Economics and Philosophy, **26** (2010) 121–148
doi:10.1017/S0266267110000167

VARIETIES OF ALTRUISM

PHILIP KITCHER
Columbia University

1. AN APPROACH TO PSYCHOLOGICAL ALTRUISM

Discussions of altruism occur in three importantly different contexts. During the past four decades, evolutionary theory has been concerned with the possibility that forms of behaviour labelled as altruistic could emerge and could be maintained under natural selection. In these discussions, an agent A is said to act altruistically towards a beneficiary B when A's action promotes the expected reproductive success of B at expected reproductive cost to A. This sort of altruism, *biological altruism*, is quite different from the kind of behaviour important to debates about ethical and social issues. There the focus is on *psychological altruism*, a notion that is concerned with the intentions of the agent and that need have no connection with the spread of anyone's genes. Psychological altruists are people with other-directed desires, emotions or intentions (this is a rough preliminary characterization, to be refined below). Finally, in certain kinds of social scientific research, the important concept is that of *behavioural altruism*. From the outside, behavioural altruists look like psychological altruists, although their motives and preferences may be very different.

In what follows, I shall not be concerned with biological altruism. The focus will be on psychological altruism, and derivatively on behavioural altruism.

Many people believe that psychological altruism does not exist, even that it is impossible. Often they are moved by a very simple line of

I am grateful to Michael Schefczyk for inviting me to participate in the St. Gallen symposium, and for some very helpful discussion. I am greatly indebted to three anonymous referees for *Economics and Philosophy* and especially to Christian List for some excellent comments that have enabled me to make considerable improvements in the final version.

reasoning: when a person acts in a way that could be appraised as altruistic, she acts intentionally; to act intentionally is to identify an outcome that one wants and to attempt to realize that outcome; hence any potential altruist is trying to get what she wants; but to strive for what you want is egoistic; consequently the potential altruist turns out to be an egoist after all. The key to rebutting this argument is to distinguish different kinds of wants and goals. Some of our desires are directed towards ourselves and our own well-being; other desires may be directed towards the welfare of other people. Desires of the former type are the hallmark of egoism, but those of the latter sort are altruistic. So altruists are intentional agents whose effective desires are other-directed. (This response stems from Joseph Butler; a contemporary formulation is given in Feinberg 1975).

I want to develop this approach to psychological altruism further, by giving a more detailed account of the character of other-directed desires, and thereby bringing into the open some of the complexities of the concept of altruism – exposing the *varieties of altruism* and the factors on which they depend. I will try to show how psychological altruism can be represented in the standard idiom of decision theory and game theory, and will eventually use my representation to suggest a way of thinking about some classical social and political questions. First, however, the basic account.

The other-directed desires that are central to the defence of the possibility of altruism are desires that respond to the altruistic agent's recognition of the impact of her actions on the situations of others. To be an altruist is to have a particular kind of relational structure in your psychological life – when you come to see that what you do will affect other people, the wants you have, the emotions you feel, the intentions you form change from what they would have been in the absence of that recognition. Because you see the consequences for others of what you envisage doing, the psychological attitudes you adopt are different. You are moved by the perceived impact on someone else.

So far, that is abstract and vague. I will motivate the underlying idea with a simple and stylized example, and will then offer a more precise definition.

Imagine that you are hungry, and that you enter a room in which some food, a pizza say, is spread out on a table before you. Suppose further that there is nobody in the vicinity who might also be hungry and want all or part of the pizza. Under these circumstances, you want to eat the pizza; indeed you want all of it. If the circumstances were a bit different, however, if there were another hungry person in the room or known to be in the neighbourhood, then your desire would be different: now you would prefer the outcome where you share the pizza with the other person. Here your desire responds to your perception of the needs and wants of someone else, so that you adjust what you might otherwise

have wanted so as to align your desire with the wants you take that other person to have.

This is a start, but it isn't sufficient to make you an altruist. For you might have formed the new want when you see that someone else will be affected by what you do, because you saw profitable future opportunities for accommodating this other person. Maybe you envisage a series of occasions on which you and your fellow will find yourselves hungry in pizza-containing rooms. You see the advantages of not fighting, and of not simply having all the food go to the first person who enters. You resolve to share, then, because a future of cooperation will be better from your point of view. For real psychological altruism, the adjustment of desires must not be produced by this kind of self-interested calculation.

I offer a definition of '*A* acts altruistically towards *B* in *C*' – where *A* is the agent, *B* the beneficiary, and *C* is the context (or set of circumstances). First, two contexts *C* and *C** will be said to be *counterparts*, just in case they differ only in that, in one (*C** say) the actions available to *A* have no consequences for *B*, whereas in the other (*C*) those actions do have consequences for *B*. *C** will then be the *solitary* counterpart of *C*, and *C* will be the *social* counterpart of *C**. If *A* forms different desires in *C** from those *A* forms in *C*, the set of desires present in *C** will be *A*'s *solitary* desires (relative to the counterparts *C* and *C**). Given these specifications:

A acts (psychologically) altruistically with respect to *B* in *C* just in case

(1) *A* acts on the basis of a desire that is different from the desire that would have moved *A* to action in *C**, the solitary counterpart of *C*
(2) The desire that moves *A* to action in *C* is more closely aligned with the wants *A* attributes to *B* in *C* than the desire that would have moved *A* to action in *C**
(3) The desire that moves *A* to action in *C* results from *A*'s perception of *B*'s wants in *C*
(4) The desire that moves *A* to action in *C* is not caused by *A*'s expectation that the action resulting from it would promote *A*'s solitary desires (with respect to *C* and *C**).

(1) tells us that *A* modifies his desires from the way they would otherwise have been, when there is an impact – more accurately, when there is a perceived impact[1] – on the wants of *B*. (2) adds the idea that the desire, and the behaviour it directs, is more in harmony with the wants attributed to *B* than it would have been if *B* were unaffected by what was done; (it is possible to modify your desires in response to the perceived wishes of another, but to do so in a way that *diverges* from their perceived

[1] I shall consider cases in which agents have mistaken beliefs later. For the time being, I suppose that the parties get things at least roughly right.

wants – that's spite!). (3) explains that the increased harmony comes about because of the perception of B's wants; it is not, say, some caprice on A's part that a different desire comes into play here. Finally, (4) denies that the modification is to be understood in terms of A's attempt to promote some desire that would have been present in situations where there was no thought of helping or hurting B; this distinguishes A from the pizza-sharer who hopes for returns on future occasions when B is in the position of disposing of the goods.[2]

Consider next how to integrate this basic account into the language of game theory and rational decision theory. I imagine that agents recognize a set of available actions and that they associate each such action with a *valuation*, a real number that measures the value they ascribe to the expected outcome of the action. We imagine A in C assigning the value v_{iC} to the outcome of the ith available action. In C, A also supposes that B would assign v_{iB} to that outcome. In C^*, the solitary counterpart of C, A would assign u_i to the outcome of the ith available action. For A to act altruistically towards B in C, the maximal value of v_{iC} must be different from u_i (condition (1)), $|v_{iC} - v_{iB}| < |u_i - v_{iB}|$ (condition (2)), this inequality obtains because of A's attribution of the v_{iB} (condition (3)), and the inequality does not come about because, in C, A envisages the ith act as producing an outcome that would be assigned maximal value from the perspective of C^* (condition (4)). Here, it is important to appreciate that the expected outcome of the ith action, when performed in C, may be distinct from the expected outcome of that action performed in C^*. The outcome resulting in C simply cannot be achieved within C^*. Intuitively, within C^*, the hungry A doesn't think of a long series of happy food-sharing. Simply leaving the pizza cannot be seen as sharing, and thus not as leading to occasions on which A would be the recipient. Confronted with C, non-altruistic A decides to share the pizza because of the expected food-sharing future. The value of that future might be represented in C^* but not as the value of the expected outcome of some available act. The value ascribed would, however, exceed all the values assigned to the expected outcomes of the available acts.

