Symmetry and the Illusion of Control as Bases for Cooperative Behavior[1]

Jeffrey Goldberg
Jeffrey@goldmark.org
http://www.goldmark.org/jeff/

Lívia Markóczy
Anderson Graduate School of Management
University of California, Riverside
Phone: +1 909 787-3908, Fax: +1 909 787-3970
Livia.Markoczy@ucr.edu
http://www.goldmark.org/livia/

Lawrence Zahn
Anderson Graduate School of Management
University of California, Riverside
Phone: +1 909 787-3727, Fax: +1 909 787-3970
Lawrence.Zahn@ucr.edu

Version 2.1
January 13, 2004

[1] You are looking at version 2.1, last modified February 7, 2004. This paper has been accepted for publication in Rationality & Society. Some version of this can be found at `http://www.goldmark.org/jeff/papers/symmetry/symmetry.pdf`

Abstract

The existence of cooperation in the face of temptation to free ride requires explanation. We discuss two psychological phenomena, “symmetry” and “the illusion of control,” which we believe underlie the “what if everyone acted that way” type of reasoning used in some types of cooperation. We provide a simple model of how these lead to cooperation. We also show how some bizarre causal beliefs, such as effect preceding cause, can follow from these phenomena. We look at some existing evidence for these phenomena and report on our studies which support the model.

> Alice laughed. "There's no use trying," she said:
> "one *can't* believe impossible things."
>
> "I daresay you haven't much practice," said the Queen.
> "When I was your age, I always did it for half-an-hour a day.
> Why, sometimes I've believed as many as six impossible things
> before breakfast."
>
> —Lewis Carroll (Through the Looking Glass)

Despite the individual cost of cooperating and the obvious individual benefits of free riding in a Prisoner's Dilemma (PD) "the analytically uncomfortable (though humanly gratifying) fact" (Hirshleifer, 1985, p. 55) is that more people choose to cooperate than existing theories can convincingly psychologically explain. The functioning of organizations, as well as society, depends on unrewarded cooperative behavior of this sort (Murnighan, Kim, and Metzger, 1983, p. 515).

We discuss two psychological constructs, "symmetry" and "the illusion of control" which we believe underlie some – certainly not all – anonymous or near anonymous cooperative behavior in PDs. Symmetry is the mundane belief (not necessarily conscious) that others are like yourself with respect to certain matters. Individuals have an illusion of control when they overvalue their influence on events correlated with their choices. Symmetry's interaction with the illusion of control leads to the belief that your own state of mind and behavior (even in private) is intimately linked to that of others in that, if you choose to think or act one way, others who are like you will make the same choice. Definitions of symmetry and the illusion of control will be made more precise as the paper progresses, and can only be fully defined in terms of the model in Section 3.

Let us illustrate the psychological phenomena we wish to explore before going any further. Imagine that you are asked to act as a reviewer for a conference. You decide to review papers because you believe that reviewing is important to your field. You mention this to a friend but the following dialogue ensues between you, a cooperator (C), and your friend a, defector (D):

D: Why are you reviewing? I'm sure that lots of people are reviewing for the conference and one person isn't going to make a difference.

C: True, one person won't make a difference, but lots of "one persons" will.

D: Yes, but you are only *one* “one person”. Let the others do it.

C: But suppose everybody thinks that way, then nobody would review.

D: But they don’t all think like that. And if they did you would be left reviewing alone, so it still doesn’t make sense to review.

C: Well if I decide not to review, other people who think like me may also decide not to.

D: That’s crazy! Most of them have probably already made their decision and nobody is watching to see what *you* do.

C: Well it might be crazy, but that is the way that I feel, and I am glad that many people feel the same way.

Many of us probably have had a “what if everyone thought like that” argument with ourselves (or feelings of that nature) before we decided whether to vote in large elections, give blood, sort our garbage into recycling bins, or donate money to a charity. In this paper we wish to show that the reasoning behind this – while apparently irrational – is fully natural, human, and only mildly irrational.

In the next section we discuss families of “solutions” to the Prisoner’s Dilemma so that we can indicate how our proposal relates to them. In Section 2 we give an informal introduction to symmetry and the illusion of control, the notions we are proposing, and why we think that they are a real part of human decision making. In doing so we will also look at some previous work on (variants of) these notions. In some ways the view presented in that section can be read as an elaboration of Elster (1989, pp. 195–201). In Section 3 we provide a more explicit version of the psychological model and how it leads to cooperation. That section also restates (and slightly simplifies) a formulation made by Brams (1975) and rediscovered by Lewis (1985) which shows the relation between the PD and Newcomb’s problem. (Those unfamiliar with Newcomb’s problem may consult Figure 4 for an example). We then briefly report on some of our own studies, which were designed to test the effects of symmetry and the illusion of control on cooperation in a prisoner’s dilemma. Finally, in the concluding remarks, we discuss some broader and more speculative issues and acknowledge some of the weaknesses of our argument here.

1 When is a Prisoner's Dilemma not a Prisoner's Dilemma?

We are interested in understanding cooperation in those situations where the individual would be better off not cooperating (no matter what the others did), but everyone would be better off if everyone cooperated than they would be if many did not cooperate. This particular kind of conflict between individual interest and the collective interest is often referred to as the "collective action problem". The paradigmatic instance of the collective action problem is the familiar Prisoner's Dilemma (PD).

The fact of the matter is that when people are in what appear to be PD situations they don't defect as often as we would expect of rational individuals. This behavior requires explanation. In a pure PD, defecting is the only individually rational option (e.g., Binmore, 1994). As a consequence, all "solutions" to the PD must effectively transform the pure PD into something resembling the pure PD, but different enough to have a cooperative solution. In looking at how the PD gets modified, it is important to recall that the payoffs are actually the player's perception of the utilities. Anything that modifies the payoffs themselves will probably modify the perception of them, but sometimes only the perception needs to change. Approaches to transforming the PD into games with cooperative solutions can be divided into three categories:

1. Allow for decisions to cooperate or defect to have consequences beyond the game itself. If a reputation for defecting will leave you excluded from future games or get you beaten up, then it is worthwhile to avoid such a reputation. A pure one-shot PD requires complete and perfect anonymity and assurance of zero consequences beyond the stated payoffs; otherwise, maintaining a reputation may take priority. One of the simplest and most discussed instances of this approach is "reciprocal altruism" (Trivers, 1971) or "tit-for-tat" (Axelrod, 1984).[1] Gauthier's (1986) "constrained maximizers" may also fall into this category without too much stretching of the notion. Discussion of "social exchange" (e.g., Blau, 1964) also falls within this category.

[1] To see why tit-for-tat is too simple and doesn't quite do all that is commonly believed of it, see Hirshleifer and Martínez Coll (1988).

