Geoffrey Brennan* and Geoffrey Sayre-McCord*

On 'Cooperation'

https://doi.org/10.1515/auk-2018-0005

Abstract: The term 'cooperation' is widely used in social and political and biological and economic theory. Perhaps for this reason, the term takes on a variety of meanings and it is not always clear in many settings what aspect of an interaction is being described. This paper has the modest aim of sorting through some of this variety of meanings; and exploring, against that background, when and why cooperation (in which sense) might be of value, or be required, or constitute a virtue.

Keywords: cooperation, cooperativeness, positive interdependence, symbiosis, mutual advantage, altruism, coordination

> "[…] *observe the accommodation of the most common artificer or day-labourer in a civilized and thriving country and you will perceive that the number of people of whose industry a part, though but a small part, has been employed in procuring this accommodation exceeds all computation*" (Adam Smith, WN I.i.11)

> "[…] *in civilized society, man stands at all times in need of the cooperation and assistance of great multitudes, while his whole life is scarce sufficient to gain the friendship of a few persons*" (Adam Smith, WN I.ii 2)

> "*Should I have my way, I should propose that we cease forthwith to talk about 'economics' or 'political economy' although the latter is the much superior term. Were it possible to wipe the slate clean, I should recommend that we take up a wholly different term such as 'catallactics' or 'symbiotics'. The second of these terms would on balance be preferred.*" (Buchanan 1964)

1 Introduction

Adam Smith emphasized, repeatedly, the importance of 'cooperation' in making people better off. He was not the first—nor, by any means, the last—social commentator to make this claim. Indeed, 'cooperation' is a term that is widely de-

***Corresponding author: Geoffrey Brennan,** School of Philosophy, RSSS, ANU; Philosophy Department UNC-Chapel Hill; Political Science Department, Duke University, e-mail: geoffrey.brennan@anu.edu.au

***Corresponding author: Geoffrey Sayre-McCord,** Philosophy Department UNC-Chapel Hill, e-mail: sayre-mccord@unc.edu

ployed in contemporary scholarship—not just in the social sciences (economics and political science most notably) and in the humanities (especially in moral philosophy and political theory) but also in evolutionary biology. Because the term is deployed in a wide variety of contexts and because it often carries a normative valence, we think it a worthwhile task to explore the various meanings it is assigned. In particular, we want to lay out a tri-partite distinction among:

1. 'cooperation' as the fact of positive interdependence between the payoffs accruing to agents in an interaction (PI)[1]
2. 'cooperation' as involving some element of coordination by participants in the attempt to produce benefits (C1)
3. 'cooperation' as a situation in which the prospect of such positive interdependence constitutes a direct, non-derivative, and effectual motive for participant actors to act in the way they do. (C2)

We offer this set of distinctions in the spirit of clarification. It is not our ambition here to try to legislate on the proper use of terms. However, we should register our ambivalence about the propriety of describing the mere fact of positive interdependence as 'cooperation'—for reasons that we shall try to make clear in what follows. Accordingly, although some scholars do use 'cooperation' in this way, that is a usage whose appropriateness we shall want to contest. We have therefore labelled it not as a possible variant of 'cooperation' but simply as PI.

One issue with which we shall be concerned relates to the precise location of normativity—that is, of the value of cooperation. Is it in its effects? Is it in the process (of coordination)? Or is it in the motives of those who engage in the process expecting the effects? For if it is the fact of positive interdependence (PI) that is of normative concern—as it might be for consequentialists and some types of quasi-contractarians—then it becomes an open question as to whether cooperation in either of the *other* types we isolate carries any direct normative weight at all.

In principle, PI is a property of an outcome. Whether a given outcome (or a set of outcomes under some given institutional arrangement) possesses that property constitutes one possible domain of enquiry. Whether that outcome comes about

1 In an earlier version of this paper, we referred to the fact of positive payoff interdependence as 'mutual advantage', in part because of its redolence with Rawlsian vocabulary (as when Rawls refers to society as a "*cooperative venture for mutual advantage*", Rawls 1971, 4). But the 'mutual' in 'mutual advantage' risks implying that the advantage depends on each acting with an eye to the benefit of the other, and thus risks blurring one of the distinctions we are working to warn against. Accordingly, we have jettisoned that term in favour of a set of words that is a bit clumsy but, we hope, clear enough as to content.

via a process that involves cooperation in either the C1 or C2 sense is a further and separate question. That further question might of course be independently normatively significant. The fact that individuals are 'cooperating' in some psychological sense might be independently desirable: 'cooperation' may have a value over and above the benefits (if any) to participating parties that the cooperation might produce. But whether cooperation is of independent value or not is a question that has to be addressed directly. It would clearly be a mistake to think that C1 or C2 had independent value simply because in some (perhaps many) instances, cooperation (C1 or C2) happened to secure PI.

The quotes from Smith and Buchanan, in the epigraph, illustrate what we see as a certain lack of clarity in this connection. Consider first Smith.

In the first chapter of *The Wealth of Nations* Smith emphasizes that in a commercial society the standard of living that each enjoys is heavily dependent on the activities of others. Indeed, Smith assures us, the number of persons on whom each relies "*exceeds all computation*". In that sense, each depends on the "*assistance and cooperation of many thousands*".[2]

Smith then claims, in the second chapter, that the division of labour, which is the central mechanism for this '*cooperation*', co-evolves with market exchange:

> "The division of labour [...] is the necessary though very slow and gradual consequence of a certain propensity in human nature that has in view no such extensive utility—the propensity to [...] exchange one thing for another." (WN I.ii.1)

Furthermore, Smith insists, the propensity to exchange arises mainly from each participant pursuing her separate advantage. In some cases, we may rely on benevolence and goodwill "*to gain the favour of those whose service*" is required. But it is only from friends that we can expect such treatment. And man's "*whole life is scarce sufficient to gain the friendship of a few persons*". The extensive division of labour that we observe is, then, for Smith, evidence that friendship is a secondary consideration in explaining the division of labour: the division of labour and the advantages that derive from it are primarily an upshot of individuals pursuing their own interests via market exchange.[3]

Smith believes that the structure of interdependence he observes in a "*civilized and thriving country*" generates vast general benefit. He speaks (optimisti-

2 Later (in chapter 2), Smith repeats the phrase—this time with "*assistance*" and "*cooperation*" in reverse order and "*great multitudes*" in lieu of "*many thousands*". It is that later version that is quoted in the epigraph.