The more formal treatment brings out an oddity in the basic account as I originally presented it. Effectively, I allow for altruism provided that the desire that moves the agent to act responds to the perceived wants of the beneficiary. Plainly, however, it would be strange if the recognition

[2] There are complexities here. A solitary desire might be a standing wish that some other person have a particular positive attitude towards you. One can want the approval or liking of others, even in contexts where there is no possibility of one's affecting the lives of those people. In particular, it is possible to have a standing wish to coerce (or to press for) the liking of another, and desires of this sort should count as solitary. I am grateful to an anonymous referee for drawing my attention to these complications; I suspect that it would require a much longer essay to work them out completely.

of potential impact on another were not to affect all one's thinking about possible outcomes. So, we might say, a more realistic form of altruism is to think of a function f that maps the u_i and the v_{iB} to v_{iC} never increasing the distance from v_{iB} and sometimes diminishing it. In other words:

> $v_{iC} = f(u_i, v_{iB})$ where $\forall j\ |f(u_j, v_{jB}) - v_{jB}| \leq |u_j - v_{jB}|$ with the inequality holding strictly for some j.

That requirement, however, would allow for altruism in cases where only some minor adjustment is made to valuations that don't affect the action performed (imagine that A simply modifies his assessment of the previously least-valued option, so that, while still highly disvalued, it is no longer in last place). We can strike a compromise between the new proposal and the original form of the basic account by demanding:

> $v_{iC} = f(u_i, v_{iB})$ where $\forall j\ |f(u_j, v_{jB}) - v_{jB}| \leq |u_j - v_{jB,}|$, and the top-ranking action in C is different from the top-ranking action in C^*.[3]

The function f introduced here will be called the *preference-response function*. In general, one can suppose that people have preference-response functions not only to individuals, but also to *collections of several individuals*. A may modify her preferences in light of the preferences she attributes to members of an ordered n-tuple of other agents. I shall not consider how to extend my treatment along these lines, and will chiefly consider dyadic interactions.[4]

The stylized pizza example allows for a very simple treatment of the preference-response function. You can take the other person's wishes into account by deciding just how much of the goods to give them. Perhaps f is a weighted average,

$$f(x, y) = \theta x + (1 - \theta)y \text{ where } 0 \leq \theta \leq 1.$$

Suppose you think that the value of a whole pizza is 10, the value of $^3/_4$ pizza is 9, the value of $^1/_2$ pizza is 7, the value of $^1/_4$ pizza is 1, and the value of no pizza is 0. You attribute the same valuations to the potential beneficiary, and contemplate three possibilities: eating it all, giving a quarter, and sharing evenly. If your valuations are determined by the parameter θ, the relevant values are: 10θ, $8\theta + 1$, 7. Giving a quarter is never the best option. For if that is to be preferable to consuming the whole thing, then $2\theta < 1$; but in that case $8\theta + 1 < 5$ (and hence < 7). Sharing evenly is preferable provided $\theta < 0.7$.[5]

[3] Intuitively, the response to B is sufficiently strong to modify the preference that is expressed in the action. I am grateful to Christian List for helping me to see that an earlier formulation was inadequate.

[4] Here, too, I am indebted to Christian List.

[5] I offered this analysis in Kitcher (1993). Here I set it in a more general approach to psychological altruism.

https://doi.org/10.1017/S0266267110000167 Published online by Cambridge University Press

This should suffice to introduce the basic account. I am now going to introduce refinements and complications.

2. SOME DIMENSIONS OF ALTRUISM

Altruism is a multi-dimensional notion. Individuals can be placed in a multi-dimensional space where complete egoism is represented by a single plane, and the various forms of altruism range over the entire rest of the space. There are many different ways to be (something of) an altruist.

My spatial metaphor introduces into the discussion of psychological altruism something akin to the way in which behavioural geneticists think about the dependency of behaviour on the environment: they say that the *norm of reaction* of a genotype is a graph (or function) that shows how the phenotype of an organism with that genotype varies with the character of the environment. This notion presupposes some way of representing the various possible environments along one or more dimensions (something which is, in practice, impossible to specify completely). By the same token, a person's *altruism profile* is a graph (or function) that represents the variation in the difference between the valuations assigned in solitary and social contexts for all potential beneficiaries (or collections of such beneficiaries) across all possible pairs of solitary and social contexts. Plainly, any complete representation of the axes is as unavailable as it is in the case of genetics. Yet it is easy to appreciate the fact that egoists are people whose altruism profile takes the form of a plane (or hyperplane), since the difference between the valuation assigned in a social context and its solitary counterpart is, for them, always and everywhere zero, no matter which individuals are involved.

Your altruism profile (where you are located in altruism space) is determined by five factors: the *intensity* of your responses to the perceived wishes of the other, the *range* of people to whom you are prepared to make an altruistic response, the *scope* of contexts in which you are disposed to respond, your *discernment* in appreciating the consequences for others, and your *empathetic skill* in identifying the desires others have. As just noted, egoists are people who never respond to anyone else in any context: their discernment and empathetic skill can be as you please, for these are never called into play.

Altruists are not like that. They modify their desires to align them with the perceived desires of at least some others in at least some contexts. Their responses may be more or less intense, in that they give weight to the perceived desire of the beneficiary, rather than the desire that would have been present in the solitary counterpart of the context in question.[6]

[6] I assume throughout the ensuing discussion that condition (4) in the analysis of altruism is satisfied: my envisaged subjects aren't calculating ways to advance their solitary desires.

My treatment of the stylized example in terms of weighted averaging provides a clear paradigm for intensity. If $f(x, y) = \theta x + (1 - \theta)y$ where $0 \leq \theta \leq 1$, then egoists are those who set θ at 1. People for whom $\theta = 1 - \varepsilon$, where ε is tiny, are only altruists in a very modest sense: they will only act to advance the wishes of others when the perceived benefits to others are enormous compared to the gains for themselves that they would forfeit – they may suffer the scratching of their finger in order to avoid the destruction of the world, but are not prepared to make larger sacrifices. People for whom $\theta = 0$, by contrast, are completely self-abnegating. They abandon their own solitary desires entirely, taking on the wishes they attribute to the beneficiary. In between, we find 'golden rule' altruists, for whom $\theta = {}^1/_2$, who treat the perceived wishes of the other exactly as they do their own solitary desires.[7]

Most altruists, indeed probably all, don't have a fixed value of θ (or, more generally, a fixed intensity of response, as measured by the extent to which the preference-response function gives priority to the perceived wishes of the other) that applies with respect to all potential beneficiaries and all contexts. There are many people to whom we would rarely make an altruistic response: these people effectively fall outside the range of our altruism. Even with respect to those to whom we are disposed to respond, there are many contexts in which we don't take their perceived wishes into consideration. Often our altruistic responses to some are coloured by indifference to others: parents who make sacrifices to help their children obtain things the children passionately want frequently don't take into account the wishes of other children (or the altruistic desires of the parents of the other children).

Someone's altruism profile will usually show a relatively small number of people to whom the focal individual responds, frequently with significant intensity, across a wide set of contexts. The beneficiaries lie at the centre of the range of altruism for the focal individual, and the scope for these beneficiaries is wide. As we consider other potential beneficiaries more distant from the centre, the scope narrows and the intensity falls off, until we encounter people to whom the focal individual makes no altruistic response at all. Henceforth, I will conceive of the range of A's altruism in terms of the metaphor of centre and periphery: the centre is the select set of potential beneficiaries for whom A's response is relatively intense across a relatively wide scope of contexts; at the periphery, the intensity of the response and the scope of contexts narrow and vanish.[8]

[7] Here I recapitulate views I advanced in Kitcher (1993); they should now be understood in light of the basic account offered in section 1.