2. Emotions and similar devices which may change the payoffs, so that the problem is no longer a PD. If guilt feelings would make Alice feel so bad for defecting, then it effectively subtracts from the gain she would get for defecting, and so she would not actually be confronted with a true PD. Feelings like guilt, empathy and obligation certainly exist and play a role in many situations. There has been a growing interest in the role of emotions in these sorts of problems and issues (e.g., Elster, 1996). Some people (e.g., Frank, 1988) argue that one of the reasons that emotions exist is to help us manage our reputations unconsciously, automatically and convincingly. Elster (1989, ch. 3) disagrees with Frank, but for our purposes it does not matter who is right, since both argue that emotion does play a role.

3. Cognitive quirks of reasoning may transform the PD into something in which a cooperative solution is viable. Symmetry will convert a PD into something known as "Newcomb's Problem" (Brams, 1975; Lewis, 1985; Nozick, 1985). The illusion of control is the irrationality that leads to a cooperative solution to Newcomb's problem. By attempting to model non-normative reasoning, this paper is in the tradition of work by people like Tversky, Kahneman and Shafir in various papers. Elster (1989, pp. 195–201) discusses notions of "everyday Kantianism" and specifically the "magical thinking" that underlies it.

The distinction between categories 1 and 2 is not simple. On the one hand, many scholars look at emotions (category 2) as a mechanism for deriving the behavior expected of solutions fo category 1 (e.g., Hirshleifer, 1987). On the other hand, category 1 could be seen as a subcategory of 2 which simply brings modifications of payoffs outside of the formal game. We introduce our typology not because we think it reflects some essential classifications of solutions, but because it provides a convenient, if rough, description of research approaches. Along these lines, it should be noted that some of the most careful work on "social norms" does not fit neatly into our three way distinction by falling into categories 1 and 2 simultaneously. Elster (1989) provides an extensive definition of norms, which he summarizes as "the propensity to feel shame and to anticipate sanctions by others at the thought of behaving in a certain, forbidden way" (p. 105). This way, a norm fits partially under 1 (anticipating sanctions) and partially under 2 (feeling shame).

In this paper we focus on something that is in the third category, not because we think it is the most important but because it is the most often overlooked by those examining the collective action problem. There is extensive work elaborating the first of the above, and a growing body of work on the second, but there is virtually no literature on the third as solutions to the PD. Cognitive psychologists certainly have noted these things, but they did not address the collective action problem specifically and their work has not yet been integrated into the collective action literature. One of our goals in Section 2 is to illuminate two discoveries by cognitive scientists and show how their interaction explains the "what if everybody thought that way" reasoning.

The reader may have noticed that we have omitted notions like socialization from the list of solutions to the PD, and some skeptics might argue that there is no need to discuss reputation, emotions, rationality, or reasoning processes when a simple and obvious explanation "that we are socialized to behave this way" would suffice. There are several reasons why we do not accept this view when investigating symmetry and the illusion of control. Our goal is to convince you that these exists and play a role in cooperative behavior. How they arise is less important for our purposes than their existence and nature, although we do offer some speculation along the way.

2 Symmetry and the Illusion of Control

In this section we will make a first pass at bringing the notions of symmetry and the illusion of control into clearer focus. We believe that each of these has been demonstrated to exist by research in cognitive science, but have often not been disentangled from each other. In this section we wish to show that each exists independently of the other, and that the interaction of the two leads to "what if everybody thought that way" reasoning.

Recall that we have characterized symmetry as a belief that how you behave in a PD-like situation is how someone else will behave. This belief interacts with another belief, the "illusion of control", which we will discuss in more detail shortly, to lead to a compound belief that how you decide will directly or indirectly influence how others decide. Alice may have very good reasons to believe that Bob will decide to vote the same way she does, but for that belief to help motivate Alice to vote she must also believe that her decision will directly or indirectly *influence* Bob's decision. In our discussion we need to separate these two beliefs (i.e., correlation versus influence) and we need to demonstrate that there is evidence for the existence of each of them.

2.1 Symmetry

Symmetry, in its most general form, is uncontroversial, indisputable and typically rational. People predict how others would behave by imagining or remembering their own behavior. This is a fundamental principle of folk psychology or "theory of mind"[2] and one instance of it is the phenomenon of "the false consensus effect" which is "the tendency for people's own habits, values, and behavioral responses to bias their estimates of the commonness of the habits, values, actions of the general population" (Gilovich, 1990, p. 623). See Gilovich (1990) for references to "a flurry of empirical studies" which demonstrate false consensus.

But does symmetry play a role in PD-like situations? Dawes, McTavish, and Shaklee (1977) have found clear evidence that it does. They showed that people who cooperate in one-shot PD experiments predicted a higher level of cooperation from others than those who defected. They took steps to rule out the possibility that those who tend to cooperate are more optimistic about others' cooperation than defectors. Dawes et al. (1977, p. 10) explained their findings in terms of individuals using their own behavior to diagnose others'.

> Individuals may decide to use their own behavior as information about what other people would do; after all if people from similar cultures tend to behave in similar ways in similar situations, and if I do this, it follows that my peers may do so also.

Symmetry, however, is not enough to explain the choice to cooperate. It may explain a prediction that the other will cooperate as well, but to get from "my actions are correlated with Bob's actions" to "I will cooperate so that Bob will cooperate" requires the illusion of control.

[2] Theory of mind is the label given to an individual's beliefs about the mental life of others. The term was first used when asking whether chimpanzees have a theory of mind (Premack and Woodruff, 1978). Any theory of mind will have to include the notion that others' minds cannot be too different from one's own, which is a consequence we need for symmetry. Our view of symmetry neither depends on nor informs any of the debate (Carruthers, 1996) surrounding theories of theory of mind.

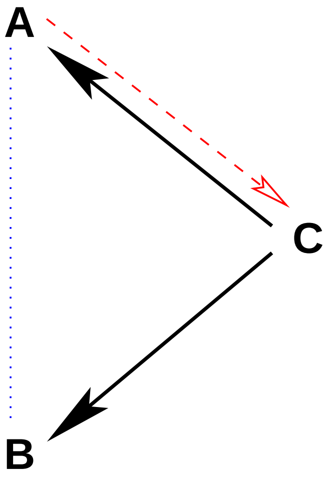
2.2 The illusion of control

We borrow the phrase "illusion of control" from Langer (1975) who demonstrated in a series of experiments that people overvalue their influence over chance events when they are given some choice. However, we will be using it in a broader sense to include influence over non-chance events. Our intended usage is similar to what has been called a "sense of control" (Hayashi, Ostrom, Walker, and Yamagishi, 1999; Watabe, Terai, Hayashi, and Yamagishi, 1996), an "illusion of influence" (Nozick, 1985, p. 127), "quasi-magical thinking" (Shafir and Tversky, 1992), confusion between "diagnostic" and "causal" relations (Quattrone and Tversky, 1984), and "magical thinking" (Elster, 1989, chap. 5).