3 In ii.3, for example, Smith considers the emergence in a primitive tribe of the armourer. "*From a regard for his own interest therefore, the making of bows and arrows grows to be his chief business.*"

cally) of the "*universal opulence that extends itself to the lowest ranks of the people*" (WN I.i.10). Of course, the structure of interdependence is 'social'—there would be no interdependence without multiple participants. But in what sense are the participants 'cooperating', if most of the persons they are supposed to be cooperating with, they do not know, do not care about, and whose very existence they may not even recognize?[4] When Smith observes 'cooperation' among the members of a civilized and thriving society, is he merely observing the fact of positive interdependence in the relations among those members? Is what he is describing really 'cooperation' at all? That is a question we shall want to return to in the subsequent discussion.

A different way of exploring the relation between cooperation and positive interdependence is in terms of Buchanan's canvassing of appropriate names for the discipline of economics (as quoted in *his* contribution to the epigraph). Buchanan is relevant in this connection because, as perhaps the most prominent and self-conscious social-contract theorist among academic economists, he is famous for his insistence that economists should concentrate on exchange (rather than rational choice or scarcity). Exchange is of interest to Buchanan in part because Buchanan sees exchange as the central source of general benefit.

It is also noteworthy that, when Buchanan considers alternative titles for 'economics'—'catallactics' or 'catallaxy' (the science of exchange) on the one hand; and 'symbiosis' on the other—he expresses a preference for 'symbiosis'. Any such preference presupposes that Buchanan sees a difference. We want to suggest one location of that difference.

In the standard biological applications, symbiosis refers to a situation in which the survival value of each of two species in an interactive situation is augmented by the presence (and more particularly some relevant behaviour or feature) of the other. There is, as we might put it, 'reciprocal benefit' (with 'benefit' in this setting understood in terms of increased evolutionary fitness). The biological case is an interesting reference point because, as normally understood, it operates without the mediation of any psychological factors whatsoever. Symbiosis does not involve attributing to the interdependent 'partners' any particular motives.[5] But if that is so, we think it is at least misleading to describe the symbi-

4 After all, Smith writes as if the extent of the division of labour is news to his readers—something that they are likely to be astounded by when it is drawn to their attention. Yet all such readers are participants in the processes that Smith describes.

5 Evolutionary biology is prone to use terms that have a natural psychological connotation to describe evolutionary phenomena. 'Altruism' is a notable example. In the 'cooperation' case, is there analogously some kind of unconscious projection of human features onto the situation? There seems to be a loose sense in which in symbiotic cases we can talk of species 'cooperating':

otic partners are 'cooperating' to produce the positive outcome they both enjoy. Is there really anything at all in the biological cases that involves something more than the brute fact of positive interdependence (in the survival-rate terms that are relevant in this particular application)? We think not.

Buchanan obviously believes that at least some human cases (those with which economics is concerned) are closely analogous to the biological. To be sure, benefit/advantage in the human case is typically understood in terms of individual flourishing rather than genetic survival. But the relevant point is independent of what form the 'benefit' takes. The characteristic feature of *human* interactions in the 'symbiotic' analogy would just be that the 'payoff' to each in the interaction is positively affected by the behaviour of the other. But do the features of such positive interdependence involve 'cooperation' in any further sense?

There are perhaps two possible 'further senses'. One thought is that the prospect of positive interdependence may need to be *recognized* by both players for the full benefits to be appropriated. Such recognition may explain aspects of the agent's behaviour even in cases where the positive interdependence does not constitute a direct motive for the relevant behaviour. The other thought is that 'behaving cooperatively' involves the agents involved in the interaction having the prospect of each benefitting as an effective motive for their acting as they do. In this last case, we think, talk of a *cooperative venture for mutual gain* really has its place. In what follows we shall want to explore both cases.

2 'Cooperation' as the Fact of Positive Interdependence?

At the risk of belabouring the point, it will be useful to say a little more about the PI case. And for this purpose, it will be useful to appeal to a simple 'game'—illustrated in normal form by the matrix below (*figure 1*).

In this interaction, 'individuals'[6] A and B each have two actions available to them; with the payoffs to each represented as a number pair (with the first num-

the outcome is after all 'advantageous' to both of the interacting species. But, one might ask, what is at stake in the move from describing the attributes of the outcome to ascribing 'cooperation' to the parties involved?

6 The 'individuals' are just the entities that interact. In the human case, they may be individual persons. But they may also be aggregates of various kinds. The precise nature of the interactors is not at this point germane.

ber in each case being the payoff to A; the second number the payoff to B). In the structure illustrated, both individuals have a dominant strategy—an action for each 'individual' that is best, independent of what the other chooses. Moreover, the outcome that is best for A (a2) is also best for B; and vice versa. This game illustrates the case of 'cooperation' understood in terms of nothing more than the recognition of the positive *payoff interdependence*—if *figure 1* is an instance of 'cooperation' at all, it casts cooperation as merely PI.

		B's action	
		b1	b2
A's action	a1	(1,1)	(2,2)
	a2	(2,2)	(3,3)

Fig. 1

In the human case, it is plausible to think of the individuals as making choices between actions; and as being motivated by a desire to have as high a payoff as possible. In this case, there is a Nash equilibrium [(a2, b2)] in this game; and this Nash equilibrium exhibits 'positive interdependence'. When A chooses a2 over a1, A is 'cooperating' with B only in the sense that A's action serves to increase B's payoff. And analogously, B, in choosing b2 over b1, increases A's payoff.

Several details of the formulation in *figure 1* are worth mentioning.