[8] In extending the spatial metaphor in this way, I effectively presuppose a way of representing the dimensions of altruism space so that individuals who often excite altruistic responses are grouped together.

Someone's character as an altruist isn't simply fixed by the factors so far considered – intensity, range, and scope – because there are also significant cognitive determinants of altruism. *A* may make no response in a particular context because *A* fails to understand the consequences for *B*; perhaps *A* does not differentiate the social from the solitary counterpart. Often this is an excusable feature of our fallibility, for the impact on the lives of others may be subtle; we may just not see that following some habitual practice – buying at the most attractive price, or investing in promising stocks – has extremely deleterious consequences for people about whose welfare we care. Evidently, however, there are grades of acuity with respect to consequences, and we admire those who recognize the intricate ways in which others can be affected, while blaming those who 'ought to have seen' the damage their actions might cause.

Similarly, there are degrees to which people are good at gauging the desires of others. Almost everyone is familiar with the well-intentioned person who tries to advance the projects of an intended beneficiary but who is hopelessly misguided about what the beneficiary wants: almost everyone has had a friend or relative who persists in giving presents that are no longer appropriate for the recipient's age or conditions of life. It would be hard, I think, to declare that people who attribute the wrong desires to their beneficiaries, or who overlook consequences for those whom they intend to benefit, are not acting altruistically when they carry out their variously misguided plans – their intentions are, after all, directed towards doing good for others – but their altruism needs to be differentiated from that of their more acute fellows. Hence *A*'s altruism profile depends on two cognitive features: *A*'s skill in understanding the nature of a social counterpart to a solitary context, and *A*'s ability to empathize with *B* (that is to ascribe desires *B* actually possesses).

A simple reaction to the prospect of human egoism is to propose that people living in community with one another – or even all people – should be altruistic. Recognizing the variety of altruism profiles shows us that this thought is far too simple. There isn't a *single* way to be an altruist, and, consequently, the commendation of altruism must be given more specific content. What kind of altruist should we urge someone to be? Moreover, is it right to suppose that the best state of the community (or for the entire species) is achieved by having each member (each person) manifest the same altruism profile? In the last two sections of this essay, I'm going to suggest that central issues in social and political theory can be approached in light of these questions. First, however, some more explanations and complications are in order.

You might think that the questions have straightforward answers. With respect to the cognitive factors, accuracy is always preferable: ideally people should be aware of potential impact for others and should understand what others want. For issues of intensity, range, and scope, we ought to aim at golden-rule altruism with respect to all people across

all contexts. The demand for cognitive accuracy is more plausible, but still not uncontroversial. Debate about the second part of the proposal arises in obvious ways. There are grounds for thinking it valuable that people should develop strong ties with some others – that the range of their altruism should have a definite centre; from Freud's worries about the 'thinning out' of our libido in the development of civilization to familiar philosophical examples about parents who wonder whether they should save the drowning child who is closer, when their own drowning child is further out and harder to rescue, there's a spectrum of troublesome cases that arouse suspicion about completely impartial altruism. Moreover, in a world with finite resources, the desires of others may conflict. If A accurately perceives that both B_1 and B_2 want some indivisible good, it shouldn't be automatic that A's desire should be formed using a function that treats B_1 and B_2 symmetrically. (We may, for example, want A to respond to aspects of the history of the situation, including what B_1 and B_2 have previously done.) None of this is to deny that there may be a level at which we want altruism profiles to respond impartially to others, but merely to deny that the impartiality we want can be adequately captured as 'golden-rule' altruism towards all people in all contexts.

3. COMPLICATIONS

I hope that offering the basic account and exposing the factors that underlie the wide variety of altruism profiles helps to bring the concept of psychological altruism into focus. It does not, however, provide a complete account of the varieties of psychological and behavioural altruism. My next task is to note some complications.

Behavioural altruism. Earlier, I suggested that behavioural altruists are people who look from the outside as if they were psychological altruists, even though their motivations may be different. I will use the term 'behavioural altruism' inclusively: psychological altruists count as behavioural altruists, but so do other people whose conduct is directed by solitary desires. The conditions for behavioural altruism are the first two laid down for psychological altruism, recapitulated here for convenience.

A acts (behaviourally) altruistically with respect to B in C just in case

(1) A acts on the basis of a desire that is different from the desire that would have moved A to action in C^*, the solitary counterpart of C.
(2) The desire that moves A to action in C is more closely aligned with the wants A attributes to B in C than the desire that would have moved A to action in C^*.[9]

[9] One could generate a slightly different concept by imposing the first *three* of the conditions on psychological altruism. So far as I can see, this makes little difference: the important point is that the fourth does not need to be satisfied.

https://doi.org/10.1017/S0266267110000167 Published online by Cambridge University Press

The source of *A*'s behavioural altruism *might* be the perception of *B*'s wants, and the action *might* be generated in a way that is independent of any of *A*'s solitary desires – *A* might be a psychological altruist. On the other hand, *A*'s desire to act as he does could be the product of his thought that behaving in this way will produce results that he wants on the basis of standing desires that are present quite independently of the contexts in which his actions affect *B*. He may recognize the value of entering into schemes of reciprocation, or be concerned with the approval of third parties, or fear punishment if he elects some other option.

For many types of social scientific inquiry, the notion of behavioural altruism is the crucial one. What is of interest is *how* people behave, not *why* they come to form the intentions to act as they do. Experimental economists are sometimes interested in showing that subjects do not belong to the fictitious species *Homo economicus*, and their ingenious studies can succeed without probing the motivations.[10] Revealing that people will share with strangers is an important result, whether what lies behind the sharing is the type of identification with the beneficiary celebrated in psychological altruism or a desire to behave acceptably in the eyes of those running the experiment, or a sense of shame at the thought of telling friends and family about how one acted, or a determination to follow particular principles and ideals of conduct.[11]

Human motivations are sufficiently varied and complex that it is doubtful that any of the most prominent experiments *could* demonstrate psychological altruism. That doesn't matter, since tendencies to behavioural altruism, whatever the underlying psychological causes, are neglected in classical economic modelling. Showing *behavioural* altruism suffices for initiating reforms in economic theory. Moreover, the machinery introduced in the previous sections can easily be adapted to represent the preference structures of behavioural altruists. With respect to this notion, too, we can usefully introduce the idea of a preference-response function, use it to characterize the intensity of the altruism, distinguish the scope and range of a behavioural altruist's profile, and even identify cognitive factors that affect the profile.[12]

[10] See the writings of Ernst Fehr and his associates. An excellent overview is provided in Fehr and Fischbacher (2005).

[11] I shall leave it open whether the notion of psychological altruism should be further developed to include cases in which preferences are adjusted because the subject feels respect for some general moral maxim. Examples of this type might involve no representation of the beneficiary or of her preferences. Such examples could be viewed as a second main type of psychological altruism or as belonging to a special category that should be confused neither with psychological altruism nor with psychological egoism.

[12] Some adjustment is required here. Behavioural altruists might be concerned with the attitudes of people other than the potential beneficiary, and so be susceptible to errors

Why, then, did I begin with psychological altruism? The short answer is that psychological altruism seems an important concept in understanding our ethical life. Moreover, without it, I don't think we can fully understand the possibility of the motivations that underlie many instances of behavioural altruism that are not cases of psychological altruism.

