**UNIVERSALIZABILITY FOR COLLECTIVE RATIONAL AGENTS:**
**A CRITIQUE OF AGENT-RELATIVISM**

Kantians argue that any sound theory of practical reason must be universalizable. Their opponents argue that insofar as universalizability is hedged enough to be defensible it is an "empty formalism." The critic presents the Kantian with a dilemma. They argue that if the only notion of a contradiction in play in the categorical imperative is simply that of logical one (as opposed to some sort of practical or teleological contradiction)[1] then the categorical imperative it too anemic to have interesting consequences. If, on the other hand, the categorical imperative employs a more robust conception of contradictions then the critics argue that the categorical imperative, so understood, is not supported by Kant's arguments. In discussing the first horn of this dilemma (concerning the implications of universalizability read in the more modest logical contradiction way) Kantians and their opponents have both focused on the possibility of universal compliance with a proposed theory of practical reason by all *individual* agents. However, it is plausible to suppose that not all possible rational agents are individuals. For it is also reasonable to suppose that *collective* rational agents are also possible. After all, we speak of nation-states, lobbying groups, churches, corporations, universities, trade unions and other groups as performing actions for reasons and as proper objects of both moral and legal responsibility. Nor is there any obvious reason not to take this talk at face value.[2]

I shall not here defend the supposition that collective rational agents actually are possible nor shall I defend Kantian universalizability read in the more modest "logical contradiction" way. Rather, my aim is simply to see what follows from these two suppositions. For perhaps universalizability is more than an empty formalism in virtue of the possibility of collective rational agents. Kantians have not really explored this

possibility, in spite of the fact that some of them emphasize the importance of learning about individual agents by way of analogy with collective agents. I argue that the possibility of such agents combined with universalizability entails the striking result that no fully agent-relative theory of practical reason could be correct. Since the refutation of agent-relativism is a rather surprising and strong result, this undermines the worry that the logical contradiction interpretation is too anemic to have any interesting implications. The Kantian can embrace the second horn of the alleged dilemma – the more modest interpretation of universalizability actually does have at least some non-trivial consequences. The refutation of ethical egoism, to mention just one historically influential fully agent-relative view is hardly trivial.

Before developing this argument, it is important to be clear about what is meant by 'agent-relative' and 'agent-neutral'. For present purposes, a principle's being agent-relative consists in its making an ineliminable and non-trivial pronominal back-reference to the agent.[3] For example, a theory whose most fundamental principle is that an agent must maximize aggregate welfare involves no pronominal back-reference and hence is an agent-neutral one. Whereas a theory whose only principle held that an agent must promote *her* welfare, or that an agent must promote the welfare of *her* friends, would contain pronominal back-reference and hence would be agent-relative. Egoism is a paradigmatic case of an agent-relative theory.[4] The back-reference must not be eliminable; it must do some real work in the theory. Nor can the back-reference be trivial. Here I primarily have in mind the way in which all principles of practical reason must at least implicitly relativize reasons to the possible actions of the agent. For not to relativize reasons in this way would mean that our theory might have the consequence

that there is reason for an agent to perform an action which is in no way a possible action of hers. This, however, would flout the plausible assumption that in some sense 'ought' implies 'can'. So if the agent-relative/agent-neutral distinction is going to be at all interesting we must define agent-relativity in terms of non-trivial pronominal back-reference. Otherwise *all* reasons will be agent-relative ones on the trivial grounds that reasons must respect the dictum that 'ought' implies 'can'.

Call the doctrine that the correct theory of practical reason is fully agent-relative, in the sense that all its fundamental principles are agent-relative, "agent-relativism." Agent-relativism, as I shall understand it here is a doctrine about "insistent reasons."[5] Insistent reasons differ from non-insistent reasons in that if one recognizes an insistent reason as a reason outweighing the other reasons in play but nonetheless does not act upon it, one is thereby irrational. Non-insistent reasons are such that one can register them as reasons and fail to act as they recommend without thereby being irrational. It is controversial whether there could be any non-insistent reasons.[6] On the other hand, there are *prima facie* plausible instances of such reasons. For example, "agent-centered prerogatives" to give one's interest disproportionate weight are perhaps best seen as non-insistent reasons.[7] Supererogation may presuppose the possibility of such reasons.[8] Fortunately, I need not settle the dispute over whether there are any non-insistent reasons. The only point is that *if* there are any, they are outside the scope of the present argument.