First, the formulation does double duty for the biological and analogous social cases—subject to two provisos:

1. that the interpretation of the 'pay-offs' in those two applications are different (survival value in the former, and material well-being or preference satisfaction, or some such in the latter).
2. That the 'actions' in the biological case represent counterfactuals (perhaps entirely hypothetical ones)—things that might (at least in principle) be other than they are—whereas in the human case, they usually represent options for choice available to the agents. So, in the human case, some account of why agents choose the actions they do is a standard feature of explanation. A typical assumption in such settings is that each human agent prefers outcomes that leave that agent better off in payoff terms and that each actor chooses

her actions on this basis. If that is the motivational structure then *figure 1* will indeed exemplify PI.[7]

The critical feature of the interaction for our purposes is not the interpretation of payoffs, but rather the general *structure* of the interaction—the 'pattern of interdependence' that it exhibits. It is clearly a feature of the interaction that, as a matter of fact, each player is led on the basis of her own independent principles of action, to act in a manner that is advantageous to the other (in whatever currency 'advantage' is understood). We think that this is the sense in which Smith uses the term. But the presence of that feature does not entail either C1 or C2: in which case, we are inclined to think that it does not merit description as 'cooperation' at all. Perhaps the term 'symbiosis' could be deployed to cover cases (whether biological or social) where the interdependence has the relevant structure. 'Cooperation' could then be reserved for cases in which PI plays some role in motivating participants' actions.

Second, and relatedly, the actions/behaviours available to the two agents—the a_i's and b_i's—are specified abstractly. This is done deliberately: illustrative examples often carry an interpretative freight that we here want to eliminate. There are many specific cases that might exemplify the structure illustrated in *figure 1* and we shall provide some examples in what follows—but at this point, we do not want to frame interpretation by appeal to any one example.

Third, it is worth noting that the interdependence exhibited in *figure 1* is in terms of payoffs—not actions. Neither player need form expectations concerning how the other is going to act in order to determine which is the payoff-maximizing action for herself. And neither player can influence the behaviour of the other by her own choice of action.[8]

Buchanan's preference for the term symbiotics suggests that *figure 1* is the kind of case he has in mind. It exhibits PI. The temptation to describe this case as 'cooperation' is understandable—it is '*co*-operation' because both agents are necessary to produce the outcome; and it is co-'*operation*' because it is the action/operation of each that is involved. But for all that the game matrix tells us, the positive effect of the action of each on the payoff of the other could come about entirely by accident. Or it could come about as the incidental upshot of circumstances that happen to be ongoing but are not systematic in any stronger sense.

---

7 Of course, each player may have positive concerns for the payoff of other actor(s) affected by the interaction: but the presence or absence of such concerns is irrelevant for securing the given outcome.

8 Clearly, in *figure 1* neither has any motive to do so: both will choose the action best for the other anyway by virtue of her own principle of action.

A simple example might be helpful here. Suppose A and B occupy adjoining lots along an English country lane. Each is a keen gardener; and each develops her garden partly for the purpose of satisfying her creative and aesthetic urges and partly because having a spectacular garden will add to the market value of her property. However, A's attractive garden also adds to the value of B's property next door and vice versa. The effect of their independent actions is to produce the most beautiful little lane in the area—though the beauty of the lane as such was never part of either's intentions. They do not coordinate in any way in producing that particular outcome. The independent action of each does however have the effect of significantly improving their property values (and for that matter contributing to the pleasure of passers by).

We might even suppose in a variant of the example that A and B are rivals in their endeavours—that despite the effect on the value of her property, A would prefer that B's garden be somewhat less attractive (and analogously in relation to B's attitudes towards A's garden). In this case, the increase in property value will overstate the benefit to each: and the fact of *positive interdependence* will be something that each regrets. Nevertheless, unless these negative effects outweigh the positive property-value effects, there will remain positive net interdependence: *Figure 1* faithfully depicts the relation between the players actions and the net benefits that each derives.

There is no C1 cooperation. There may of course be C2 cooperation. Player A may have an additional motive for choosing a2 over a1—namely, that it also gives a benefit to B. And B may be similarly motivated in choosing b2 over b1. But these extra motivational considerations are 'surplus to requirements'. C2 cooperativeness does not change the behaviour of the parties involved—nor the structure of positive interdependence that we see as the central feature.

3 PI—compared to C2?

One natural point of contrast here is with an alternative game structure—exemplified by the prisoner's dilemma (illustrated in the game matrix in *figure 2*). Again, in order to focus on the structural features, we shall abstract from district attorneys and prisoner's confessions (or any of the many alternative applications of PD thinking in social and biological contexts) and focus directly on the structure of interdependence. In this family of interactions, the interdependencies are negative. In the biological case, the maximization of A's fitness reduces B's fitness; and vice versa. In the human case, A's payoff-maximizing action reduces the payoff to B; and vice versa.

		B's action	
		b1	b2
A's action	a1	(2,2)	(0,3)
	a2	(3,0)	(1,1)

Fig. 2

This interaction does not exhibit the structure characteristic of *figure 1*. As in *figure 1*, each player has dominant strategy (a2 in A's case and b2 in B's case); and there is therefore a unique Nash equilibrium, (a2,b2). But this equilibrium outcome generates for each a lower payoff than is achievable in outcome (a1,b1): it is just that (a1,b1) is inaccessible to independent payoff-maximizing play by each.[9]

Of course, in the social setting, this second case is utterly familiar. As Heath (2006) remarks:

> "If individuals simply seek to satisfy their own preferences in a narrowly instrumental fashion, they will [sometimes][10] find themselves embroiled in collective action problems: interactions with an outcome that is worse for everyone involved than some other possible outcome. Thus they have reason to accept some form of constraint over their conduct in order to achieve this superior, but out-of-equilibrium outcome. A social institution can be defined as a set of norms that codify these constraints. Simplifying somewhat, one can then say that social institutions exist in order to secure gains in Pareto-efficiency." (Heath 2006, 313)

Heath's description of this 'utterly familiar' case invites a couple of quibbles. First, the final sentence seems to imply an explanatory agenda—as if the existence of various social institutions is to be explained in terms of the capacity of those institutions to secure gains in Pareto-efficiency. Considerably more would need to be said about the processes by which social institutions are established before that explanatory agenda could be sustained.[11] To be sure, Buchanan sometimes

9 Or, in the biological analogue, the fitness of species A would be higher if B's fitness were not maximized; and vice versa.