For consider the experimental subject who shares because he thinks that those running the experiment would disapprove of his walking away with as much as possible, or because he anticipates the difficulty of telling his wife about the details of the experiment. Mundane motives of this kind are only possible once there is a recognized social system of norms, including a directive to share. I claim (bluntly!) that human beings could not have arrived at any such system unless they had antecedently had (limited) dispositions to psychological altruism. Those dispositions enabled our ancestors to live together, but their limitations produced social friction. Tens of thousands of years ago, some of them began 'the ethical project', discussing and introducing rules for behaviour within a small group, and our current systems of norms are the outgrowth of their efforts. Ethics, as I conceive it, is a social technology whose initial function was to amplify our (weak) psychological altruistic dispositions.[13]

Hominid societies needed psychologically altruistic dispositions to get this project off the ground.[14] Once the project is in place, the motivations for behaviour can become more complex, and more difficult to probe. Many, probably the vast majority, of our actions are affected by a number of distinct factors, and it is often probably impossible to acquire convincing evidence that the fourth condition for psychological altruism is met. For many explorations of human social behaviour, the wise researcher gives up on the question and is content to frame the inquiry in terms of behavioural altruism. Nevertheless, psychological altruism remains crucial, since without it the range of behavioural altruism would be greatly limited.

Wants, interests, and paternalism. From the beginning, I have supposed that altruistic *A* responds to the perceived *wants* of *B*, but this is plainly not always the case. The altruistic mother doesn't align her wants with the wishes of the young child who is vigorously resisting the necessary

about the preferences of these other people. If you want to impress a third party with your actions, your success is dependent on identifying what the person will find impressive.

[13] This view is explained and defended in Kitcher (forthcoming).

[14] The work of Jane Goodall (1988) and of Frans De Waal (1996), provides evidence for *some* instances of psychological altruism among our evolutionary cousins (and among our hominid ancestors). I don't suppose either that hominid tendencies to psychological altruism were extensive, or that they furnish the core of our ethical attitudes; see my contribution to De Waal (2007) and Kitcher (forthcoming).

medicine. Parallel to the account I have outlined, we could approach altruism in terms of responses to the perceived *interests* of others, which I will understand, for present purposes, to be the wants those others would have if they were clearly (and coolly) to deliberate on the basis of all the facts.[15] With respect to young children, it often seems evident that accommodating wishes rather than interests is a defective form of altruism – perhaps not even worthy of the name. Should such paternalism be preferred across the board?

You might say this: to be an altruist is to identify with the other person, and that is to take her seriously as an agent (at least once she is mature); hence, even though one may think her wishes misguided, as unlikely to promote what she would want were she to be in the ideal position of full information and calm reflection, those actual wishes are to be respected. Or you might say something different: to be an altruist is to care about the other's good, and that isn't what she actually – and myopically – wants, but rather what she would want were she better situated to judge; so one should align one's desires with the desires one supposes that she would form, were her ignorance, or limitation of vision to be remedied. Both types of considerations have enough force to incline me to an inclusive view: in many contexts, it is reasonable to see either option as altruistic; there are paternalistic and non-paternalistic forms of altruism. Not in all contexts, however. On some occasions, it would be arrogant to substitute one's own judgement about what the intended beneficiary would want, given the benefit of an idealized perspective. If *A* has evidence that would support the judgement that *B* has thought hard about her valuations of outcomes, if *A*'s own reflections on those outcomes are hasty and uncritical, then *A* is quite wrong to override *B*'s expressed wants, even though, by chance, *A*'s particular judgement on this occasion would be closer to what *B*, given full information and cool reflection, would actually desire. By the same token, if *A* has excellent evidence that *B* is missing a crucial item of information, if there is no opportunity to present the salient facts to *B* – and thus induce a change in *B*'s desires with which *A*'s own valuations could then be aligned – then responding to *B*'s actual wants would seem to rest either on indifference to her welfare or on disrespect for her powers of rational revision. Hence I suggest that, in some circumstances, only alignment with the wishes counts as genuine altruism, while in others only alignment with the interests is altruistic.

[15] This is only a gesture towards a proper definition. It is notoriously hard to explain just what the situation of the ideal deliberator is supposed to be. As Thomas Schelling has also argued, there may often be serious difficulty in deciding what the 'real interests' of a subject are, in figuring out 'who is Jekyll and who is Hyde'; see Schelling (1984). The altruistic response that considers the interests of the other is akin to Smith's conception of sympathy, where we don't see the world as others see it but react to their objective situation as we conceive it (Smith 1984; henceforth TMS).

How do we tell whether paternalism is appropriate? Here the preceding discussion of behavioural altruism can prove helpful. Once norms are in place, they reshape our altruism profiles, pointing us towards responding to particular people in particular contexts. A further part of the pointing consists in directing us either towards the actual wants, or to those that would be acquired, given a better grasp of the facts. Once the ethical project is well underway, the concept of psychological altruism is articulated according to ethical maxims – but, as I have claimed, a prior, pre-ethical, notion of psychological altruism was needed to help that project off the ground.

Higher-order altruism. Among the wishes that a potential beneficiary B may express, or A may recognize B as having, are wishes directed towards another person. In some instances, A's altruism may take the form of an action-directing desire to benefit B^*, even though no response of that intensity would arise towards B^* in that context independently of the recognition of the altruistic response B makes to B^*. Part of our altruism consists in promoting the altruism of those towards whom we are altruistic.

Often, of course, the person towards whom you make an altruistic response also responds altruistically to you. In the notation of the last paragraph, $B^* = A$.[16] Hence arises an example of an important phenomenon that I will call *higher-order* altruism. Sometimes it is altruistic to allow others to express their altruism towards you, even though that means that your own solitary wishes are satisfied.

It might appear that responding to someone's altruistic desires to respond to oneself is essentially selfish, and that I am simply decorating egoism with a more attractive name. On the contrary, there is an important difference between people who are genuinely directed by their desire to promote the wishes of another, and those who aim at some personal benefit. In many instances, of course, there are reasons to wonder about the purity of your motives, to ask if the choices you make are centred on a friend's wishes to be kind to you, rather than on anticipation of the outcomes produced by the kind outcomes. Ironically, the self-scrutiny expressed in questions like this can often indicate the presence of psychological altruism: perhaps the real psychological altruists are those who worry most about whether they are acting selfishly! I suspect that there are examples of psychological altruism in which people help their friends to help them, and that in most, if not all, of these examples, it is difficult for anyone, including the altruist, to be sure about the motivations. That signals, once again, the points I have been making about the relative tractability of the notion of behavioural altruism.

[16] If the term had not already been pre-empted, 'reciprocal altruism' might be a more appropriate label for the types of altruism considered in the next paragraphs.

When there is a pattern of repeated interactions, with mutual responsiveness, in which the parties support one another's desires, in which they sometimes allow altruistic responses to themselves to go forward, a new and important form of higher-order altruism can emerge. There are first-order goals that the individual actions advance (*A* acts to give *B* what *B* wants), and second-order ends that are supported (*A* permits – or fosters – *B*'s desire to give *A* what *A* wants). Beyond this, both parties can find the entire pattern of the interactions valuable in its own right, so that, quite independently of the first-order goals that are achieved, each attaches a value to the *process* of mutual accommodation. In some circumstances, that value can be sufficiently great to support a preference for an outcome in which *neither* of the participants achieves any primary goal.

Literature abounds with intricate examples in which characters make mutual sacrifices that they value because of the ways in which what is done reflects the concerns each has for the other; a poignant, but extremely intricate example is the decision of the Ververs, Maggie and Adam, to live on different continents (at the end of Henry James' *The Golden Bowl*). I will use a far simpler (and slightly mawkish) example, from O. Henry's famous short story, *The Gift of the Magi*.

Jim and Della, a happily married young couple, who live in New York, are quite poor. Jim has inherited a pocket watch, but he cannot wear it because he lacks a chain. Della has abundant hair that would be further enhanced by an ornamental comb. Christmas approaches and each wants to give the other a present: Jim spots an appropriate comb, and Della discovers a good chain. Neither has enough money to buy the desired gift. But, on Christmas Eve, Jim pawns his watch to buy the comb, and Della cuts off and sells her hair to buy the chain. They exchange presents: Jim receives a useless chain and Della gets a useless comb.