The argument developed here against agent-relativism hinges on the way in which the action of a collective can be constituted by the action of a privileged individual within it under certain conditions. Properly understood this is a very plausible idea. For example, if a President of a constitutional democracy declares war in the name of his

country against another country in such a way that his declaration is recognized as legitimate by his nation's constitution then his country has declared war and that declaration is (at least partly) constituted by the President's action. Of course, not all (or even most) actions of individuals with special roles in a collective will constitute an action by the collective. When the President brushes his teeth his country does not brush its teeth! The argument developed here requires only that *sometimes*, in certain kinds of cases, the action of a collective depends on the action of a particular individual who occupies an important and constitutionally recognized role in the collective. Moreover, I argue that in some such cases it will be impossible for both the individual and the collective to act as an agent-relative theory like egoism would demand. For agent-relativity can give the individual and a collective of which she is a part conflicting ends. This, however, is just to say that agent-relative theories flout universalizability in the relevant sense. Here the possibility of collectives is crucial. For in cases involving only individuals, the fact that an agent-relative theory provides us with conflicting ends does *not* show that it is impossible for both of us to act as the theory demands. You should do your best to win our chess game, and so should I. There is no impossibility here because we can both do our best to win even though we cannot both win. Admittedly, we must hold that our reasons are to do our best to win, rather than simply to win, but this is a plausible assumption (defended in more detail below). Indeed, the move to trying or "doing one's best" is itself very Kantian insofar as Kant's test applies to maxims, where maxims indicate one's commitment to try to promote (or respect) some end (though again my aim here is the exploration of broadly Kantian ideas rather than Kant exegesis). It is only when one agent is a proper part of another agent that we begin to find impossibilities

of joint compliance even at the level of trying one's best, and hence a problem with Kantian universalizability. For in such cases, it may be that the collective as much as *doing its best* to achieve the end it has been given by the agent-relative theory in question is constituted by the action of an individual member of that collective's *not doing its best* to achieve the end that it has been given.

**I. The Argument From Compliance Universalizability.**

At least one interesting version of Kantian Universalizability is constituted by the thesis that a theory of practical reason is sound only if every possible set of rational agents could universally accept and follow it. It is not hard to find this idea at work in Kantian moral and political philosophy. Kant famously held that reason demands that one always "act as if the maxim of your action were to become by your will a universal law of nature."[9] Any maxim which fails this test is impermissible according to Kant. This immediately raises the question of how Kant conceived maxims. A person's action is a function of the maxim she has adopted, where the maxim incorporates what the agent takes to be the good reason(s) for performing actions of that type.[10] So understood, a person's maxims include her most fundamental practical principles as well as various derivative ones. For a fundamental principle of practical reason would simply be an articulation of what considerations one takes to be reason-giving in their own right, and given our characterization this would count as a maxim, albeit a rather abstract one. Kant discusses such high-level maxims, and the structure of more fundamental maxims to more derivative ones in his *Religion Within the Limits of Reason Alone*.[11]

On a standard reading of the Universal Law of Nature formulation, it entails that an agent must consider whether it would be possible for the laws of human nature to

include a universal law corresponding to her maxim, keeping the rest of the features of the world as fixed as possible.[12] A maxim's being a universal law of nature, in this sense, requires that every rational agent accepts and always follows the maxim. So in determining whether a given maxim could be made a universal law of nature, one must determine whether it would be possible for all rational agents to accept and follow that principle. Furthermore, it is quite plausible to suppose that a *fundamental* principle of practical reason's status as sound or unsound is a necessary truth. Indeed, this presumably is part of what distinguishes fundamental principles of practical reason from more derivative principles, as the derivation of those lower-level principles might well go via contingent premises. In which case, the mere *possibility* of a set of agents who could not will a fundamental principle as a universal law of nature must entail that the principle is unsound.[13] For otherwise we would get the implausible result that such a principle is sound in one possible world but unsound in another. Finally, it is relatively clear from Kant's discussion of the "Realm of Ends" formulation of the categorical imperative that it is really whole sets of principles, rather than individual ones, that must be tested for universalizability in this way.[14] So not only must each of our fundamental principles individually pass the universalizability test, they must also collectively pass it. We must, in other words, test entire theories of practical reason rather than atomistically testing each component individually. This entails the following Universalizability Thesis:

Kantian Universalizability: A theory of practical reason is sound only if every possible set of rational agents could universally accept and comply with its demands.