A grossly simplified version of Calvinist doctrine, to use an example from Quattrone and Tversky (1984) and Elster (1989), exhibits the illusion of control to a very high degree: (1) those who are among the "elect" are elected by the deity before birth; (2) the elect will live with God in heaven; (3) those who are elected will live virtuously; (4) therefore, if one wants to be among the elect (and live with God), one should decide to live virtuously. Another instance might be the patient whose minor symptoms of something potentially serious disappear when they are being examined by a doctor. The confusion between diagnostic correlation and causal relationship is ever present in the human mind. In each of these cases, people are manipulating the consequence of something in order to affect its cause or some other consequence of a common cause.

Consider the diagram in Figure 1. A and B are caused by C as indicated by the solid arrows. In the case of the Calvinist, we can have A be "virtuous living", B be "going to heaven", and C be "being among the elect". According to the theology A and B are caused by C with no other causal relations involved. Yet a Calvinist will make efforts to live virtuously, presumably in the belief that A can sometimes cause C. Recall that beliefs are not necessarily conscious or accessible beliefs. This belief in A influencing C is indicated by the dashed arrow in the figure.

Consider another case. Imagine that there is some inherited gene that has two effects. It decreases one's chance of heart disease and it increases one's tolerance for cold. In Figure 1 we can assign "having the gene" to C, "tolerance for cold" to A, and "non-tendency for heart disease" to B. People aware of the existence of such a trait might, if they fell victim to the illusion of control, try to display a higher tolerance for cold than those



C is the cause of both A and B as indicated by the solid arrows. A has some illusory control over C as indicated by the dashed arrow. The correlation between A and B is indicated by the dotted line.

Figure 1: Causal and non-causal relations

unaware of such a gene. Quattrone and Tversky (1984) performed exactly such an experiment, telling people about such a gene, and found exactly the increased tolerance for cold that the illusion of control would predict.

2.3 Interaction between symmetry and the illusion of control

Symmetry and the illusion of control, when combined, provide the basis for a "what if everyone thought that way" argument. This interaction is illustrated in an experiment by Quattrone and Tversky (1984) on the "voter's illusion". They investigated why people may vote in situations where they know that their single vote is very unlikely to matter in and of itself:

> [I]f one votes, then one's politically like-minded peers, who think and act like oneself, will also vote. Conversely, if one abstains, then one's like-minded peers will also abstain. Because the preferred candidates could defeat the opposition only if the like-

> minded citizens vote in larger number than do the unlike-minded citizens, the individual may conclude that he or she had better vote. That is, an individual may regard his or her single vote as diagnostic of millions of votes, and hence as a sign that the preferred candidates will emerge victorious. [p. 244]

That is an expression of both symmetry and the illusion of control. Symmetry is recognizing that the decision to vote or abstain will be diagnostic of the actions of like-minded people. The illusion of control is the belief that voting will make it more likely that others will vote. That is

> an individual may regard his or her own decisions as diagnostic of the decisions likely to be made by other "like-minded" persons. If the individual recognizes that beneficial outcomes would ensue if very many like-minded persons select a particular alternative, even if the choice is costly, not witnessed by other and not likely by itself to affect the final outcome. In these circumstances, the choice is made to "induce" others who think and act like oneself . . . [p 244]

Again looking at Figure 1, A can be "me voting", B can be "others like me voting" and C can be "whatever properties of socialization, norms and mindset lead me and others like me to vote". While we do think that symmetry and the illusion of control may well help explain some voting behavior, it is only one of many psychological mechanisms which may (e.g., Jankowski, 2002; Schuessler, 2000).

The illusion of control combined with symmetry can lead to some bizarre causal beliefs. The illusion of control, when applied to decision making and the decisions of others, entails a (usually unconscious) belief in backwards causation. But it can be even worse than that: it can also entail a belief in causation backwards in time. Imagine that Alice and Bob are both thinking about attending a public rally for some cause they both support. For various logistical reasons Bob must make his decision earlier than Alice makes hers (maybe he has further to travel), but doesn't communicate his decision to Alice (maybe they don't know each other). Returning to Figure 1, we can let A be Alice's decision to attend, B be Bob's, and C can be whatever common cause accounts for the correlation between their decisions. But in this case, B precedes A in time. So A is exerting an influence which has an effect prior to the time of A.

2.4 Controlling the illusion

Many people have failed to notice that the illusion of control is operative in their reasoning. Even when it is explicitly pointed out, some people do not acknowledge it. It appears to be not only unconscious, but very inaccessible to introspection. Dawes et al. (1977), when discussing something very much like symmetry in PDs, fail to note that their argument depends on the illusion of control. More strikingly, philosophers and logicians studying the problem have also fallen victim to the illusion. One of the most spectacular and appealing attempts to "solve" the PD (Davis, 1985a,b; Hofstadter, 1983; Rapoport, 1966) tries to provide a rational basis for the illusion of control by arguing that if both players are rational they *must* come to the same decision. The fallacy of the argument (Binmore, 1993, 1994; Brams, 1975; Gibbard and Harper, 1985) would not be particularly difficult to see, but is obscured by the illusion of control. Our own discussions with people have illustrated some of the difficulties: "When I help someone who has collapsed on the street I do so for rational and selfish reasons." "What are those reasons?" "Well, I would want someone to help me if I collapsed on the street."

Elster (1989, p. 195) notes that "Psychologically, if not logically, there is a short step from the thought 'If *I* don't do it, why should anyone?' to the thought 'If I don't do it, nobody will'." He also notes that some people don't recognize that the extra step is necessary.

In order to see that the step is necessary, consider again Figure 1. Imagine that the solid arrows from C to A and B involve complete determination. This way A and B will be perfectly correlated. You cannot choose a particular A because you want a particular B unless there is some arrow leading away from A. The illusion of control provides that extra arrow. Alice may believe that her voting is perfectly correlated with Bob's, but she could only use that as part of her motivation to vote if there in an arrow from her decision.

The illusion of control may be, partially, a consequence of a more fundamental belief in free will. Effectively there are three ways A's decisions and B's decisions can be correlated: (1) A's decisions can cause B's; or (2) B's decisions can cause A's; or (3) both A's decisions and B's decisions are caused by the same thing, C. But given A's belief in her own free will, she can't entertain options 2 or 3. Her free decision cannot be controlled. Only option 1 remains to her (although A does not have to cause B directly, it can be by some intermediate C). Go back to Figure 1 and let A be A's decision, B be B's decision, and C be the common cause of the decisions.[3]

In Section 3 we will omit the intermediate cause C, and effectively have the dashed arrow go from A to B. It is irrelevant whether the influence is from A directly to B, or from A to B via C, and it is simpler to discuss in terms of it being direct.