The solitary counterpart of the options is easily represented (where Della's actions correspond to columns and Jim's to rows).

	Cut	Keep
Pawn	$<-5, -5>$	$<-5, 10>$
Keep	$<10, -5>$	$<0, 0>$

Assume Jim accurately assigns the parameter θ^* to Della, and that she, equally accurately, assigns the parameter θ to him. No matter how many times they adjust their preferences by taking into account the values assigned by the other, iterating the application of the preference-response function, the values of <Pawn, Cut> and <Keep, Keep> will always remain at $<-5,-5>$ and $<0, 0>$ respectively (since the weighted average of a number and itself is always that number). The values of <Pawn, Keep> and <Keep, Cut> will depend on the extent of the iteration: at the first stage, for example, Jim will assign $-5\theta + 10(1 - \theta)$ to <Pawn,

https://doi.org/10.1017/S0266267110000167 Published online by Cambridge University Press

Keep> and Della will assign $-5(1-\theta^*)+10\theta^*$ to that outcome; at the second iteration, Jim will assign $-5\theta^2+10\theta(1-\theta)$ $-5(1-\theta)(1-\theta^*)+10\theta^*(1-\theta)$ to the same option, and so forth; it is not hard to see that both Jim's and Della's assignments to <Pawn, Keep> and <Keep, Cut> will both always be larger than -5. This reflects the fact that, from both their points of view, however long they modify their preferences in response to the altruism of the other, in response to the other's altruistic response to their individual tendencies to altruism, and so on, the analysis given by my proposed framework always counts <Pawn, Cut> as inferior *for both* than one of the successful acts of gift-giving.

O. Henry, however, concludes his story with a scene in which Jim and Della are portrayed as happy – 'Of all who give and receive gifts, such as they are wisest. Everywhere they are wisest. They are the magi', in the story's concluding lines. That conclusion cannot be reached if one views higher-order altruism simply as iterated accommodations of altruistically modified valuations. I suggest that it points to the important fact about altruism I have noted, to wit that the original solitary value ascribed to an outcome is sometimes negligible in comparison to the value that whatever outcome be reached be the product of a serious process of mutual engagement with the wishes of another person. The original value of the comb is different from the value of the comb-as-offered-by-Jim, just as the original value of the chain is different from the-chain-as-offered-by-Della. When Della and Jim exchange gifts, each sees that the other has made a sacrifice, and the value of the sacrifice is far greater for them than any of the values that might have been attributed independently (in solitary counterparts) to the material articles involved. The form of the valuations is thus:

	Cut	Keep
Pawn	$<-5+V, -5+V>$	$<-5\theta+(10+V)(1-\theta),$ $-5(1-\theta^*)+(10+V)\theta^*>$
Keep	$<(10+V)\theta-5(1-\theta),$ $(10+V)(1-\theta^*)-5\theta^*>$	$<0, 0>$

When V is very large in relation to the values of the material goods ($V >> 10$), it is easy to see how <Pawn, Cut> can be the preferred option for both, so long as θ and θ^* are not too large.

In general, some interesting and important types of human altruism will ascribe extra value to outcomes because of the ways in which they result from mutual perception and mutual accommodation. This point is not only significant for the understanding of altruism; it should be a recognized feature of any sophisticated form of consequentialism that the value of an outcome may vary according to the character of the causal process that produced it.

https://doi.org/10.1017/S0266267110000167 Published online by Cambridge University Press

I don't suppose that the considerations I have adduced here exhaust all the complications of the notions of psychological and behavioural altruism. Nevertheless, in the concluding sections of this essay, I want to explore two ways in which the ideas I have been presenting can be applied to classical discussions of human social interactions.

4. ADAM SMITH'S LEGACY: THE TWO-SECTOR SOCIETY

Once you recognize the varieties of altruism, there is a very obvious way to think about contemporary affluent societies. Those societies are divided into two sectors, in the sense that there is a set of contexts within which tendencies to altruism are encouraged and viewed as praiseworthy, and another set within which they are discouraged, or, at best, considered to be neutral. Economic transactions fall within the second of these sets. Early social training encourages individuals to have altruism profiles that take a very particular form: in 'public' (market, economic) contexts, they are not supposed to make *any* altruistic responses to the people with whom they deal; in 'private' contexts (family and friendly relations), quite different things are expected. This is one among many possible ways of shaping altruism. Does it have any especially privileged status?

It is useful to approach this question by beginning with Adam Smith and with what early commentators dubbed 'das Adam Smith Problem'. Famously, Smith wrote two books, one of which (TMS) offers a theory of human behaviour and a theory of human ethical behaviour, based on the thought that people have a natural disposition to feel sympathy for others, while the other (Smith 2001; henceforth WN) treats members of society as rational self-interested agents whose goals are typically positively correlated with personal wealth (they belong to the fictitious species *Homo economicus*). Smith's characters look like psychological altruists in my sense (with a pronounced tendency to paternalism, since Smithian people anticipate how they themselves would feel if they were in the other's shoes).

The Problem arises from the difference between the two conceptions of human nature. Many defenders of Smith have tried to wave it away on the grounds that Smith was pursuing different projects in his two works: it is frequently said that one is entitled to make different assumptions about people when one is considering how they ought to behave than when the topic is how they actually do behave. This response is, however, extremely superficial. It is essential to the account of morality developed in TMS that human beings are, by nature, inclined to sympathy with one another. Such inclinations are the basis of – although they are not to be identified with – the refined moral sentiments whose proper development is Smith's specific concern. Now if people do indeed have this basic nature, then

that fact about us cannot simply be forgotten in articulating the analyses of WN. When Smith constructs, as he often does, informal models of how rationally self-interested agents would carry out their business under various conditions, we have no right to expect that these models conform to the behaviour of actual people who, *ex hypothesi*, have basic inclinations that are in no way represented in the models. In the terms I have used in offering an approach to altruism, we ought to expect that the valuations assigned by real – altruistic – agents will diverge from those attributed to *Homo economicus* because people typically modify their solitary desires to align them with the wishes of certain others. If we take TMS seriously, we ought to wonder if the (pre-global-warming) Scottish wine-grower who is considering shifting to a different line of work (woollens, say) would be moved by the envisaged plight of the horticultural staff who will lose their jobs: will his valuations be defined purely by profit, or will they respond to the wishes of his workers?

In what may be the most famous line from WN, Smith announces what can easily appear to be a denial of the account of our altruistic nature offered in TMS:

> It is not from the benevolence of the butcher, the brewer, or the baker, that we expect our dinner, but from their regard to their own interest. (Smith 2001:15.)

Notice that Smith doesn't deny here that the butcher, the brewer and the baker *are* altruists (and the next sentence hints that they do possess 'humanity'); instead, he suggests that we don't expect their tendencies to altruism to respond in the context of doing business. He is identifying a sphere of trade and commerce as lying outside the scope of the altruism of the members of the societies he is analysing. We can bring TMS and WN into harmony with one another by recognizing the varieties of altruism, and suggesting that commercial and full-fledged capitalist societies contain agents with special altruism profiles, agents for whom economic contexts suppress altruistic responses: in such contexts the preference-response function is the identity function, or, in the averaging simplification, $\theta = 1$.[17] That raises an obvious question: Are there any reasons to think that this collection of altruism profiles is likely to emerge or be maintained?