10 Our insertion. It is unnecessary either to Heath's project or our own to suppose that such 'collective action problems' are ubiquitous (as Hobbes might have it).

11 To observe that interacting individuals 'have reason' to accept constraints on their behaviour is, as the prisoner's dilemma itself shows, not necessarily sufficient to motivate them to accept such constraints. Moreover, the arrangement that maximizes A's payoff is one in which B is constrained to choose to 'cooperate' but A is not! Only if A's accepting a relevant constraint is *necessary* in order to get B to accept it will A increase her payoff by accepting the constraint.

talks in such terms;[12] but most social contract theorists, we take it, are engaged first and foremost in a justificatory exercise. In that sense, an alternative rendering might be more apt—something along the lines of:

> "one can then say that social institutions are justified in terms of their securing gains in Pareto efficiency[13]—or perhaps, justified to the extent that they generate such gains."

Second, when Heath refers to the Pareto-dominated outcome (a2, b2) as 'out-of-equilibrium', he is referring to the interaction in *figure 2*. But of course, that interaction is one in which the effects of the relevant 'social institution' are absent. The logic of the argument requires that the 'Pareto superior' outcome must be an equilibrium outcome in the alternative interaction where the 'social institution' or 'set of norms' is in operation. The task of designing or supporting social institutions might then be conceived in terms of transforming the interaction in *figure 2* into an interaction of the type in *figure 1*: the aim is to replace prisoner's dilemma type interactions with 'symbiotic' ones.

The relevance of 'cooperation' specifically in the analysis of prisoner's dilemma situations is complicated by a tendency on the part of economists (and game theorists generally) to label the *actions* in *figure 2* as 'cooperate' (for a1 and b1) and 'defect' (for a2 and b2). And this usage seems natural enough, given the usual set-up for PD cases. Yet the 'cooperative action' is understood simply in terms of its outcome, as that action which if undertaken by both (all) players would involve the 'cooperative'—specifically the PI—outcome. But there are complications here that ought to be recognized.

In the first place, a1 is the 'cooperative action' (so understood) for A only if B chooses b1. It may well be the case that the cost imposed on A by B's 'defecting' is so large that A may not be justified in choosing a1 (on aggregate payoff-maximizing grounds) unless A is reasonably sure that B will choose b1. Such a case is illustrated by *figure 2'*. There, the aggregate payoff when both defect is considerably larger than the aggregate payoff[14] for the off-diagonals (a1, b2) and (a2, b1); so, choosing a1 in the absence of some restriction on B's choice might not be justified on aggregate payoff grounds. In that sense, the link between aggregate payoff and the a1 action is not always so clear.

12 See for example Buchanan 1990.

13 Actually, it is not 'gains in Pareto efficiency' as such that do justificatory work but the net gains in preference satisfaction (in terms of which Pareto efficiency is defined).

14 The relation between aggregate payoff and 'mutual advantage' is a further source of complication. We say a little about that distinction in section V.

		B's action	
		b1	b2
A's action	a1	(2,2)	(-10,3)
	a2	(3,-10)	(1,1)

Fig. 2'

Equally, the 'off-diagonal' payoffs might be such that the best outcome for *aggregate* payoff maximization is where one player chooses the 'cooperate' action and the other the 'defect'; and this fact can be transformed into specifically *mutual* advantage by players taking turns in 'exploiting' the other. The structure is illustrated in *figure 2"*. Clearly, in this case, both players will be better off with an arrangement in which (a1,b2) and (a2,b1) alternate, than one in which (a1,b1) emerges all the time. In this case, a necessary condition of maximal PI is that 'defect' is chosen by one or other player in every round of play. The 'natural' link between the so-called 'cooperative action' and the action that produces a so-called cooperative outcome (that outcome that exemplifies PI) is severed.[15]

		B's action	
		b1	b2
A's action	a1	(2,2)	(0,10)
	a2	(10,0)	(1,1)

Fig. 2"

There could of course be an alternative understanding of the 'cooperative *action*'—namely, that which is chosen under a cooperative *motive*. Conceivably, for instance, one solution to the 'prisoner's dilemma' may lie in both players having a concern to produce the collectively best outcome, rather than their own individually best outcome. But in modelling that possibility, we would have to distinguish 'inclinations to choose actions' (where the concern to produce the collectively best outcome finds expression) on the one hand, from payoffs to each,

15 It should perhaps be emphasized that the variants illustrated in *figure 2'* and *figure 2"* retain the essential features of the prisoner's dilemma—namely, that there is a Nash equilibrium that is Pareto dominated in payoff terms.

on the other—the latter being relevant for the benefits each receives and the former being relevant for behaviour. To take just one example of how this might come about, individuals might engage in 'we-thinking'[16] when settling how to act. They might ask, not: what is best for me? but rather: what is best for *us*?[17] Having established the jointly best *outcome*, each might then ask: 'What do I have to do in order to play my part in producing this jointly best outcome?' and then perform that action. We take it that this kind of calculus would reflect a strong form of 'cooperation' in a direct psychological sense: each takes into account the payoff to the other in calculating how to act. But of course, what is best for each may well be that the *other* adopts a 'we-thinking' mode of calculating action: my choosing that 'cooperative' approach makes sense in payoff-maximizing terms *for me*, only if my doing so increases the likelihood of your choosing that approach (and then only if the best for us outcome is also better for me, or at least not worse). In other words, maximizing my payoff does not entail my adopting the 'we-thinking' mode of calculation. Furthermore, as we shall show in section V, there are lots of relevant cases in which the 'we' relevant for we-thinking is ill-defined. When the 'we' for defining cooperation falls short of the set of persons affected by outcomes, the 'cooperative' disposition may produce an outcome that does not reflect the PI property.[18]

In the light of these complications, consider again Heath. Heath declares his purpose to be that of exploring the '*benefits of cooperation*'—and he thinks it important to do so at a lower level of abstraction than is customary in social contract philosophy. He complains that Gauthier and Rawls (as archetypical examples) '*make no attempt at all to specify how cooperation improves the human condition*'. Now, if cooperation is understood in PI terms then the question of '*how cooperation improves the human condition*' seems bizarre: cooperation understood as PI is actually constituted by the increases in the payoffs to (all relevant) players. Heath's enterprise of exploring the 'benefits of cooperation' would then simply be that of exploring the 'benefits of general benefits'—which seems like a singularly pointless exercise! But if, on the other hand, cooperation is to be understood as the widespread presence of a motivating concern to produce benefits to others as

16 This case is advanced in Economics by Bacharach 2006, and Sugden/Gold 2007; and connects to work in Philosophy by Bratman 1992 and Gilbert 2001.