2.5 The necessity of symmetry and the illusion of control

Our view is just one of many stories that can be told to explain cooperative behavior in collective action problems. So is there actually a need to explain what is already explained? Shafir and Tversky (1992) conducted an experiment in which cooperation that could be explained by existing views is factored out. Although they didn't set out to test symmetry and the illusion of control, their work provides one of the clearest demonstrations so far that they are not only psychologically available, but that they play a real role in cooperation in one-shot anonymous PDs.

In some experimental groups subjects were told ahead of time what the other player had decided. When subjects were informed that the other player had defected they also defected all but 3% of the time. When they were informed that the other player had cooperated they defected in all but 16% of the circumstances. When they were not told beforehand what the other player did, they cooperated 37% of the time. The rate of cooperation more than doubled when the subjects did not know the other's decisions, but knew that the other had already decided.

If people were rewarding known cooperation 16% of the time, we would expect that when it is unknown whether the other has cooperated, the cooperation rate would be no more than 16%. That is, if people were simply rewarding or punishing cooperation or defection (or are behaving like Gau-

[3] If our belief in free will can be taken for granted, then this may go towards satisfying the task set by Nozick (1985, p. 127): "in Newcomb's example there is the *illusion* of influence. The task is to explain in a sufficiently forceful way what gives rise to this illusion so that, even as we experience it, we will not be deceived by it." Irrespective of whether our decisions are caused by external events or internal and even random ones, as long as they are caused by something, we do not have free will in the most strict sense. However we do maintain a good enough approximation of free will along with a psychological experience of it (Dennett, 1984) to force an illusion of free will which defies normal causation.

thier's (1986) constrained maximizers or following some social norm to cooperate with cooperators), we would expect that the decision under uncertainty would be to cooperate between 3 and 16 percent of the time, but certainly never more than 16%. The doubling of cooperation (from 16% to 37%) cannot be explained either by reputation management or emotion. Symmetry together with the illusion of control, however, might be an explanation, as Shafir and Tversky (1992, p. 458) suggest:

> [W]hen the opponent's response was not known, many subjects preferred to cooperate, perhaps as a way of "inducing" cooperation from the other. Because subjects naturally assume that the other player – a fellow student – will approach the game in much the same way they do, whatever they decide to do, it seems, the other is likely to do the same.

In a repeated PD, symmetry and the illusion of control, if they play a role, could lead to increased cooperation following mutual defection. If Alice and Bob both defect in one round, they may each take that as evidence that they think alike. So, they might be more likely to cooperate after mutual defection than would be explicable by other factors influencing their decisions. Zahn and Wolf (1998) appear to have found exactly that result.

2.6 Something old, something new, . . .

As should be abundantly clear from this section, there has been a great deal of work on what we are calling symmetry and the illusion of control. There has been much less on the interaction between them, and still less on the relationship between them and the kind of cooperation we are investigating. Almost all of the pieces of the argument we are making have been put forward by others, but on those rare occasions when some of the pieces have been put together, the argument has been hesitant or merely hinted at. Shafir and Tversky (1992) come very close to making our argument, but end up proposing an alternative. Quattrone and Tversky (1984) also came extremely close to putting symmetry together with the illusion of control, but because they were primarily interested in the illusion of control they didn't develop the argument further.

	B cooperates	B defects
A cooperates	(w, w)	(v, t)
A defects	(t, v)	(u, u)

Figure 2: Prisoner's Dilemma with w, u, t and v as defined in the text

Much of the discussion of the PD in the philosophy literature (of which more is cited in the next section) has also been exploring the pieces, occasionally putting them together, but never fully psychologizing the results. By translating some of the work in philosophy to a psychological model, we can construct, in the next section, a clear model of "what if everyone thought that way" reasoning.

3 The Expected Utility of Symmetry

This section sketches a model of how symmetry helps transform the PD into something in which cooperation is a viable option given the illusion of control. Figure 2 shows the familiar PD where w is the payoff for both cooperating, u is the payoff for both defecting, t is the payoff for defecting while the other cooperates, and v is the payoff for cooperating when the other player defects. The ordered pair (x, y) indicates a payoff of x to the first player and y to the second player, where $t > w > u > v$ and $w > (t + v)/2$.

Suppose that Alice suspects that there is a high chance, $p_s > .5$, that Bob will come to the same decision in a PD as she does. Following Lewis (1985), we can recast the PD into what we see in Figure 3. Here, p_s is what we will call the symmetry probability.[4]

[4] A more general way to treat the symmetry probability would be to maintain two probabilities (Lewis, 1985), one for when cooperating and another for when defecting. For simplicity, we assume that these will be the same and wrap them up in one probability. However, breaking it up into these two probabilities is necessary if one wants to generalize to n-person prisoner's dilemmas.

	B does	
	same as A	not same as A
A cooperates	(w, w)	(v, t)
A defects	(u, u)	(t, v)
	p_s	$1 - p_s$

Figure 3: PD recast with symmetry probability

In this figure the rows still represent A's payoffs for defecting and cooperating, but the columns represent whether B does the same as A (first column) or B acts differently than A (second column). Each column is also associated with a specific probability.

If the probability of B doing the same as A is 1 (i.e., if there is a perfect correlation between A's actions and B's), then we can ignore the second column of Figure 3 entirely, leaving only the first column and presenting A with a choice between w (for a cooperate–cooperate outcome) and u (for defect–defect). If the symmetry probability is .5 (no correlation between what A does and what B does) then the problem is the simple PD, in which, for the one-shot case, it makes sense for players to defect.

When viewed this way (subject to the illusion of control), the expected utility of cooperating will be seen as

$$p_s w + (1 - p_s)v \tag{1}$$

and the expected utility of defecting will be

$$p_s u + (1 - p_s)t \tag{2}$$

This will make cooperating the better choice when

$$p_s > \frac{t - v}{(w + t) - (u + v)} \tag{3}$$

It must be noted that it is only through the illusion of control that we justify calculating the expected utility this way (Brams, 1975; Gibbard and Harper, 1985).[5]

3.1 Symmetry in n-person situations

Throughout we have discussed exclusively the two-player PD, but the type of analysis we've presented here can be easily extended to other games such as Chicken and the n-player PD. We briefly sketch that here.

For this discussion we will consider a simplified variant of the public good games sometimes called "social dilemma" games (e.g., Dawes et al., 1977). In a typical one of these, a group of n players is given a stake, s, and are told that they can contribute any part of that to a pool. Anything they don't contribute, they can keep. But anything that they do contribute will be multiplied by some amount, k, such that $1 < k < n$. The results will be distributed to all the players evenly, irrespective of how much they contributed. So, if a total of c is contributed to the pool, then kc will be divided equally among the n players. However, each will also get to keep whatever portion of their own stake, w that they withheld.