I can envisage a Smithian answer. Imagine a society in which commercial transactions are sensitive to altruistic responses. In that society, for each seller, there's a class of potential customers – the class of friends, F – for whom, in contexts of trade, $\theta < 1$. Suppose that p is the price

[17] Smith might allow that there are occasional deviations – nepotistic corruption, for example – but that these are negligible.

that a seller who made no altruistic response (for whom $F = \emptyset$) would sell his wares. If p_i is the price that the ith member of F would be content to pay (I will assume its value varies with the person's needs and background resources, but that it is typically less than, and never greater than, p), then the seller sets the price for this person at $\theta p + (1 - \theta)p_i$ (which will, of course, typically be less than p). Imagine that the seller has N goods, and that the minimum he needs to keep himself going is Np^* (often, though not always, $p^* < p$). If the seller has n friends, then the price he must set for people who are not in F will be

$$\left\{Np^* - n\theta p - (1 - \theta) \sum p_i\right\} /(N - n).$$

The more friends he has, and the stronger his altruistic response, the higher the price he will have to demand for those outside his circle of beneficiaries.

Now when times are hard and competition is severe, p may be close to p^*. A seller with a large number of friends will have to set a high price to non-friends, and may meet with no takers. Thus, instead of obtaining the Np^* he needs to keep going, he will only receive $n\theta p + (1 - \theta)\sum p_i$, which is strictly less than np, and will be too little when p^* is sufficiently close to p (more exactly: when $p^* > np/N$). His non-altruistic counterparts, by contrast, who set their prices at p are likely to be more successful. Hence there will be selection pressure against the expression of altruism in the trading context, and that selection will be more intense against the forms of altruism that exhibit greater intensity. So we can expect the intensity of altruism to diminish – yielding, in the ideal limit, a group of sellers who belong to *Homo economicus*.

This argument depends on several assumptions. First it supposes that, for any given part of the economy, there will be a sizeable collection of people who don't fall within the F-class of any seller; these people will have to pay higher prices, and (it is assumed) resent paying the very highest prices in hard economic times. Reflection on ties within affluent societies from the eighteenth century to the present makes the assumption appear reasonable: I may know a baker and you a brewer, others may know a butcher or a candlestick-maker, but, for most of us, most of the time, economic transactions will take place with relative strangers. Yet there are conceivable socioeconomic arrangements that would defeat the assumption. Suppose that suppliers intend only to sell to members of their F-classes ($n = N$). No trouble need arise if $n\theta p + (1 - \theta)\sum p_i > np^*$. But that inequality might be satisfied if life is relatively stable (p^* doesn't approach p, p_i doesn't get periodically depressed), or if limited mobility for the buyers doesn't set constraints on p. It is not hard to see how the economic arrangements of isolated villages or of feudalism might allow for this to occur.

https://doi.org/10.1017/S0266267110000167 Published online by Cambridge University Press

Second, the argument also supposes that the intensity of the altruistic response varies with pairs of sellers and potential friends, but, for each such pair, not across the varying economic circumstances. An alternative way to introduce altruism into economic life would be to tune the intensity of the altruistic response to the distance between p^* and p. As p^* approaches p, the seller is subjected to greater constraints, and might scale back the intensity (or the scope) of his altruism. So, as times get harder, θ increases, with the result that $n\theta p + (1 - \theta)\sum p_i$ is maintained at a constant distance from Np^*:

$$\theta = (Np^* - \sum p_i - k)/(np - \sum p_i) \text{ where } k \text{ is a constant.}$$

If the altruism of the *buyers* is expressed in their willingness to give a little more to accommodate the hardships for the suppliers (i.e. to increase the p_i), that would have a mitigating effect on the tendency for altruism to diminish in bad times.

It is now possible to envisage a different form of economic arrangement that would allow for the expression of altruism within the trading zone. In its original form, the Smithian argument doesn't address the possibility that sellers do not differentiate the *F*s from the non-*F*s: they set one price for all. That price is adjusted as the times are demanding or severe. If p^* were to fall (that is, the profits obtained from full sales would rise relative to the fulfilment of subsistence needs), the sellers would intensify their altruistic response (diminish θ). They would sell to each potential buyer at the price $\theta p + (1 - \theta)p_i$, thus reflecting facets of the situations of those potential buyers ('from each according to his ability').

Does an analogue of the Smithian argument apply here? Effectively, sellers are altruistically returning to their buyers a part of the excess profit they would otherwise have received. It is plausible to think that a seller with a more limited altruistic tendency would reap greater profits, that these could be invested in streamlining the production process, enabling greater market share in the future, and that more altruistic suppliers would thus be driven out of the market. Yet there are possibilities for regulative institutions that would tell against any such strategy – that would initially compel behaviour in accordance with more intense altruism profiles and that might ultimately lead to the widespread inculcation of those more intense profiles (properly brought up sellers just wouldn't think of hedging their altruism – that would be shameful or blameworthy).[18] Similarly, one can imagine ways in which potential

[18] The possibility of maintaining a system of commerce based on socially shaped altruism can be viewed as a central question in Shakespeare's *Merchant of Venice*. I attempt a reading of this sort in an essay currently in progress.

https://doi.org/10.1017/S0266267110000167 Published online by Cambridge University Press

buyers might be publicly assessed and assigned to classes that allowed smooth adjustment of prices in accordance with the values they assigned to p_i.[19]

It would be reasonable to object, however, that conditions of trade that encouraged, or even required, altruistic responses would be vulnerable to disruption. The coordination of those altruistic responses to changing circumstances and to the perceived needs of others appears more delicate than the simple mechanisms Smith envisages, based on his exclusion of altruism from commercial life. Yet there are also reasons to worry about the two-sector society that Smithian capitalism bequeaths to us. Does the attempt to demarcate spheres, with altruism having considerable scope in one and negligible scope in the other, distort our psychological and social lives? Is it psychologically possible for people to sustain the altruism profile that Smith takes as appropriate in the 'public' (market) sphere, together with the sort of profile he (and other moralists with very different meta-ethical views) would take to be apt in the 'private' contexts? In particular, does the Smithian division erode the possibilities of various forms of higher-order altruism, thus diminishing the values to which human beings can aspire? A further version of the 'Adam Smith Problem' provides reasons for taking these issues seriously.

In a famous passage late in WN, Smith considers the educational needs of workers in societies with developed division of labour, and comes dangerously close to conceding the human costs of the mechanisms he has typically commended.[20] This is only one of several occasions on which Smith recognizes, in WN itself, the possibility of a non-economic standard against which the arrangements of the economic sphere could be judged. TMS is more consistently forthright. There, even in the famous passage in which Smith concludes (optimistically) that the fundamental goods of human life are inevitably distributed relatively equally ('as by an invisible hand'),[21] he emphasizes the importance to us of tranquillity. Throughout TMS, this genuine good is contrasted with the official value of WN – the betterment of our (economic) condition – and sometimes linked to the exercise of human sympathy. Thus, in a passage in which he is considering someone who obtains economic success,

[19] Each of us might carry a passbook that assigned us to an economic category; we would show the book to sellers, and they would give the price appropriate to the category.

[20] See Smith 2001: 840, 846, and compare 9–10. In his *Economic and Philosophic Manuscripts of 1844*, Marx fastens on precisely this point in formulating his fourfold thesis of the alienation of labour. Although the vocabulary changes, the same response to the dehumanizing power of capitalism pervades Volume 1 of *Capital*.

[21] TMS IV-i-10. Smith 1984: 184–185.

Smith writes:

> If the chief part of human happiness arises from the consciousness of being beloved, as I believe it does, these sudden changes of fortune seldom contribute much to happiness. (TMS I-ii-v-1; Smith 1984: 41.)[22]

We might amend the slightly narcissistic formulation, to propose that continued relationships in which people respond to the wants of others and value mutual accommodation (recall the discussion of higher-order altruism in Section 3) are central to the human good. *If* the elimination of altruistic responses from the economic sphere (and possibly from other parts of public life, as well) damages our capacities for altruistic responses in other contexts – perhaps through the dehumanization Smith glimpses (and from which he averts his eyes) late in WN, perhaps because the boundary between the altruism-free zone and our personal lives is porous or indefinite, perhaps because it is psychologically hard to maintain altruism profiles that combine rich enjoyment of mutual relationships with the attitudes of a competitive marketplace – then the alleged delicacy of economic arrangements that take human altruism seriously may be outweighed, on the non-economic scale of value that really matters to us, by the damage inflicted on our pursuit of what is central to our happiness.

