both players a way to occasionally defect without any risk of a collision. In contrast, the mixed-strategy equilibrium has some positive probability of ( $D$ ,  $D$ ).<sup>13</sup> More generally, an optimal mixed-strategy profile can never be more efficient than the optimal correlated strategy. That is because a mixed strategy is a probability distribution over each of the pure strategies of a game (e.g.,  $C$  and  $D$ ), while a correlated strategy is a probability distribution over the set of pure-strategy profiles – ( $C$ ,  $C$ ), ( $C$ ,  $D$ ), ( $D$ ,  $C$ ), and ( $D$ ,  $D$ ). Imagine a traffic light that uses a coin toss to decide when to turn green. Suppose instead each driver used a coin toss to decide when to go. The latter would eventually result in a collision (as well as wasted gas while both drivers wait for the other). The traffic light’s correlated strategy avoids both these problems by randomizing over profiles rather than strategies.

Correlated strategies expand the set of possible outcomes beyond the set of pure-strategy profiles (i.e., the cells of the payoff matrix). Pure-strategy profiles are a subset of correlated strategies – the special cases where one outcome has probability 1 and all others are zero. Including the interior of the probability distribution extends the set of counterfactual alternatives, the customary practice in deciding if an outcome is Pareto deficient is to compare the outcome to the other pure-strategy profiles, which arbitrarily excludes all correlated strategy profiles except those at the corners of the distribution. This exclusion systematically underestimates the price of individual rationality.

This is demonstrated by a version of the Deadlock game in which  $T_i + S_i > 2P_i$ . For example, consider a Deadlock game with payoffs  $T_i = 500$ ,  $P_i = 2$ ,  $R_i = 1$ , and  $S_i = 0$ . Assuming both players are self-interested, the unique dominant-strategy Nash equilibrium is ( $D$ ,  $D$ ), yielding both players payoff  $P_i = 2$ . There is no other pure-strategy profile that would benefit one player without hurting the other, and this has led most commentators to

conclude that the Nash equilibrium in Deadlock is Pareto efficient. Yet clearly both players could have done better than to settle for  $(D, D)$ , by choosing unilateral cooperation where the cooperator is chosen under a “veil of ignorance” by the toss of a fair coin. This correlated strategy ensures each player an expected payoff of  $(500+0)/2 = 250$ , which dominates the Nash equilibrium at  $(D, D)$ , whose payoff is  $P_i = 2$ . Clearly, the players have paid dearly for the privilege of exercising unbridled individual rationality in this game. Limiting collectively rational outcomes to the set of pure-strategy profiles obscures the potential tension between individual and collective rationality in the game of Deadlock, even when the game is played by egoists.

Suppose  $P$  and  $R$  are swapped, making the Deadlock game into a Prisoner’s Dilemma, and the players are made into altruists. The dominant strategy Nash equilibrium is now mutual cooperation, which has led most analysts to conclude that altruists can escape the dilemma. However, mutual cooperation is Pareto dominated by a correlated strategy that pays 250 instead of 2. The altruists do a shade better than egoists (who earn 1) but far below what they could have earned had they surrendered their unfettered autonomy and allowed their fate to be decided by the flip of a coin. The players are trapped not by self-interest but by the inability to coordinate two independent coin tosses to see who will unilaterally defect. They need the Leviathan to toss a single coin that instructs both players. In short, expanding the horizons of collective rationality to include correlated strategies leads to the remarkable result that not only egoists but altruists as well can be trapped in a Prisoner’s Dilemma by the exercise of individual rationality. Conversely, restricting Pareto improvement to pure-strategy profiles leads to underestimation of the tension between individual and collective rationality, among altruists as well as egoists.

## **CONCLUSION: SOCIAL DILEMMAS AND THE PROBLEM OF ORDER**

In the course of this discussion, alternative specifications of the conditions that are necessary for a social dilemma have been compared. Dawes requires an equilibrium of best replies to *every* partner strategy, while sociologists have increasingly settled on an equilibrium of best replies to the *actual* partner strategies. However, the games that were examined suggest that a social dilemma requires only a rationalizable outcome of best replies to

*possible* partner strategies. This definition implies a broader understanding of individual rationality that includes cases where the best reply is based on a mistaken belief about the strategies of others.