17 'What is best for us' might be interpreted aggregatively. But it might also be interpreted as constrained by the requirement that both have her payoff increased. We shall say a little about that ambiguity in section V below.

18 One way of putting the point is to suggest that the prisoner's dilemma structure of interdependence may be replicated in interactions *between* groups, even though individuals are motivated 'cooperatively' within groups.

well as to oneself, then there would be a possible agenda—though it doesn't seem to be Heath's agenda. It would have two parts: the first to explain how the given motivations work to solve various predicaments in particular cases; and second, how those motivations might be extended more broadly and more deeply across the population. What in fact Heath does is to direct attention to different '*mechanisms of social benefit*'—with the mechanisms in question being understood as "*different ways in which individuals can help each other to achieve each other's objectives, whatever those objectives may be*" (315). But before we get to 'mechanisms', it seems we need to be clear on what is at stake in individuals '*helping each other to achieve each other's objectives*'. We need to distinguish whether people 'helping each other' is a matter of fact about the outcomes their interactions produce (PI); or is meant to represent a description of their intentions (C2). Or yet something else (C1 perhaps).

4 Cooperation as Coordination—from PI to C1?

So far, we have explored the idea of cooperation through the lens of two familiar structures of interaction—in both of which players have a dominant strategy and hence a unique Nash equilibrium. This property means that each can act entirely independently of the action choices made by the other. Each may or may not be ignorant of how the other is going to act; but any such information is irrelevant to either's choice of action (though not to size of payoff). It is not even necessary that either be aware that the actions of the other have any impact on the payoff she receives: each will in fact be led to act in a manner that does or does not benefit the other—but neither C2 nor C1 is entailed.

		B's action	
		b1	b2
A's action	a1	(1,1)	(0,0)
	a2	(0,0)	(2,2)

Fig. 3

There is, however, a different family of cases in which players interact in a manner such that each's choice of action *does* depend on the choice the other makes.[19] The simplest form of such a game is the coordination game illustrated in *figure 3*.

In the game as depicted, there is an outcome that is better for both than any other—namely, (a2,b2). It is tempting to conclude that, in this case, there is a need for 'coordination' among players to achieve the best outcome: players need to 'cooperate' to secure mutual advantage. But that is not quite so. Suppose A and B act independently—with no attempt outside the game to coordinate their behaviour. A might reason that since she does not know what B will do, it is just as likely that B will choose b1 as b2; and since the expected value to A is higher by choosing a2, that is what A will rationally choose. B may reason analogously. In that case (a2,b2) may emerge as the outcome without any explicit coordination.

Of course, A may rightly consider that she can more reliably secure benefit for herself by explicit coordination. And so may B. A may say to B: 'I will choose a2 if you choose b2.' And B has every reason to 'cooperate'. But equally, A may move first in full knowledge that B will choose b2 if A chooses a2. In such a case, could we say that A and B are 'cooperating'? Isn't such a case structurally identical to that in *Figure 1*, where once A has acted, B has a clear best action which also happens to be good for A? Does B cooperate with A in this case in any sense differently from that applying in *figure 1*?[20]

Put another way, what would be required to ensure that *figure 3* or some related interactive structure would involve cooperation in some real sense? Consider the interaction illustrated in *figure 4*. Here, there are two outcomes that involve net benefit—(a1,b1) and (a2,b2)—but the benefit in question accrues to only one of the players in each case. A has an interest in B's choosing b1; B has an interest in A's choosing a2. But A's moving first, or making a prior announcement that she intends to choose a1, gives B no incentive to choose b1. And analogously for B in choosing b2. Or at least this is so, unless A cares positively about the payoff to B (or vice versa).

19 These are often termed 'strategic' interactions.

20 One way of thinking about *figure 3* is that it raises questions about the necessity of 'cooperation' only under the assumption that action is simultaneous and pre-play communication is ruled out.

		B's action	
		b1	b2
A's action	a1	(5,0)	(0,0)
	a2	(0,0)	(0,5)

Fig. 4

There is clear scope for genuine cooperation in this case. A and B can enter into a turn-taking arrangement—a kind of 'market exchange'—in which A undertakes to alternate a1 and a2, if B will alternate b1 and b2. In that case, there is an additional 'action' admitted into the game—namely, 'coordinate with the other to secure (a1,b1) and (a2,b2) in turn'. Each has reason to adopt that strategy whenever the game is ongoing. There may be difficulties in securing compliance with the 'contract' if there is an end-point in the sequence of interactions. But here there is no positive temptation to defect. There *would* be such a temptation if the payoffs in the off-diagonals were (1,1) rather than (0,0). Then we would require some additional motivation to secure cooperation reliably (some form of C2).

But notice that the 'cooperation' relevant to the coordination cases mapped by *figure 3* does not require that either player care intrinsically about the payoff that accrues to the other. Indeed, one or other (or both) may care *negatively* about the other's payoff—such that a gain of 2 to B involves a (possibly purely psychic) loss to A of 0.8, say. In this case, A will still increase her net return by 1.2 by coordinating with B. Coordination here requires that each take into account what the other will do (which in turn requires some information about the payoffs to the other under various outcomes) adjusting her own behaviour in the light of that information. If, for example, A knows that B is a 'we-thinker' in the terms already canvassed, then A will predict that B will choose b2 and hence has every reason to choose a2. In doing so, A makes B better off (than if A had chosen a1) but the benefit to B is incidental to A's choice. Put another way, B's benefit is relevant to A's choice of action (and to that extent, B's benefit explains A's choice); but A need not value B's benefit intrinsically. A and B can be sufficiently 'socially alert' to manage to coordinate—but such social alertness does not amount to 'cooperation' in the C2 sense.