While Kantian Universalizability is in my view quite plausible, it is not a platitude. Kantian Universalizability (henceforth, "KU") requires that any fundamental principle be one with which we could all comply, *in the sense that there is a possible world in which*

*all of us together comply with it.* Those who believe in the possibility of strong moral dilemmas will perhaps balk at universalizability in the sense in which I am here invoking it. For if we combine universalizability so understood with a familiar agglomerativity principle, according to which "A ought to x" and "A ought to y" entails "A ought to x and y" then it would follow from KU that strong moral dilemmas are impossible. Unsurprisingly, the view that strong moral dilemmas are impossible is a deeply Kantian view, and if that follows from the assumption of KU, then I must take on that assumption as well. Again, I am not here going to try to defend the Kantian view or its implications. Rather, my aim is to take the Kantian view as given and explore its implications.

An even weaker universalizability requirement would be that, for each one of us taken individually, *there is a possible world in which he or she complies with it*; this would be compatible with there being no single possible world in which we all comply. This latter form of universalizability follows from a plausible interpretation of the dictum that 'ought' implies 'can'. For if there is no possible world in which I accept a standard, then there is no sense to be made of the thought that I ought to adopt it. Call this "Distributive Universalizability" (henceforth, "DU"). For example, if twenty agents each have decisive reason to drink some water, then DU could be satisfied if there were twenty different possible worlds, such that in each possible world one and *only* one of the twenty drinks some water. Whereas KU requires that there be a *single* possible world in which *all twenty* drink some water.

KU requires the possibility of universal acceptance and compliance, but the present argument requires only a universal compliance constraint. So I shall rely on the following constraint which KU entails:

Compliance Universalizability: A theory of practical reason is sound only if every possible set of rational agents could universally comply with its demands.

Compliance Universalizability (henceforth, "CU") presents a surprisingly serious problem for agent-relativism, given the possibility of collective agents. A defining feature of agent-relativist theories is that their principles all involve an ineliminable and non-trivial back-reference to the agent. This means that any two agents' reasons may pit them against one another. So far, this is no problem for the agent-relativist, for the fact that we have reason to behave antagonistically toward one another in no way undermines the possibility of our both so behaving. It may mean that if we both act on our reasons that each will be worse off than if we cooperated. However, the possibility of everyone's being individual rational leading to all being worse off is hardly news, given the familiarity of prisoner's dilemmas. In my view, prisoner's dilemmas in themselves do not undermine agent-relativism.[15] Rather, they reveal that given agent-relativism, if everyone accepts the correct view and acts rationally then all will do worse than if everyone acted irrationally or accepted an incorrect view.

It might seem that in competitive contexts that agent-relativist theories like egoism fall afoul of universalizability quite apart from the possibility of collective rational agents. Suppose, for example, that we are in competition for a scarce resource. In a suitably specified case, it might seem that given egoism what I have reason to do is acquire and consume all of the resource, and what you have reason to do is acquire and consume all of the very same resource. If, however, that is an accurate description of our reasons, then it seems that egoism flouts universalizability. For by hypothesis, the resource is not one which we can both fully consume, in which case it will be impossible for both of us to do what there is most reason for us to do (namely, consume all of the

resource). However, the case is actually underdescribed and this apparent problem for universalizability will vanish on any plausible way of filling out the description. First, there is the question of who (if anyone) will succeed if both agents try to get the resource. Since by hypothesis the resource is useless unless one has a monopoly on all of it, there is no possibility of both agents' succeeding. So either one of them will succeed or neither of them will succeed (perhaps their competition will destroy the resource). Suppose first that only one of them (the more powerful one) would succeed if both tried to get the resource. Then unless there are further relevant facts then the more powerful one should get and consume the resource and the other should acquiesce (as her efforts would be futile). Since this is clearly a logical possibility, universalizability is respected on this specification of the case. Of course, it might be that acquiescence by the weaker agent is a bad idea from her point of view because it will give the impression that she is easily pushed around, or for some other reason. In that case, though, the more powerful agent has most reason to get the resource and the other agent has reason to try to get it for herself even though she will fail. Since it is clearly logically possible in such a case for the more powerful agent to get the resource and the weaker agent to try but fail to get it we once again have no problems with universalizability. Alternatively, let us suppose that if both agents try then neither will succeed but if only one agent tries then that agent (whichever one it is) will succeed. Again, unless other facts are given then each of them has most reason to get the resource if and only if the other agent does not try to get it, since there is no obvious point in trying to get the resource if one's efforts will fail. In that case, though, universalizability is respected in virtue of two possible worlds: (a) the world in which A gets the resource and B acquiesces and (b) the world in which B gets