The concept of collective rationality has also been broadened. Social dilemma research has restricted collectively rational outcomes to the set of pure-strategy profiles – almost always mutual cooperation. Yet it is seen how players can sometimes do better than mutual cooperation if there are enforceable institutional arrangements for tossing a coin to decide who gets to defect. By opening up the solution set to include correlated strategies, the possibility that altruists can be trapped in a Pareto deficient outcome in a Prisoner's Dilemma if the  $T$  and  $S$  payoffs are sufficiently large is allowed.

Individual rationality based on rationalizable best replies has something in common with collective rationality based on Pareto-efficient correlated strategies – the problem of coordination. Rationalizable strategies broaden the set of deficient outcomes to include those that obtain through miscalculation and incorrect beliefs about the strategies of others, as in the game of Chicken. Correlated strategies broaden the set of efficient outcomes to include those that require coordinated moves, such as flipping a coin or waiting for the light to turn green before going.

These extensions, in turn, have important implications for the Hobbesian problem of order with which we began: Why does individual rationality sometimes lead to collectively irrational outcomes? Is the problem that people are egoists and therefore sometimes get slapped by what Hardin aptly calls “the back of the invisible hand”? Or is the problem that people are autonomous yet interdependent, and therefore sometimes fail to get all their ducks lined up in the most efficient way? If egoism is assumed away individual rationality will still lead to collectively irrational outcomes, as when there are too many cooks in the kitchen. But if instead, autonomy is assumed away, the Leviathan can simply constrain individuals (whether altruist or egoist) to implement a collectively rational outcome. This simple thought-experiment reveals what is most fundamental about the problem of social order. It is not that human rationality is directed toward individual self-interest; it is that rationality is directed toward the interdependent interests of autonomous individuals.

## NOTES

1. Games can also be depicted in *extensive form*, as a game tree, usually used to indicate that players choose sequentially, with the branches of the tree charting all possible paths from the first move to the payoffs at the end.

2. We index the choices and payoffs for each player to remind readers that a game like Prisoner's Dilemma, with the four payoffs ( $T$ ,  $R$ ,  $P$ , and  $S$ ), is actually a special case, one that has become so familiar that social psychologists often assume that the payoffs must be the same for both players. No such requirement exists; indeed, the "Sucker's" payoff ( $S_1$ ) for one player in a Prisoner's Dilemma can be larger than the partner's "Temptation" payoff ( $T_2$ ). Greater awareness of this will hopefully open up new explorations of what happens when players have very different payoffs. (The other two payoffs abbreviate the "Reward" for mutual cooperation and "Punishment" for mutual defection.)

3. A strategy profile is a set of strategies, one for each of the players. For example, in the Prisoner's Dilemma ( $C$ ,  $C$ ) is a strategy profile in which both players cooperate.

4. From the point of view of a given player, a strategy is a "best response" if no alternative strategy has a higher payoff, given the strategies of the other players. It is not necessary that alternative strategies would earn a lower payoff.

5. Heckathorn (1996) identifies four of these games in his map of the game-space of collective action, although he refers to Deadlock as the "Altruists Dilemma."

6. If the players are highly myopic, they may play each round of an indefinitely repeated game as if there were no future, in which case, they play the supergame as if it were simply a series of unrelated one-shot games. However, laboratory experiments using games that are repeated for a fixed number of rounds and then terminated have shown that the opposite is more likely the case (Selten & Stoecker, 1986). These experiments show that most people play a finite series of one-shot games as if they were a supergame. In a Prisoner's Dilemma supergame, foreknowledge of the endgame eliminates any incentive to use conditionally cooperative strategies that might engineer a tacit collusion. Nevertheless, laboratory experiments show that most people ignore the implications of the endgame up until the final few rounds, such that they play as if it were a supergame even when it is not.