If we are to understand the role that dispositions to be cooperative of the C2 kind play in promoting positive interdependence, we should, it seems, focus on the prisoners' dilemma case illustrated in *figure 2*. Certainly, we do not, in general, need players to be motivated to be 'cooperative' in order for those agents to coordinate their actions. Coordination may exemplify PI; and it seems to require

a certain attentiveness on the part of each player to the other's actions in order to secure the benefits on offer. But such attentiveness to the other's actions need not involve any positive evaluations of the benefits of one's actions to others—no desire to promote others' well-being as such alongside or in addition to one's own. In other words, the *mutuality* of benefit/advantage as such need play no role in motivating agents who successfully coordinate: each can be motivated simply by the benefit/advantage to herself.

5 Cooperative Motives?

Much of what we have said so far may appear straightforward enough. There is certainly a clear distinction between the biological setting in which symbiosis is present; and a human setting in which the behaviour of interacting agents is motivated by a desire to be 'cooperative'. What may not be so obvious is that the same ambiguity about cooperation—the same distinction between PI and C2—is in play in the human case. For it is one thing to argue that certain social institutions can solve failures to secure positive outcomes by operating on agents' choices; and entirely another to suppose that such social institutions work by operating on agents' motives—that is by making them more cooperative in the C2 sense.

Again an appeal to game theoretic representations is helpful here. But (as foreshadowed in our discussion of Heath's exposition in section III) there is a complication. We are going to require two sets of 'payoffs' for each outcome (each cell in the matrix representation): the set that defines the 'advantage' to each player; and the set that tracks what motivates the *behaviour* of the agents. In the biological case, these are the same because 'behaviour' is just a property of the emergent evolutionary equilibrium; but in the human case, where social institutions, norms and the like may operate to solve 'collective action problems', the two come apart and must be represented independently.[21]

It is worth noting that tension between the currency of 'advantage' and the currency of expected preference satisfaction implied by this dual 'payoff' structure. If a particular set of preferences is to be *justified* by appeal to the currency of 'advantage', then one cannot hold to the idea that preference satisfaction is the exclusive, foundational ground of normative analysis (as economists are often prone to do).

When a social institution exists and modifies the *choice-relevant* payoffs in a manner that does indeed 'solve' the diagnosed collective action problem, that

21 Simon Blackburn makes exactly this distinction in ch 5 of his *Ruling Passions* (1998).

social institution induces cooperation in the PI sense. Whether C2 (or even C1) cooperation is involved, however, is simply an open question.

The argument that follows involves two steps. The first says a little more about how 'cooperative' motives/concerns are to be defined. The second investigates the logical relations between players having cooperative motives/concerns so defined and the delivery of positive interdependence. The aim is to show that having cooperative motives is neither necessary nor sufficient for PI.

(1) Defining cooperativeness:

C2 cooperativeness is a motivational matter by stipulation. It appears as an element in the 'utility functions' of the players. One way to capture cooperativeness so understood might be in terms of generalized altruism or benevolence: each cares positively about the payoff to the other (to a greater or lesser extent). The extent can be measured, say, by the number of dollars of benefit A is prepared to sacrifice in order to secure a dollar's worth of benefit to B—where payoffs are expressed in terms of dollars.

Another way of capturing cooperativeness might be in terms of each player being motivated by a desire for PI. In this latter case, it is necessary that both actors have positive benefits from the interaction—with benefit defined in terms of the currency of advantage. In this case, where we are assuming the goal of an interdependence that is positive for all, cooperativeness requires that the payoff to the other(s) is a motivating consideration only if the payoff to oneself is not reduced: 'mutual advantage' here is construed in objective payoff terms and is subject to the constraint that all enjoy some level of 'advantage' (some increase in benefit, or at least not a decrease).

To clarify, consider the case of 'we-thinking' mentioned earlier. A key step in the 'we-thinking' exercise is to establish what outcome is 'best for us'. That outcome might be derived simply by aggregating the payoffs to the various players; or alternatively, outcomes might be constrained by the requirement that no-one have her payoff reduced (aggregation subject to vector dominance). We think there is a distinction between what we might call 'mutual advantage' and merely 'aggregate advantage' at stake here: and we shall follow the idea of mutual advantage. So we shall say that an outcome can only be 'better for us' if it is better in payoff terms for each of us; and correspondingly that an outcome is 'no worse for us' if it is no worse for each. (Generalized altruism involves no such restriction.)[22]

22 For the difference this makes to Hume's understanding of justice, see Sayre-McCord 2016.

(2) Mutual advantage => cooperativeness?

Does securing mutual advantage in this sense require that the agents are motivated to secure mutual advantage? It seems not.

It is well-known from the empirical literature (Fehr/Gachter 2000 for example) that the availability of punishment possibilities significantly reduces the proportion of players in two-person prisoner's dilemmas who choose the 'non-cooperative' action. One might think that a capacity to punish would simply replicate the prisoner's dilemma structure at the punishment imposition level of the game at least in the n-person version; after all, punishing a defector in order to reduce the likelihood of defection is to provide a public good to all players. However, the experimental literature suggests that, given a chance to punish, individuals will often do so, even where it is costly to the punisher.[23] So, provided players in the substantive game believe that the probability of being punished if they chose the 'defect' option is sufficiently higher than the probability of being punished when they choose the 'cooperate' option,[24] then they will have an incentive to choose the 'cooperate' option. And they do so choose in significant proportions when punishment possibilities are introduced.

Of course, this does depend on at least some players believing that others are prepared to 'punish' and to discriminate among their targets according to how those people acted in the substantive game. But neither those punishers nor their possible targets seem to be motivated by a desire for mutual advantage as such. For punishers, it may well be a simple desire to punish[25]—and for players, it may be the desire to avoid punishment—that is securing the relevant behaviour.