The game parallels the two person game in a number of ways. The worst overall outcome is if everybody withholds their stake and the best overall outcome is if everyone contributes their stake. Each individual is better off by withholding their own stake.

We will use a simplified version for showing how this interacts with symmetry and the illusion of control. In our simplified version, individuals have the single choice of contributing their stake or withholding it, instead of having the choice of how much of their stake to contribute. As a consequence, we can just consider the stake to be 1 unit without any loss of generality.

Each player ends up with the amount they withheld (0 or 1), plus one nth of the multiplied sum of contributed stakes. If p is the portion of people who contribute their stakes, then the payoff for contributing your stake is $k(pn)/n$ or kp. The payoff for not contributing is $1 + kp$. Without symmetry and the illusion of control, players know that if they contributed p will be $1/n$ higher than if you don't, since it will contain your own contribution, and

[5] Our model does not explicitly mention the illusion of control but sneaks it in through calculating the expected utility in the way we do. For our purposes it is enough to say that the illusion of control is an all or nothing thing.

you are one of the n participants. With the condition $k < n$ it is always better to withhold a contribution.

But now take symmetry and the illusion of control into account. Alice may believe that if she cooperates other people like her in the game will also cooperate (symmetry), and uses that as motivation to cooperate (illusion of control). Thus she believes that if she cooperates the portion of people cooperating will grow by more than just herself.

We will use p_c to indicate what portion of cooperators a person expects if they themselves cooperate, and we will use p_d to indicate what portion of cooperators a person expects if they themselves defect. It will be worthwhile to cooperate when

$$kp_c \geq 1 + kp_d \tag{4}$$

which is exactly when

$$p_c - p_d \geq \frac{1}{k} \tag{5}$$

So if a person's symmetry beliefs are sufficient to satisfy Equation 5 and if the person is subject to the illusion of control, we can have these lead to cooperation in a situation where defecting is rational.

4 Our studies

We have argued that there is a large body of experimental work which supports our view on symmetry and the illusion of control, but that work was not conducted or evaluated in the light of the theory we have presented here. For example, we believe that Shafir and Tversky (1992) have already demonstrated that the illusion of control plays a role in cooperation in PDs, although they came to a different conclusion.

Recall that Shafir and Tversky (1992) found that subjects playing PDs cooperated more when their opponent's decision was still unknown (37%) than when their opponent's decision to cooperate was known (16%). Although Shafir and Tversky (1992, p. 458) acknowledge that something like symmetry and the illusion of control might play a role, they ended up concluding that something else, limitations of deciding under uncertainty, was responsible.

> The presence of uncertainty, we suggest, makes it difficult to focus sharply on any single branch [of a decision tree]; broadening the focus of attention results in a loss of acuity. The failure to appreciate the force of [dominance], therefore, is attributed to people's reluctance to consider all the outcomes, or to their reluctance to formulate a clear preference in the presence of uncertainty about the outcomes.[p. 457]

Shafir and Tversky (1992) did not investigate the differences between the uncertainty hypothesis and the illusion of control hypothesis, but we do.

If people see their choices as occurring before (in some unspecified sense of "before") the other person's choice, they should be more inclined to invoke the illusion of control. We are aware of a number of unpublished[6] studies which have attempted to do this by setting up problems where subjects were told either that the others had already made their choice, or others had yet to make their choice. We should expect that the illusion of control would play a larger role where the subject believes that they are going first. This is because, following the White Queen in the epigraph of this paper, it is harder to believe two impossible things (backwards causality and causality backwards in time) than just one impossible thing (backwards causality). However with very few exceptions those (unpublished) studies failed to show clear results. Some exceptions include Watabe et al. (1996) which is described by Hayashi et al. (1999) and Morris, Sims, and Girotto (1995).

We believe that our timing effect experiments failed for three reasons:

1. The effect that we are going after is relatively weak. Thus anything which substantially complicates the task of the subjects is likely to overwhelm the effect we are trying to isolate.

2. By having players play in different orders, the symmetry of the situation is broken. Subjects may no longer see that the other player is being faced with the same decision problem. This will also weaken the effect.

3. People who accept backwards causality are inclined to accept causality backwards in time with little extra effort. That is, once you believe the first impossible thing, believing the second is harder, but not much harder. At least that is the case for these two impossible things.

[6] It is a serious problem for the field that negative results are very difficult to publish.

In order to circumvent these problems, we sought something like a timing effect experiment which would be for the subjects nearly as simple as a regular PD, would not break the symmetry of the situation, and would allow us to isolate those people who don't take both backwards causation and causation backwards in time as a bundle.

The first two goals were achieved by changing not the timing of who makes a decision first, but in the order in which the decisions are presented. The same PD can be described by the trees in Figures 5 and 6. The presentation of the former may lead the subjects to consider the other's decision problem in the context of their own decision, while in the latter they will think of their own decisions in the context of the other's decision. The former then lends itself more easily to seeing the other player's decision as being causally dependent on the subject's. The games (from the text of the decision) are clearly simultaneous and symmetric. So this design is fully symmetric and nearly as simple from the subjects' perspective as a simple PD.

Hypothesis 1 *Those presented with a PD as a tree in which their own decision is at the root (Figure 5) will be more likely to cooperate than those presented with a tree in which the other player's decision is at the root (Figure 6).*

To deal with the third problem, we identified those who are inclined to take on both backwards causation and causation backwards in time by presenting them with a non-supernatural version of Newcomb's problem. Those who take the one box must accept both backwards causation and causation backwards in time together.

Hypothesis 2 *The effect described in Hypothesis 1 will be less pronounced among those who select one box in Newcomb's problem.*

Our final hypothesis is

Hypothesis 3 *There should be a correlation between cooperation in a prisoners dilemma and taking one box in Newcomb's problem.*

We report on two studies we conducted, one in 1998 and the other in 2001, which collectively support all of these hypothesis. We only report on a small part of the 1998 study, as most of it centered around a failed attempt to get timing results.

4.1 The 2001 study

Subjects were various undergraduate students in an introductory business course at the University of California, Riverside.

We presented subjects with two tasks. First a Newcomb's problem task as in Figure 4 and then with one of four varieties of a prisoners dilemma.

We presented subjects with text describing a PD and presented them with one of two trees describing the problem. There were two variants of the text, the second variant shown in brackets.