7. "In anderer Terminologie: das elementare Vertrauensspiel ist ein soziales Dilemma, in dem individuelle Rationalität (Gleichgewichtsverhalten) und kollektive Rationalität (Pareto-Optimalität) auseinanderfallen" (Rapoport, 1974).

8. In game theory, the Nash equilibrium is the solution concept, hence Weesie's definition is equivalent to that of Kollock and Raub and Buskens.

9. In a Nash equilibrium, neither player has an incentive to unilaterally change strategy. Hence, mutual defection is not an equilibrium in Chicken since each player would prefer to swerve than suffer a head-on collision. The deficient equilibrium in Chicken is a mixed-strategy equilibrium that is Pareto dominated by mutual as well as unilateral cooperation, as we explain below. Deadlock also has a deficient equilibrium if the players are altruists, such that mutual cooperation is dominated by mutual defection.

10. A subgame consists of all possible moves from the original game following from the last move that was played.

11. For a slightly different treatment of rationalizability in extensive-form games, see Pearce (1984, p. 1044).

12. For example, suppose there are two roads to the Cape, one shorter than the other. If everyone takes the shorter road, the traffic jam leaves everyone wishing they had taken the longer road. But then, if next weekend they all take the longer road,

the traffic jam leaves everyone wishing someone had directed traffic to an optimal level on each road.

13. A familiar example is a traffic light at a busy intersection. Suppose the light gives each street the green signal half the time, while the other is red. If neither driver has any incentive to run the light, the outcome is a correlated equilibrium. Suppose instead each driver flips a coin to decide whether to drive through the intersection. If neither driver has any incentive to deviate from the instructions of the coin, the outcome is a mixed-strategy equilibrium. The latter carries a positive probability of occasional collisions (as well as wasted gas while both sides idle), which is why we prefer traffic lights over coin tosses to regulate traffic. Note that a traffic light could also use a coin to decide whose turn is next, instead of letting each street have green exactly half the time. Both devices generate a correlated equilibrium, but the equal-time device dominates the probabilistic version by always giving drivers enough time to clear the intersection. However, the probabilistic light still dominates the probabilistic mixed-strategy equilibrium.

## ACKNOWLEDGMENTS

The order of the authors is random, based on comparable contributions. We would like to thank Vincent Buskens and Andreas Flache for useful comments and suggestions. The second author wishes to thank the National Science Foundation (SBR 0241657 and SES-0432917) for support during the time that this research was conducted. Send correspondence to Michael Macy, 372 Uris Hall, Cornell University, Ithaca, NY 14853 (mwm14@cornell.edu).

## REFERENCES

- Aumann, R. (1974). Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1, 67–96.
- Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.
- Barnett, W. P., & Hansen, M. T. (1996). The red queen in organizational evolution. *Strategic Management Journal*, 17, 139–157.
- Bernheim, B. D. (1984). Rationalizable strategic behavior. *Econometrica*, 52, 1007–1028.
- Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology*, 31, 169–193.
- Dawes, R. M. (1991). Social dilemmas, economic self-interest, and evolutionary theory. In: D. R. Brown & J. E. K. Smith (Eds), *Frontiers of mathematical psychology: Essays in honor of Clyde Coombs* (pp. 53–79). New York: Springer-Verlag.
- Dawkins, R. (1976). *The selfish gene*. Oxford University Press.
- Fleishman, J. A. (1988). The effects of decision framing and others' behavior on cooperation in a social dilemma. *Journal of Conflict Resolution*, 32, 162–180.
- Hardin, R. (1982). *Collective action*. Baltimore, MD: Johns Hopkins University Press.