In many cases, of course, punishment is meted out not by 'vigilantes' operating under spontaneous inclination but by institutions designed expressly for such purposes. The legal system might be justified and/or rationalized on the basis of promoting mutually (or generally) advantageous outcomes; but the individuals who are subject to the system may well comply with the law primarily because of the sanctions that the law imposes.

23 This is also one way to interpret the results of ultimatum games.

24 It seems that in laboratory experiments at least sometimes players are punished even when they choose the 'cooperate' action. It is worth noting that many of the laboratory experiments are constructed in such a way that 'punished' players have no opportunity for retaliation—which one might suspect would inhibit the tendency to invoke punishment in the first place.

25 The picture is complicated somewhat by a tendency to refer to punishment in such prisoner's dilemma settings as 'altruistic' punishment. But as elsewhere we think that this terminology confuses effects and causes: it may be that such punishment has the effect of producing beneficial action. It is doubtful whether such effects constitute the punishers' intentions (if only because punishers seem to punish many who do play the 'cooperative strategy' in the substantive game)

One specific case of some interest to us is that in which the 'punishment' in question does not take the form of inflicting material losses on victims but simply with the assignment or withdrawal of esteem. The thought in this case is that observers form judgments of others' behaviour—and assign approval or disapproval accordingly. The underlying general motivational assumption is that individuals wish to stand high(er) in the opinions of others entirely for its own sake: they value the esteem of others directly and intrinsically.

There is some evidence that the desire for esteem so understood can be a very powerful motivator—especially when aggregated across large numbers of observers. In such a case, individuals may be led to choose the 'cooperate' strategy in prisoner's dilemma settings because so acting garners esteem (or they may refrain from 'defecting' because defecting generates disesteem). Individuals so motivated do not exhibit a concern to be cooperative as such: they are not motivated directly by the mutual advantage on offer but by the esteem/disesteem that attaches to different actions. However, the fact that choosing the 'cooperate'-strategy is a source of positive esteem (or choosing the 'defect'-strategy is a source of disesteem) reflects the fact that observers approve of actions that promote the advantage of the people with whom an agent interacts.

Significantly, forming those attitudes does not show that those observers would themselves be motivated to act to promote mutual advantage when they themselves face the temptations of the prisoner's dilemma payoffs: it shows only that they think that A's acting to promote the advantage of those A interacts with in prisoners' dilemma situations is a 'good thing' and that A deserves approval when A so acts. Equally, those affected persons need not (and probably typically are not) the same persons as supply the esteem and disesteem. In other words, forming the attitudes observers do form need not reflect the interests of those observers in any way. Disapproving of the man that beats his dog does not become implausible because the observer is not himself a dog—or might risk being beaten! The point here is that though a general favourable attitude to mutual advantage and the desire to promote it is implicated in the 'economy of esteem', it is not appropriate to attribute a 'cooperative disposition' to either demanders or suppliers of esteem: to do so is to mistake effect for cause.

More generally, a concern for cooperativeness among participants is not required for securing mutual advantage in the prisoner's dilemma setting. A concern for positive esteem (and to avoid disesteem) may be sufficient.

(3) Cooperativeness => mutual advantage?

But suppose players in broad social dilemma situations do have a concern to act cooperatively: suppose, that is, that players are directly motivated by mutual ad-

vantage (as distinct from just their own advantage). Is that disposition sufficient to generate PI?

Clearly not. First and trivially, though mutual advantage might be one motivating factor it may not be sufficiently strong to induce individuals to choose the 'cooperate'[26] strategy: temptations from individual interest may just be too strong. But more importantly, even where the cooperative concern is strong enough to motivate action (it represents an 'all things considered' desire rather than a 'pro tanto' desire) the actual effects may not be to promote mutual advantage. Hence the important truism: 'The road to hell is paved with good intentions.'

Moreover, it is a common observation that firms in an industry are related to one another via a prisoners' dilemma structure: all would benefit by forming a monopoly cartel and sharing the resultant monopoly profit—but cannot do so because each has a strong incentive to break any cartel deal and undercut other firms. It is equally well-known that were firms sufficiently 'cooperatively disposed' to make the cartel deal stick, that would serve to make consumers worse off by an amount larger than the monopoly profit generated. In other words, maximal advantage would be ruled out by (sufficiently strong) cooperative concerns among producers.

Take another example. The conduct of successful military campaigns involves a significant level of cooperativeness among the individual participants in each combatant force. As one of us has tried to argue elsewhere[27], the members of a military force (at pretty much any level of aggregation) are bound to one another by a prisoners' dilemma structure: best for all if everyone pulls his weight equally, but best for each if others bear the main burden of attack. Accordingly, between two warring parties, there is, in evidence, considerable cooperativeness within parties—but in almost all cases, the relation between parties is Pareto pessimal: even the victors lose in absolute terms.

Put more generally, when we talk of 'we-thinking', much depends on the 'we' that the participants have in mind. Specifically, unless the 'we' relevant for motivation is the same as the 'we' relevant for justification, action motivated by we-thinking can readily lead to suboptimal (to say the least) outcomes. Yet to describe any case where the set of persons whose benefit motivates an actor falls short of the entire set of affected people as one in which there is no motive for being co-

26 We have already complained that the use of terms 'cooperate' and 'defect' to refer to actions is liable to be confusing in any setting where 'cooperation' and 'cooperativeness' are themselves precisely the terms under dispute. However, such terminology is so heavily entrenched in the PD literature that we shall follow it here—making it clear at every point however that it is the action, not the outcome nor the motive for action, that is the relevant referent.

27 See G. Brennan and G. Tullock 1982.

operative seems bizarre. As Smith might put it, a man might exhibit a highly cooperative concern for his 40 or so friends and yet lack such a concern in relation to the countless multitudes on whom he depends for his style of living. To deny that such a person has a 'cooperative' disposition—or claim that he is no more 'cooperative' than his solitary and surly neighbour—just seems perverse.