> A and B are two different choices or "strategies" that are described below. You are asked to pick either Strategy A or Strategy B. Your response will be randomly paired with the response of someone else who is making their decision on this problem at the same time that you are. If you both pick strategy A then you each get 75 points. If you both pick strategy B then you each get 30 points. [*first alternative*: If you pick strategy A and the other person picks strategy B, then you get 25 points and the other person gets 85 points. If you pick strategy B and the other person picks strategy A then you get 85 points and the other person gets 25 points.] [*second alternative*: If the other person picks strategy A and you pick strategy B, then you get 85 points and the other person gets 25 points. If the other person picks strategy B and you pick strategy A then you get 25 points and the other person gets 85 points.]
>
> This decision situation is indicated in the following diagram.

The two variants of the diagram are shown in Figures 5 and 6 respectively.

Both versions represent identical PDs. The only difference is the order in which the players' choices are presented. In the "you first" version the subject's choice is described first while in the "other first" version the subject's choice is presented second. In neither case is there any asymmetric information. The difference is only one of how the problem is described.

4.2 Results of the 2001 study

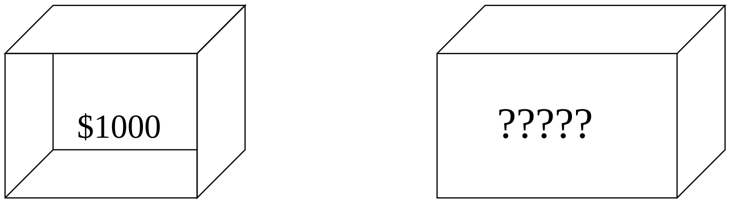
We found that those presented with the "you first" version of the trees cooperated significantly more than those presented with the "other first" versions.

In the figure are two boxes. One is open and you can see that it contains $1000; the other box is closed and you cannot see into it. The closed box contains either $100,000 or nothing. Your choice will be between

(a) taking only the closed box, *or*

(b) taking *both* boxes.

If the problem were as simple as this, it would be obvious that the best choice would be to (b) ("take both boxes"). But the problem is not that simple.



Imagine that there is a super-intelligent space alien which is an expert on human psychology and can, after a brief examination, predict individual human behavior extremely well. It has examined you some time in the past, and has either put $100,000 into the closed box or left it empty. If it thought that you would take both boxes, it left the closed one empty. If it thought that you would take only the closed box, it put the $100,000 in it.

You know that this alien psychologist is extremely accurate at this kind of prediction, and has done this with hundreds of people before you and has never (yet) made an error. But you also know that the procedure is carefully audited, and the money is already placed (or not placed) in the closed box before you are presented with this choice.

Based on this, please answer whether you take (**A**) only the closed box, or (**B**) both boxes.

Please circle the strategy you choose.

A **B**

Figure 4: Newcomb's problem as presented

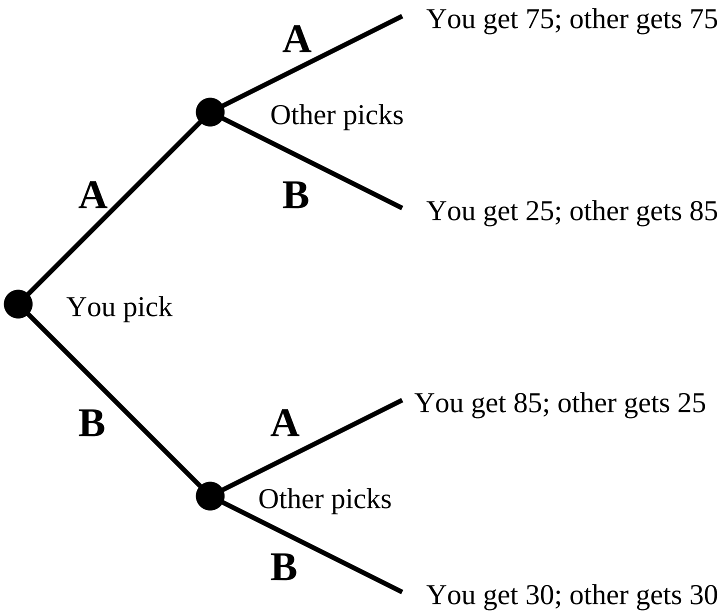

A
A
B
B
A
B
You pick
Other picks
Other picks
You get 75; other gets 75
You get 25; other gets 85
You get 85; other gets 25
You get 30; other gets 30

Figure 5: The "you first" presentation of the decision tree

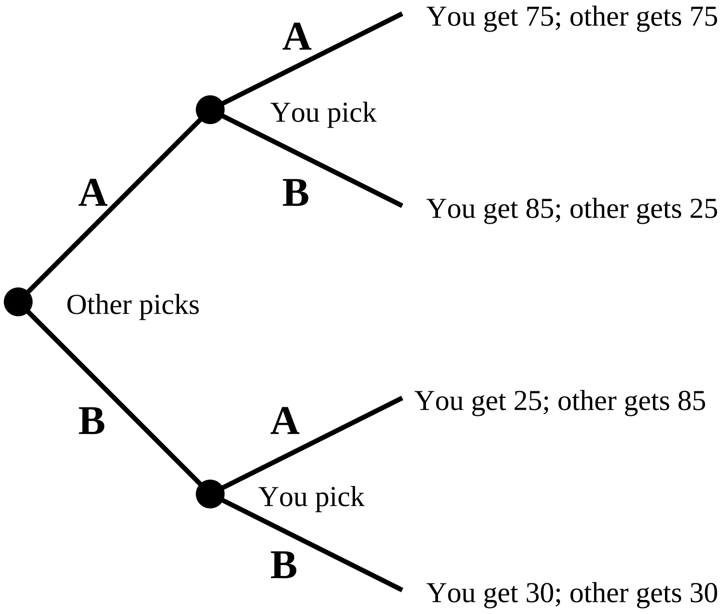

A
You get 75; other gets 75
You pick
A
B
You get 85; other gets 25
Other picks
You get 25; other gets 85
B
A
You pick
B
You get 30; other gets 30

Figure 6: The "other first" presentation of the decision tree

	Coop	**Defect**	**Total**
Subject first	58	36	94
Other first	32	39	71
Total	90	75	165

Table 1: Cooperation and PD presentation

One box

	Coop	**Defect**	**Total**
Subject first	35	23	58
Other first	12	12	24
Total	47	35	82

Two box

	Coop	**Defect**	**Total**
Subject first	23	13	36
Other first	20	27	47
Total	43	40	83

Table 2: Cooperation and timing by Newcomb's response

The data are summarized in Table 1 which yield a $\chi^2 = 4.512$ and a one-tailed $p = 0.025$ by Fisher's exact test.[7] This confirms Hypothesis 1.