- Heckathorn, D. D. (1991). Extensions of the prisoner's dilemma paradigm: The altruist's dilemma and group solidarity. *Sociological Theory*, 9, 34–52.
- Heckathorn, D. D. (1996). The dynamics and dilemmas of collective action. *American Sociological Review*, 61, 250–277.
- Helbing, D., Schoenhof, M., Stark, H.-U., & Holyst, J. (2006). How individuals learn to take turns: Emergence of alternating cooperation in a congestion game and the prisoner's dilemma. *Advances in Complex Systems*, 8, 87–116.
- Kollock, P. (1998). Social dilemmas: The anatomy of cooperation. *Annual Review of Sociology*, 24, 183–214.
- Komorita, S. S., & Lapworth, C. W. (1982). Alternative choices in social dilemmas. *Journal of Conflict Resolution*, 26, 692–708.
- Komorita, S. S., & Parks, C. D. (1994). *Social dilemmas*. Dubuque, IA: Brown & Benchmark.
- Latané, B., & Rodin, J. (1969). A lady in distress: Inhibiting effects of friends and strangers on bystander intervention. *Journal of Experimental Social Psychology*, 5, 189–202.
- Liebrand, W. B. G. (1983). A classification of social dilemma games. *Simulation & Games*, 14, 123–138.
- Maynard, S. J. (1982). *Evolution and the theory of games*. UK: Cambridge University Press.
- Olson, M. (1965). *The logic of collective action*. Cambridge, MA: Harvard University Press.
- Ostrom, E., & Walker, J. (2003). *Trust and reciprocity: Interdisciplinary lessons for experimental research*. New York: Russell Sage.
- Parks, C. D., Henager, R. F., & Scamahorn, S. D. (1996). Trust and reactions to messages of intent in social dilemmas. *Journal of Conflict Resolution*, 40, 134–151.
- Pearce, D. (1984). Rationalizable strategic behavior and the problem of perfection. *Econometrica*, 52, 1029–1050.
- Rapoport, A. (1974). Introduction. In: A. Rapoport (Ed.), *Game theory as a theory of conflict resolution* (pp. 1–14). Dordrecht: Reidel.
- Raub, W., & Buskens, V. (2004). Game-theoretic models and empirical applications in sociology. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 44, 560–598.
- Roughgarden, T., & Tardos, É. (2002). How bad is selfish routing? *Journal of the ACM*, 49, 236–259.
- Schroeder, D. A. (1995). *Social dilemmas. Perspectives on individuals and groups*. Westport, CN: Praeger.
- Sell, J., & Son, Y. (1997). Comparing public goods with common pool resources: Three experiments. *Social Psychology Quarterly*, 60, 118–137.
- Selten, R. (1965). Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit. *Zeitschrift für die Gesamte Staatswissenschaft*, 12, 301–324.
- Selten, R., & Stoecker, R. (1986). End behaviour in sequences of finite prisoners dilemma supergames: A learning theory approach. *Journal of Economic Behaviour and Organisation*, 7, 47–70.
- Simpson, B. (2004). Social values, subjective transformations, and cooperation in social dilemmas. *Social Psychology Quarterly*, 67, 385–395.
- Smith, A. (2008 [1789]). *An inquiry into the nature and causes of the wealth of nations*. Oxford: Oxford World's Classics.
- Smithson, M. J., & Foddy, M. (1999). Theories and strategies for studying social dilemmas. In: M. Foddy, M. Smithson, S. Schneider & M. Hogg (Eds), *Resolving social dilemmas: Dynamic, structural, and intergroup aspects* (pp. 1–14). Philadelphia, PA: Psychology Press.

- Stanley, E. A., Ashlock, D., & Tesfatsion, L. (1994). Iterated prisoner's dilemma with choice and refusal of partners. In: C. G. Langton (Ed.), *Artificial life III* (pp. 131–175). Reading, MA: Addison-Wesley.
- Tullock, G. (1974). *The social dilemma: The economics of war and revolution*. Blacksburg, VA: University Publications.
- Van Valen, L. (1973). A new evolutionary law. *Evolutionary Theory*, 1, 1–30.
- Vermeij, G. J. (1987). *Evolution and escalation: An ecological history of life*. Princeton, NJ: Princeton University Press.
- Weesie, J. (1994). Incomplete information and timing in the volunteer's dilemma. A comparison of four models. *Journal of Conflict Resolution*, 38, 557–585.
- Yamagishi, T., & Cook, K. S. (1993). Generalized exchange and social dilemmas. *Social Psychology Quarterly*, 56, 235–248.