the resource and A acquiesces. Each of these outcomes is logically possible in such a case, and in each of these outcomes both agents have acted as they ought to have acted, all things considered. So the availability of either of these worlds would be sufficient to show that universalizability is not flouted in virtue of such cases. Again, though, it might be that one of the agents should try even when her efforts are doomed to fail. Again, it might still make sense to try because of the impact of acquiescence on one's reputation, for example. In that case, though, each agent has most reason to try to get the resource even if the other agent also tries to get it.[16] It would be a mischaracterization of their reasons in to claim that each has reason actually to get the resource even if the other agent tries to get it. For by hypothesis if each agent tries then neither will succeed. So to claim that each agent ought to get the resource even when the other agent tries is to run afoul of the very plausible thesis that 'ought' implies 'can'. Again, since it is clearly logically possible for both agents to try to get the resource in such a case we still have no problem with universalizability. Similar moves are obviously available in the case of resources that do not require a monopoly for their utility. Those cases just add the possibility that sometimes each agent should get as much of the resource for herself as possible given the other agent's efforts, but that just makes the satisfaction of universalizability easier. It seems that none of these kinds of cases really suggests that agent-relative theories like egoism cannot satisfy universalizability.

When we consider collective agents, though, certain cases of antagonism *do* provide the resources with which it can be shown that agent-relativism flouts universalizability. As I shall explain in more detail below, the solution invoked above for the cases of individual competition is unavailable in certain cases involving collective

agents. In the relevant cases involving collectives, the appeal to reasons to try to x rather than reasons to x is of no use. For when we consider the possibility collective agents, the possibility of antagonism that agent-relativism brings with it might be between a collective agent and some individual(s) who (partially) constitute the collective. My primary contention shall be that the possibility of antagonism between different agents that inevitably accompanies agent-relativism is sure to be compatible with universalizability *only if the agents who can be at odds with one another do not stand in a part/whole relation*. The possibility of collective agents just is the possibility of such part-whole relationships, though. Given the possibility of collective rational agents, the following sort of scenario seems possible for any form of agent-relativism:

(1) A collective agent, C, has decisive agent-relative reason to X at time t. (where a reason is decisive just in case the agent ought, *all things considered*, to act as the reason recommends)

(2) An individual, I, who is a member of C, has decisive agent-relative reason to Y at time t.

(3) If I Ys at time t, then given the relevant facts, it would follow that C does not X at time t.

(4) If C Xs at time t, then given the relevant facts, it would follow that I does not Y at time t.

Such cases seem possible for any agent-relativist theory precisely because for certain kinds of collective rational agents, what the collective counts as doing is often a direct function of the action(s) of an individual who has a privileged position within the collective. It is for this reason that cases in which (3) and (4) are true can be generated against the backdrop of (1) and (2). In such cases, CU is violated. For given this description of the case, it will be impossible for both agents to comply with the agent-relative theory in question. Insofar as individual I acts as the theory demands, it follows

from the case that collective C does not, and vice-versa. There just is no possible world in which, given their circumstances, I and C comply with the theory in question - which is to say that the principle fails to satisfy CU. The structure of the argument is as follows. First, for each proposed agent-relative theory, assume for reductio that the theory is sound. Then construct a case of the sort defined by (1)-(4) applying to the theory. This shows that the theory is inconsistent with CU, for (1)-(4) guarantee that for at least *some* possible set of rational agents it is impossible for them all to comply with the theory. Given CU, the theory is not sound after all, completing our reductio.