There is little doubt that our tendencies to altruism can be shaped by social conditions. During the known history of ethical practice, it is very clear that altruistic tendencies have been modified to widen the *potential* scope of altruism – the xenophobic writing-off of those who live next door gives way to successive stages at which ever more people are seen as possible beneficiaries. This means that any analysis of social possibilities ought to go beyond simple ideas that we have a fixed egoistic nature or a fixed altruistic profile of a very particular kind. Yet it would be equally futile to insist that our character in these respects is indefinitely plastic, that for any given altruistic profile, we can devise a social environment that will inculcate it. Consequently, some appealing ideas for social arrangements, including economic arrangements that introduce altruism into our commercial life, may be debarred by developmental constraints. By the same token, however, the two-sector society that would accord with Smithian ideals – tranquillity and altruism flourish in the private sphere, while the absence of altruism in commercial life promotes economic efficiency – may also rest on a type of altruism profile that is impossible for us.

Pursuing these questions in terms of the approach to altruism I have outlined enables us to appreciate two significant, and neglected,

[22] Smith goes on to recommend that one ascend 'more gradually to greatness', but his proposal about 'the chief part of human happiness' commits him to the conclusion that our relations with others are worth more than economic gains, however achieved.

projects. One is to attempt to understand the psychological possibilities that constrain the social systems we might devise: what kinds of altruism are available to us? The second is to expand the styles of economic analysis: what sorts of arrangements work best (in terms of economic and non-economic values) for (realistically) altruistic agents? Here I have only thought in a highly preliminary way about one tiny sub-case of the second issue (it would be overstating even to say that I have scratched the surface). It is, however, an interesting irony that the overall position of the great thinker who is often seen as the inventor of *Homo economicus* should provide us with reasons for transcending the type of analysis we derive from him.

5. ROUSSEAU'S HOPE: THE IDEAL OF SOCIAL SOLIDARITY[23]

I close with a second application of the approach to altruism offered here. If you think of the political state as legitimized through some social contract, presumably tacit, then the simplest versions of the story, those offered by Hobbes and Locke for example, see the bargain as one in which you trade certain possibilities for something of greater importance to you (security for Hobbes, an impartial order in the administration of punishment for Locke). Rousseau stands in an ambiguous relation to this tradition, because he seems to be offering a solution to a more ambitious problem than anything Hobbes and Locke attempt to solve. That problem is stated as if it were an exercise in geometry:

> Find a form of association that defends and protects the members of a society and their goods, with the entire common power, and through which each person, uniting himself with all others, nonetheless obeys only himself, and remains as free as he was before. (Rousseau 1987: 148)[24]

Hobbes and Locke would be quite happy with the statement of the problem until Rousseau adds the final two clauses, demanding that the citizen obey only himself and remain as free as he was before. They would be puzzled by the thought that anyone can get the goods – security, orderly justice – without giving up something, without acknowledging the authority of another (or others) and without forfeiting part of his freedom. Yet Rousseau thinks that his version of the social contract is

[23] The ideas of this section developed from an occasion on which I heard an illuminating presentation on Rousseau by John Collins, and they have benefited from subsequent discussions with Collins. As we both learned later, a similar way of reading Rousseau had been offered much earlier by W. D. Runciman and Amartya Sen (1965) (in an article that seems to have been unjustly neglected.)

[24] I shall henceforth refer to this edition as R. I have amended the translations slightly, better to capture the sense of the French original.

https://doi.org/10.1017/S0266267110000167 Published online by Cambridge University Press

a solution to just this ambitious problem, even though the citizens who are party to the contract must alienate themselves and their rights to the entire community ('the complete alienation of his rights to the whole community'; R: 148). How can this be?

Rousseau offers the perplexing answer that the process of alienation is to be understood as subordination to the General Will, and that this subordination is to be conceived as a gain in freedom. Compounding the apparently paradoxical character of this reply, he suggests, in a passage that has excited much commentary, that the community must have the power to compel those who refuse to conform to the General Will, and that this amounts to forcing them to be free:

> In order that the social contract should not be an empty formula, it must tacitly include this commitment, which is alone capable of giving force to the others, to the effect that anyone who refuses to obey the General Will is constrained to do so by all members of the society: which means simply that he is thereby forced to be free . . . (R: 150).

I suggest that Rousseau is offering an ideal of social solidarity, that he regards this ideal as importantly connected with human freedom and with human equality.[25] I will use my treatment of altruism to explicate this ideal, which is, I suggest, important to our conceptions of possible forms of society.

Rousseau supposes that his problem only arises in a particular type of circumstance, and that there are preconditions on the kinds of groups that can solve it. The pressure to form a society stems from an environment in which individuals can no longer meet their needs by acting independently.[26] I suggest a different characterization: there are pressures for novel forms of social formation (in addition to those that already exist) when there are recognizable opportunities for optional games[27] in which the payoffs for cooperative interactions are greater than the expected payoffs for solo activity. The important instances are cases in which the payoffs from interactions in which others fail to cooperate are

[25] It is no accident that the Revolutionaries, influenced by Rousseau, adopted the slogan 'Liberté, Egalité, Fraternité'. Perhaps Anglo-Saxon political theory has viewed liberty and equality as in tension with one another, not so much because of its preference for negative rather than positive freedom, but through the neglect of Rousseau's mediating idea of *Fraternité* – or solidarity, as I would put it.

[26] As briefly noted in Section 3, this is the basic condition in which I take primitive dispositions to altruism to have evolved in our evolutionary ancestors.

[27] An optional game is one in which agents can choose whether and with whom to interact; if they interact, they play a game defined by a playoff matrix; if they do not interact (if they opt out), they act in a way that delivers some fixed value. For discussion, see Kitcher (1993).

worse than what the cooperating person would have obtained by acting alone. So, for example, there might be repeated opportunities to play an optional PD

	C	D
C	<9, 9>	<0, 10>
D	<10, 0>	<1, 1>

Interact
Solo ⟶ 5

In Rousseau's original version, a decision to interact with others is potentially life-saving; in my generalization, it is potentially life-enhancing.

According to Rousseau, there are preconditions on any group that can form a social contract: before it can decide on its socio-political arrangements it must first constitute itself as a group. I interpret him as requiring that, whether the games that arise for the potential group are optional or compulsory, there is a shared conception of the collectively best outcomes, and a shared commitment to bringing about the outcome viewed as collectively best. We start from a situation in which we can recognize a future sequence of interactions that might offer outcomes that are better for all of us than a future in which we acted independently. To subordinate ourselves to the General Will is to prefer, in each instance, the outcome in the potential interaction that we all agree on as collectively best.

The values represented in the optional PD above are the '*private* wills' of the individuals involved. But once these people have entered into the contract, subordinating themselves to the General Will, they undergo a psychological shift, coming to prefer the outcome that is collectively best. We can think of this in terms of the approach to psychological altruism I have developed, supposing (for simplicity) that both parties are golden-rule altruists ($\theta = 0.5$). (It would also be possible to consider Rousseau's account for agents with different preference-response functions, including those that accord with the Smithian division of society into two sectors.) The pertinent values for each are thus:

	C	D
C	9	5
D	5	1

Interact
Solo ⟶ 5

Those who subordinate themselves to the Social Contract commit themselves to a particular sort of altruistic response to others. They see

https://doi.org/10.1017/S0266267110000167 Published online by Cambridge University Press

the other members of the group (those who have been recognized as sharing a common conception of the collective good and a commitment to promoting it) as potential partners in future interactions, so that the values ascribed to outcomes of those interactions are modified in such a way that the collectively best outcome always receives the highest measure. If *P* is the set of partners in the contract, then my participation in the contract commits me to an altruism profile that, on any occasion of future interaction, optional or compulsory, will ascribe highest value to the outcome that all members of *P* view as the collectively best option. The adoption of that altruism profile consists in my social solidarity with the members of *P*.