To test Hypothesis 2 we looked at how strong this effect was among those who took one box in Newcomb's problem and those who took two. The data are presented in Table 2. For those who took only one box in Newcomb's problem we see no timing effect here at all ($N = 82$, $\chi^2 = .743$ *ns*) while for those who took both boxes, there is a strong effect ($N = 83$, $\chi^2 = 3.717$, one tailed $p = 0.044$). This supports Hypothesis 2.[8]

To our surprise we did not find a relation between taking one box in Newcomb's problem and cooperating in the PD as seen in Table 3. Note that this also includes those presented with a PD as a standard table instead of as a tree. Our data do not show any support for Hypothesis 3. This is

[7] We report the χ^2 for information, but calculate p using Fisher's exact test throughout.

[8] Because the effect of tree presentation is close to zero for the one-boxers, we simply look at the strength of the effect for the two-boxers as also indicating the difference between these two cases.

	Coop	**Defect**	**Total**
One box	66	57	123
Two boxes	65	67	132
Total	131	124	255

Table 3: Newcomb's and cooperation in 2001 study

	Coop	**Defect**	**Total**
One box	43	9	52
Two box	25	30	55
Total	68	39	107

Table 4: Newcomb's and PD for 1998 study

peculiar because on previous failed attempts to show timing effects, we always found a strong support for Hypothesis 3, which brings us to a previous study.

4.3 An earlier study

In an earlier study involving 107 full-time MBA students at Cranfield University in the UK, we also queried Newcomb's problem along with the PD. The wording of Newcomb's problem was similar to that presented in Figure 4 except that the values were £1000 and £100,000 for what may be in the boxes. Here we found an overwhelmingly strong relation between cooperation on the PD and taking one box in Newcomb's problem ($N = 107$, $\chi^2 = 16.0$, one tailed $p < 0.0001$). See Table 4.

5 Discussion and Conclusions

In this paper we have done three things: (1) We proposed that some cooperative behaviors that occur in collective action problems can be explained by the psychological phenomena of the illusion of control and symmetry; (2) We provided a description of these mechanisms and how they interact to produce the kind of effect we believe exists; (3) We sketched a model of how symmetry and the illusion of control can affect decision making. In doing this we have discussed and drawn together many similar notions which have been suggested by others. Once symmetry and the illusion of control are clearly articulated and disentangled from each other, they and their interactions are surprisingly simple.

Elster (1989, p. 186) says that arguments that state that people cooperate because they are following norms or are irrational are "utterly trivial, unless the specific norm of cooperation or the specific type of irrationality is defined in a way that is independently meaningful." We agree entirely. We have not only defined symmetry (which is not on its own irrational) and the illusion of control in ways that are independently meaningful of the cooperation they are to explain, but we have discussed some independent evidence for them as well. By doing so we have placed on a solid foundation a theory of one of the cognitive mechanisms for some types of cooperative behavior.

We must clearly note that the core ideas we have presented are well represented in other places. The relationship between Newcomb's problem and the prisoner's dilemma was first made clear by Brams (1975) along with an understanding of the causal illusion that underlies Newcomb's problem. Other work in both psychology and philosophy have touched upon various aspects of this, and the recent work reported by Hayashi et al. (1999) shows how, behaviorally, this can be shown to play a role.

5.1 The name of the game

We have not, in fact, provided a pair of psychological parameters for symmetry and the illusion of control that underlie the parameters that we have used in the three separate games we've worked through. However, given the illusion of control, it is not difficult to see that an underlying probabilistic parameter for symmetry could get the parameters needed in the three cases. Because there are several non-equivalent ways that could be done, and because we can see no reason to select any one over any of the others, we are content – for the time being – to apply the symmetry notion to games on a case by case basis. Once our ability to isolate and measure these is much more refined than it is now, we can start arguing over the nature of the underlying symmetry parameter (or parameters). But at this point it is premature to do so.[9]

[9] Zahn and Wolf (1998) have proposed a symmetry parameter based on one particular model of the decision process.

5.2 Origins of symmetry and the illusion of control

While symmetry and the illusion of control may very well exist, the question remains as to whether they are parts of human nature or are determined by purely social factors. There are reasons for believing that they are not simply learned as arbitrary cultural artifacts, but that they exist as human universals. If the "what if everyone thought that way" reasoning were an arbitrary cultural artifact it should exist in only a small portion of cultures, but we predict that it is present in a wide variety of distinct cultures. To say, however, that we expect something to be a human universal does not mean that we expect that there won't be individual or cultural variation in its expression (Brown, 1991; Markóczy and Goldberg, 1997; Tooby and Cosmides, 1992).

We expect to find that "what if everyone thought that way" type arguments are present in a variety of cultures. This could be tested by presenting informants from a variety of cultures with a list of possible responses to, say, a child picking a flower in a public garden and asking them to rank or rate these responses with respect to appropriateness in their culture. We would predict that "what if everyone picked the flowers for themselves" would be considered an appropriate part of the care-giver reaction in more than just a small minority of the cultures.

Additionally we have suggested in Section 2 that this everyday Kantianism is a consequence of the interaction of two psychological devices that have much broader scope (theory of mind for symmetry and free will for the illusion of control). If we are correct about this, it means that this fact about our ethical nature is neither a cultural artifact nor a direct adaptation, although it may be a consequence of other adaptations.

5.3 What's next?

It is clear that things like symmetry and the illusion of control have been discussed in a variety of different academic disciplines. We hope that recasting all of those discussions in terms of our present view will allow greater integration of the discussion. Such integration will help us pursue the greater questions including (1) to what degree does the reasoning described here play a role in actual behavior; (2) how widespread are the circumstances in which this kind of reasoning can play a role; (3) and how does this mechanism compare in importance to other motives for cooperation, such as those listed by Elster (1989) and Markóczy (forthcoming).

References

Axelrod, Robert (1984). *The evolution of cooperation.* New york: Basic Books. Cited on: 3

Binmore, Ken (1993). Bargining and morality. In *Rationality, Justice and the Social Contract: Themes from "Morals by Agreement"* (eds. David Gauthier and Robert Sugden), chapter 8, pp. 131–156. Ann Arbor: University of Michigan Press. Cited on: 10

Binmore, Ken (1994). *Game Theory and the Social Contract. Volume I: Playing Fair.* Cambridge, Mass: MIT Press. Cited on: 3, 10

Blau, Peter (1964). *Exchange and Power in Social Life.* New York: Wiley. Cited on: 3

Brams, Steven J. (1975). Newcomb's problem and the prisoners' dilemma. *Journal of Conflict Resolution*, 19(4): 596–612. Cited on: 2, 4, 10, 15, 25

Brown, Donald E. (1991). *Human Universals.* New York: McGraw Hill. Cited on: 26

Carruthers, Peter (1996). Simulation and self-knowledge: A defence of theory-theory. In *Theories of Theories of Mind* (eds. Peter Carruthers and Peter K. Smith), chapter 3, pp. 22–38. Cambridge: Cambridge University Press. Cited on: 6