I have so far given the argument at a high level of abstraction to emphasize its generality. However, a specific instance of the strategy is crucial to appreciating its force. Suppose that nation-states are collective rational agents.[17] Let us suppose, in particular, that the U.S. is a collective rational agent. Admittedly, the mere supposition that collective rational agents are possible does not entail that nation-states are collective rational agents. However, it will help to have a concrete example, even though any particular example one chooses will inevitably be somewhat controversial. Hopefully, it will be clear how the argument would go through just as well even if one chose any of a number of different examples. In allowing that the individual members of a collective stand in a part/whole relation to the collective to which they belong, I am assuming that the individual remains an individual. The assumption is *not*, in other words, that the individual members of a collective must abandon their individuality insofar as they really are a member of the collective. To assume that collective rational agents are possible in the sense operative here, one must allow that the individual members of a collective can persist as separate individuals with reasons of their own in spite of their membership in a

collective, and without splitting the individual into two (or more) literally distinct agents ("the individual qua individual" and "the individual qua member of the collective"). Not all views which hold that collective or group persons are in some sense possible would allow that they are possible in this sense, but a many would.[18] Again, however, I cannot here defend the assumption that collective rational agents in this sense are possible. The point of the preceding discussion is only to clarify what that assumption involves.

With the assumption that the U.S. is a collective rational agent in hand, let us suppose for reductio that rational egoism is correct. According to rational egoism, there is just one substantive axiom of practical reason, according to which each agent has reason to do whatever is in that agent's interest, and has decisive reason to do whatever is most in his or her interest. Finally, let us suppose that the facts are as follows. Given egoism, the U.S. has decisive reason to adopt a gas tax at the present time. For if such a measure is not passed, there will be an oil shortage and global warming, and this eventually will seriously set back the U.S. national interest. The President, however, has decisive reason, according to egoism, to veto the gas tax, because if he does not he will be very unpopular. Perhaps most Americans either disagree about the relevant empirical facts about warming and oil shortages. Or we could assume that warming and oil shortages will not have any serious negative impact until long after present generations are dead and that most folks are much more concerned about their own welfare than distant future generations. Further, suppose that the President cares a great deal about his popularity and that it is in his interest to maintain it. So it is in the President's interest, all things considered, to veto the gas tax, and it is in the national interest for the U.S. to pass it. So, given egoism, the U.S. should pass a gas tax and the President should veto it.

However, let us also suppose, as seems quite plausible, that if the President exercises his veto power at that time then the U.S. simply will not count as enacting a gas tax at that time. Nation-states are the sorts of collectives whose actions can sometimes be a strict function of the actions of some privileged individual in the collective, and this is a prime instance of this phenomenon. Further, if the U.S. does enact the legislation at the present time, it follows that the President did not exercise that veto power. Roughly, the idea is that the President's signing or vetoing of the bill would, given the context and background of constitutional rules, constitute the U.S.'s enacting or not enacting the legislation, and that there are no other actions available to any other agents at the time that would constitute the U.S.'s enacting or not enacting the legislation at that time.[19] The assumption is that for collective agents with rich institutional structures (like a nation-state), the performance of certain kinds of actions by the collective is constituted by certain individual(s) following the relevant institutional rules (signing the bill into law, e.g.).[20] Given this description of the case, it is strictly speaking impossible for everyone to comply with egoism. For if the President complies with egoism's demands, it follows from the description of the case that the U.S. does not comply with egoism's demands, and vice-versa. Egoism is not universalizable. [21]

I have deliberately been vague about how we should understand collective agents and their actions because the argument developed here should work on any of a number of different plausible conceptions of collective agents and their actions. To develop the argument as premised on some particular conception of collective rational agents would unnecessarily de-emphasize its generality. Moreover, any particular theory of collective agents and actions will inevitably be more controversial than the general idea that such