It is now possible to understand the notorious idea of 'forcing someone to be free'. Suppose that an occasion of interaction arises, and I choose an action other than the one that would have led to the outcome viewed as the collective good. That would occur, Rousseau quite plausibly thinks, because I have chosen in accordance with my private will; in other words, I have failed to make the modification of my solitary preferences to which I originally committed myself. If my original entry into the contract was reflective and uncoerced, then I'm lapsing from the attitude I wanted myself to have. One might even say that I have taken on an attitude I wanted to reject. So an intervention to bring me back on track can be viewed as freeing me from obstacles to what I really want – so that I am 'forced to be free'.

My earlier discussion of higher-order altruism suggests a further connection between solidarity and freedom. If the continued participation in interactions that engender the collective good gives rise to the ascription of added value precisely because the outcomes have been produced through joint activity – mutual recognition and accommodation – then the social contract and the altruism profile it requires can be viewed as giving people access to worthwhile ends they could not otherwise have achieved. The framework of laws instituted after the contract makes me free in holding me to my original purpose (as in the discussion of the last paragraph) but it also promotes that joint activity with others that I come to identify as having a value independent of the specific rewards we reap. It contributes a distinctive form of negative freedom and an intelligible form of positive freedom as well.

As Rousseau recognizes, there is also an important link between solidarity and equality. Although he allows for division of labour, he cautions against allowing wide divergences in wealth:

> If you want to give stability to the state, make the extremes of wealth and poverty as close as possible; don't allow for people who live in opulence or for those who are indigent. (R: 170, note 12.)

https://doi.org/10.1017/S0266267110000167 Published online by Cambridge University Press

The danger envisaged here is apparent if you think about games in which there is an asymmetry in the payoffs. Consider, for example:

	C	D
C	<5, 25>	<0, 26>
D	<10, 0>	<2, 2>

If you imagine two people (or two groups) playing this game just once, it is easy to envisage them agreeing that the outcome <C,C> is collectively optimal. If the interaction is repeated several times, however, it is possible to appreciate a rival account of the common good. Suppose that the column player has no way to transfer resources to the row player. Then, if one thinks of aiming at the common good as always playing C, it is evident that the long-run distribution will be highly unequal. That strains the thought that this is the *common* good, inviting an alternate conception that allows the row player to play D with some agreed-on frequency. (Suppose the players play the game 60 times. If their notion of the common good agrees that column should always play C and row should play C exactly one third of the time, then the resultant distribution will be <500, 500>. On the other hand, if the conception of the common good requires always playing C, that distribution will be <300, 1500>, a significantly larger total amount [1800 as opposed to 1000] but very unequally distributed.) In general, large inequalities are likely to introduce asymmetries that threaten a conception of the collective good which always insists on reaching the outcome with the largest aggregate value.[28]

Here's a way to formulate Rousseau's ideas as recommendations for social arrangements. When people find themselves in conditions where there is a potential sequence of future interactions (which we can conceive as optional games from the standpoint of the present, although, once the option to engage in some of them has been chosen, others might then become compulsory), they may recognize the possibility that adopting certain altruism profiles would lead to outcomes beneficial for all. They should then want to realize the altruism profiles in question,

[28] I think this possibility lies behind Rousseau's cryptic remark about the continued existence of the General Will in the majority (R: 206). Imagine that one starts out from the assumption that the interaction represented in a game with highly asymmetric payoffs will occur once (or only a very few times). So members of the group agree that playing C reaches the outcome that is collectively best. As time passes, however, the number of opportunities for playing the game increases. Those who constantly receive the lower payoffs may then easily modify their conception of the common good, while others do not revise it. If the fortunate are in the majority, they may reasonably be seen as insensitive to the *common* good. Hence the General Will is no longer 'in the majority'. This line of thought can be extended to make sense of Rousseau's apparently perplexing claim about the inerrancy of the General Will (R: 155–156).

that is to develop social solidarity with one another. This will, of course, presuppose that all parties share a conception of the common good, and that all are prepared sincerely to commit themselves to realizing it. As indices of shared commitment, they can introduce institutions that direct future behaviour, and they should regard the institutions in questions as enabling, and not burdensome. As their interactions unfold, they may hope that institutional constraints become ever less necessary, and that the appropriate altruism profiles become sustained by emotional mechanisms, felt ties to the others involved.[29] Realizing social solidarity generates important kinds of freedom; maintaining social solidarity requires guarding against large inequalities.

It is worth asking whether the ideal of social solidarity is realizable in contemporary societies. There are several obvious difficulties. First, as Rousseau clearly recognizes, introducing social solidarity at different levels, within nested groups for example, is dangerous. For there can arise conditions under which the smaller group can aim at a collective benefit greater than that available to the more inclusive group, leading to the serious possibility that the broader social solidarity will be undermined (R: 167–168). Second, in any society in which interactions occur among members who are never, or only rarely, in direct contact with one another, it is eminently possible that any sense of a shared common good will wither, or that, at very least, the parties will have legitimate doubts about whether that sense exists. The altruism profiles that would enable smooth achievement of the common good may prove hard to sustain. The political ideal Rousseau offers may thus require self-ruling groups of quite limited size, and, for the reasons he gives, partial associations founded on a narrower social solidarity may be quite injurious to the social unit of which they are a part.

As with the discussion of Smith's legacy, my aim here has simply been to indicate how the notion of psychological altruism, understood in the way I have outlined, can be used to articulate normative claims about social organization and to prepare the way for empirical investigation of the conditions that would satisfy the norms. I draw a similar conclusion in this case too: political theory should explore the consequences of different achievable forms of human altruism, and psychology should inform us as to what parts of altruism space we may ever hope to inhabit.

[29] Even when the relevant sympathies are not initially in place, it is possible that they should develop. It is possible to regard Rousseau's educational programme, offered in *Émile*, as directed in part towards increasing the receptivity of citizens to the valuable emotions. (I am indebted to Fred Neuhouser for discussions that have brought home to me how important *Émile* is to Rousseau's social and political concerns.)

REFERENCES

De Waal, F. 1996. *Good Natured*. Cambridge, MA: Harvard University Press.
De Waal, F. 2007. *Primates and Philosophers*. Princeton, NJ: Princeton University Press.
Fehr, E. and Fischbacher, U. 2005. Human altruism – proximate patterns and evolutionary origins. *Analyse and Kritik* 27: 6–47.
Feinberg, J. 1975. Psychological egoism. In *Reason and Responsibility*, ed. J. Feinberg, 501–512. Belmont: Wadsworth.
Goodall, J. 1988. *The Chimpanzees of Gombe*. Cambridge, MA: Harvard University Press.
Kitcher, P. S. 1993. The evolution of human altruism. *Journal of Philosophy* 90: 497–516.
Kitcher, P. S. forthcoming. *The Ethical Project*. Cambridge, MA: Harvard University Press.
Rousseau, J-J. 1978. *The Social Contract*. In *The Basic Political Writings*, 141–219. Indianapolis: Hackett Publishing Co.
Runciman, W. D. and A. Sen 1965. Games, justice, and the general will. *Mind* 74: 554–562.
Schelling, T. 1984. *Choice and Consequence*. Cambridge, MA: Harvard University Press.
Smith, A. 1984. *Theory of Moral Sentiments*. Indianapolis: Liberty Fund.
Smith, A. 2001. *The Wealth of Nations*. New York: The Modern Library.