We think that such considerations establish fairly convincingly that a widespread presence of cooperative dispositions among a population is neither necessary nor sufficient for the securing of PI.

Of course, such logical relations do not establish that there might not be, as a matter of fact, a strong positive correlation between cooperativeness understood as a motivation and the extent of mutual (or general) advantage within a relevant population. Nor does it establish that normative considerations more generally may not be implicated in generating general advantage – even where the role those considerations play is not directly motivational. (That is the message we take from the esteem example.) Moreover, even where there is a positive correlation between cooperativeness, understood psychologically, and the realisation of general advantage, that is not sufficient to show that trying to inculcate a desire for cooperativeness into a population should be a high priority in the pursuit of general advantage—or even much of a priority at all. That requires its own set of arguments.

But perhaps normatively there is more at stake than securing advantage. Perhaps we should care (and perhaps people do care) not just about securing advantage but also about the processes by which general advantage is secured. Perhaps we prefer (or should prefer) to operate in communities in which a concern for mutual advantage is abundant, not because such concern is more effective in producing mutual advantage, but for its own sake. Were that so, it would give additional reasons to be clear about the distinction that we have been emphasizing: that between the *fact* of PI—conceived as a property of social outcomes—and the psychological concerns/motives that participants bring to their deliberations in producing those social outcomes. This might be a simple distinction but it seems to us to be often enough ignored; or obscured. And to return to our point of departure, we think Adam Smith's description of interdependence within a commercial society, and Buchanan's emphasis on 'social symbiosis', and Joseph Heath's elaboration of the 'benefits of cooperation' all miss something important in not drawing it.

6 Conclusion

The object of this paper has been to draw a sharp distinction between three possible meanings that the term 'cooperation' might be assigned.

The three meanings are:

1. Cooperation as a property of an outcome—one in which there is recognizable mutual positive interdependence among those whose actions jointly produce that outcome. The positive interdependence may be represented in terms of payoffs in a game theoretic structure. Cooperation understood as positive interdependence is we think best described simply as 'positive interdependence' (PI). It is usefully illustrated by the case of symbiosis in biological settings where questions of the precise motivation of interacting participants simply do not arise.
2. Because such cooperation can in principle arise as an entirely accidental feature of an interaction, one might want to restrict the use of the term 'cooperation' to situations in which each participant takes the prospect of PI into account in choosing action. But 'taking into account' does not require 'being motivated by'. That distinction is illustrated by 'coordination' games. In such games, the structure of interdependence is such that
   (a) Each has an incentive to choose her actions in the light of what action the other chooses;
   (b) The coordination does promise PI. Each seems likely to be made better of by others acting in the way that they will;
   (c) But the prospect that others are made better off plays no necessary role in motivating the actions of any.

   The characteristic case of this type is what we term 'cooperation as coordination' (C1).
3. Both the foregoing cases are to be distinguished from the case where there is a 'cooperative concern'—where the prospect of PI enters as part of the motivational apparatus of the participants. This third case is by stipulation a matter of agent psychology. We call it 'cooperativeness' or C2 cooperation.

There is a related set of distinctions in play in this paper—according to the object to which the adjective 'cooperative' refers. So, we might think of a cooperative *outcome*; or a cooperative *action*; or a cooperative *motivation*. It may be tempting to think that these uses of the term 'cooperative' are intimately connected—so that for example a cooperative action is just one that produces a cooperative outcome; or one that is motivated by a cooperative concern. That temptation is one that is

to be resisted. As we have been at pains to explain, the connections at stake are problematic.

There are two questions in relation to C2. One deals with the relation between C2 and PI. Our claim is that there is no direct logical relationship. That is, there should be no logical presumption that C2 is productive of general advantage. The second is that, to the extent that justification of a social institution lies in the general advantage it produces, there is no necessary normative reason to support C2. It may be that there are either normative or preference-based reasons relating to the processes whereby advantage is secured. Nothing we have said here rules out the possibility that being cooperatively disposed is a 'virtue' that should be promoted for its own sake; or that there might be a general preference, within a community of interacting agents, for dealing with others for whom PI is a direct concern. But the fact that C2 does not produce maximal advantage in at least some cases carries with it the possibility that any such normative commitment or preference will come at the expense of general advantage itself.

Our aim in this paper has been to clarify issues in relation to 'cooperation'—to draw distinctions that casual use of the term tends to obscure. It has not been our ambition to legislate over meanings. However, it should be clear that we think that clarity would be served if the term 'cooperation' were reserved for the motivational case. To call an outcome 'cooperative' because it produces general (or even universal) benefits—and equally to count *any* process that produces such an outcome as cooperation—just courts confusion.

Acknowledgment: We are extremely grateful to Hartmut Kliemt, Iskra Fileva, and the editors of this journal for comments, as well as to the Department of Philosophy at Groningen University and the Workshop on Ethics and Economics at the University of Colorado at Boulder, for opportunities to present earlier versions of this paper. Remaining errors are our own responsibility.

References

Bacharach, M. (2006), *Beyond Individual Choice*, ed. R. Sugden/N. Gold, Princeton

Bratman, M. (1992), Shared Cooperative Activity, in: *The Philosophic Review* 101, 327–341

Brennan, G./G. Tullock (1982), An Economic Theory of Military Tactics, in: *Journal of Economic Behaviour and Organization* 3, 225–242

Blackburn, S. (1998), *Ruling Passions*, Oxford

Buchanan, J. (1964), What Should Economists Do?, in: *Southern Economic Journal* 30, 213–222

— (1990), The Domain of Constitutional Economics, in: *Constitutional Political Economy* 1, 1–18

Gilbert, M. (2001), Collective Preferences, Obligations and Rational Choice, in: *Economics and Philosophy* 17, 109–119

Heath, J. (2006), The Benefits of Cooperation, in: *Philosophy and Public Affairs* 34, 313–351

Sayre-McCord, G. (2016), Hume on the Artificial Virtues, in: *Oxford Handbook of David Hume*, edited by Paul Russell, Oxford, 435–469

Sugden, R./N. Gold (2007), Collective Intentions and Team Agency, in: *Journal of Philosophy* 104, 109–137