Davis, Lawrence H. (1985a). Is the symmetry argument valid? In *Paradoxes of Rationality and Cooperaton: Prisoner's Dilemma and Newcomb's Problem* (eds. Richmond Campbell and Lanning Sowden), chapter 15, pp. 255–263. Vancouver: University of Vancouver Press. Cited on: 10

Davis, Lawrence H. (1985b). Prisoners, paradoxes, and rationality. In *Paradoxes of Rationality and Cooperaton: Prisoner's Dilemma and Newcomb's Problem* (eds. Richmond Campbell and Lanning Sowden), chapter 2, pp. 43–49. Vancouver: University of Vancouver Press. (Originally published in *American Philosophical Quarterly*, Volume 14, no 4, 1977). Cited on: 10

Dawes, Robyn M., Jeanne McTavish, and Harriet Shaklee (1977). Behavior, communication, assumptions about other people's behavior in a common dilemma situation. *Journal of Personality and Social Psychology*, 35(1): 1–11. Cited on: 6, 10, 15

Dennett, Daniel C. (1984). *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge, MA: MIT Press. Cited on: 11

Elster, Jon (1989). *The Cement of Society: A study of social order*. Studies in Rationality and Social Change. Cambridge: Cambridge University Press. Cited on: 2, 4, 7, 10, 25, 27

Elster, Jon (1996). Rationality and the emotions. *The Economic Journal*, 106: 1386–1397. Cited on: 4

Frank, Robert H. (1988). *Passions within Reason*. New York: Norton. Cited on: 4

Gauthier, David (1986). *Morals by Agreements*. Oxford: Clarendon Press. Cited on: 3, 11, 12

Gibbard, Allan and William L. Harper (1985). Counterfactuals and two kinds of expected utility. In *Paradoxes of Rationality and Cooperaton: Prisoner's Dilemma and Newcomb's Problem* (eds. Richmond Campbell and Lanning Sowden), chapter 7, pp. 133–158. Vancouver: University of Vancouver Press. (Originally published in Hooker, Leach, and McClennen (eds.) *Foundations and Applications of Decision Theory, Volume I*, 1978, D. Reidel, Dordrecht). Cited on: 10, 15

Gilovich, Thomas (1990). Differential construal and false consensus effect. *Journal of Personality and Social Psychology*, 59(4): 623–634. Cited on: 6

Hayashi, Nahoko, Elinor Ostrom, James Walker, and Toshio Yamagishi (1999). Reciprocity, trust, and the sense of control: A cross-societal study. *Rationality and Society*, 11(1): 27–46. Cited on: 7, 17, 25

Hirshleifer, Jack (1985). The expanding domain of economics. *American Economic Review*, 75(6): 53–70. Cited on: 1

Hirshleifer, Jack (1987). On the emotions as guarantors of threats and promises. In *The Latest on the Best* (ed. John Dupré). Cambridge, Mass: MIT Press. Cited on: 4

Hirshleifer, Jack and Juan Carlos Martínez Coll (1988). What strategies can support the evolutionary emergence of cooperation. *Journal of Conflict Resolution*, 32(2): 367–398. Cited on: 3

Hofstadter, Douglas (1983). Metamagical themas: The calculus of cooperation tested through a lottery. *Scientific American*, 248(6): 14–18. Cited on: 10

Jankowski, Richard (2002). Buying a lottery ticket to help the poor: Altruism, civic duty, and self-interest in the decision to vote. *Rationality and Society*, 14(1): 55–77. Cited on: 9

Langer, Ellen J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, 32(2): 311–328. Cited on: 7

Lewis, David (1985). Prisoners' dilemma is a Newcomb problem. In *Paradoxes of Rationality and Cooperaton: Prisoner's Dilemma and Newcomb's Problem* (eds. Richmond Campbell and Lanning Sowden), chapter 14, pp. 251–255. Vancouver: University of Vancouver Press. (Originally published in *Philosopy and Public Affairs*, Volume 8, no 3, 1979). Cited on: 2, 4, 13

Markóczy, Lívia (forthcoming). Multiple motives for cooperation. *International Journal of Human Resource Management.* Cited on: 27

Markóczy, Lívia and Jeff Goldberg (1997). The virtue of cooperation: A review of Ridley's *Origins of Virtue. Managerial and Decision Economics*, 18: 399–411. Cited on: 26

Morris, Michael W., Damien L. H. Sims, and Vittorio Girotto (1995). Sources of cooperation in the one-shot prisoner's dilemma: Distinguishing between causal illusion and ethical obligation. *Journal of Experimental Social Psychology*, 34(5): 494–512. Cited on: 17

Murnighan, J. Keith, Jae Wook Kim, and Richard Metzger (1983). The volunteer dilemma. *Administrative Science Quarterly*, 38: 515–538. Cited on: 1

NOZICK, ROBERT (1985). Newcomb's Problem and two principles of choice. In *Paradoxes of Rationality and Cooperaton: Prisoner's Dilemma and Newcomb's Problem* (eds. Richmond Campbell and Lanning Sowden), chapter 6, pp. 107–133. Vancouver: University of Vancouver Press. (Originally published in *Essays in Honor of Carl G. Hempel*, 1969, D. Reidel, Dordrecht). Cited on: 4, 7, 11

PREMACK, DAVID and G. WOODRUFF (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1: 515–526. Cited on: 6

QUATTRONE, GEORGE A. and AMOS TVERSKY (1984). Causal versus diagnostic contingencies: On self-deception and on the voter's illusion. *Journal of Personality and Social Psychology*, 46(2): 237–248. Cited on: 7, 8, 12

RAPOPORT, ANATOL (1966). *Two-person game theory: The essential ideas.* Ann Arbor: The University of Michigan Press. Cited on: 10

SCHUESSLER, ALEXANDER A. (2000). Expressive voting. *Rationality and Society*, 12(1): 87–119. Cited on: 9

SHAFIR, ELDAR and AMOS TVERSKY (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology*, 24: 449–474. Cited on: 7, 11, 12, 16, 17

TOOBY, JOHN and LEDA COSMIDES (1992). The psychological foundations of culture. In *The adapted mind* (eds. Jerome H. Barkow, Leda Cosmides, and John Tooby), chapter 1, pp. 19–136. Oxford: Oxford University Press. Cited on: 26

TRIVERS, ROBERT L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, 46: 35–57. Cited on: 3

WATABE, M., S. TERAI, N. HAYASHI, and T. YAMAGISHI (1996). Cooperation in one-shot prisoner's dilemma based on expectations of reciprocity. *Japanese Journal of Experimental Social Psychology*, 36: 183–196. (In Japanese). Cited on: 7, 17

ZAHN, G. LAWRENCE and GERRIT WOLF (1998). Cooperation and coordination in superior–subordinate relations. *Computational & Mathematical Organization Theory*, 3(4): 249–265. Cited on: 12, 25