agents and actions are possible. Nonetheless, it is worth pausing to see how the preceding argument is compatible with a number of existing accounts of collective agents and actions. First, consider David Copp's account of collective actions as a species of secondary actions. A paradigm of a secondary action is when Jones buys a house in virtue of the action of someone whom Jones has invested with power of attorney. It is true that Jones bought the house even though Jones did not sign the deed himself. Plausibly, the actions of Jones's fiduciary constitute Jones's buying of the house. Copp argues that we should understand collective actions in a similar way. The constitution of a collective's action by the action of some privileged individual(s) can happen in at least two ways on Copp's account. First, the constitution of a collective's action might supervene on the "facts about the constitutional rules or laws, laws and bylaws of organized collectives."[22] Second, the constitution of a collective's action may supervene on facts "about the composition and dynamics of, or patterns of interpersonal relations within, given unorganized collectives."[23] As Copp notes, this second mechanism comes into play only when the group in question is not tightly organized with rules and conventions in the way that nation-states are organized. Copp's account is clearly very amenable to the argument developed here. For it is clear both from Copp's account itself and his deployment of it to various examples that the account is tailor-made to explain how the action of the President might, under suitable circumstances, constitute the action of the United States. Copp at a number of points discusses the case of one nation-state declaring war on another in virtue of the actions of an individual in a privileged position within the nation-state:

> For example, the country of Exemplar, a constitutional monarchy, declared war on Germany inn 1939. This action is attributable to Exemplar on the basis of the

> Prime Minister's, Mr. Dux's, action of issuing a formal proclomation…I contend that the one action 'constitutes' the other.[24]

In the gas tax example, my suggestion is simply that the President's vetoing the legislation at least partly constitutes the United States' refraining from adopting a gas tax while the President's signing the bill into law would partly constitute the United States' adoption of a gas tax. I further stipulate that the case is to be understood in such a way that only the President's action could at that point in time constitute the relevant collective actions; so long as the story is told in the right way this should also fit well with Copp's account. Indeed, the gas tax case is very similar to the kinds of cases discussed by Copp and his account provides a nice model of such cases.

Second, consider Margaret Gilbert's (rightly) influential account of social actions. On Gilbert's account, a group action through the decision of a special representative (like a President) is possible, though it depends on a background of joint belief quite generally by the members of the group that the representative in question is authorized to make such decisions for the group. Here is Gilbert:

> Often we ascribe an action to a group as a whole when most group members are not directly involved…All Russians did not share in the act of invading Czechoslovakia in the simple way in which you and I may share in the act of travelling together. Most Russians did not take part in the invasion. Many may not have even heard about it. This is even more obviously true for so-called covert operations…in order for us to feel comfortable with the idea that a certain group invaded Czechoslovakia, there surely must be a sense in which whoever organized the invasion, and whoever took part in it, was the authorized representative of the group as a whole. In order for this to be so, something like this must be true: members of the group jointly accept that certain decisions of a certain few are to count as our* decisions. Something like that often is true and it seems that in such situations, at least, we can reasonably allow that the group itself has made the decision or performed the action in question.[25]

Gilbert's account is in many ways similar to Copp's account. Both maintain that special subgroups can make decisions on behalf of the larger group insofar as there are rules or

conventions that authorized their doing so. Copp leaves the idea of rules and conventions implicit and intuitive whereas Gilbert offers a detailed theory of such conventions in terms of the joint acceptance of rules and beliefs. On her account, a group embraces a view insofar as most members have indicated their willingness to let the view stand as the view of the group and it is clear that she takes a similar line on rules understood as group decisions.[26] This account also seems compatible with the argument developed here. Gilbert's theory of joint acceptance is complicated and subtle, but we need not get into the details of the account here. For so long as we tell the story in the right way, it will be very plausible to suppose that most Americans have in Gilbert's sense indicated their willingness to let the Constitution stand as the group's decision about how to make other decisions. If necessary, we could just stipulate that all citizens competent to understand the question had explicitly considered the question and said in some official context that they accepted the rules of the Constitution as representing a group decision. We probably do not need to tell the story in such an extravagant way, though. Gilbert holds that in the right circumstances acquiescence can count as an indication of acceptance. A citizen's not objecting to the Constitution when it is in place and guiding practice and when she has the right to free speech might plausibly be taken as an instance of the sort of acquiescence Gilbert has in mind. Moreover, it is clear enough that Gilbert wants her account to handle these kinds of cases. For it is clear from her discussion of various examples (as in the preceding quotation) that she intends her account to allow for group actions in cases like the gas tax case.[27]

One interesting feature of Gilbert's account is the idea that "group membership is not 'normatively neutral'" (Gilbert, p. 415). She suggests that insofar as one sees oneself

as a member of a group that one must take oneself as having some reason to do one's fair share in promoting the group's joint aim(s).[28] This seems to suggest that being a member of a collective agent is incompatible with being an egoist. For membership in a collective agent requires one to recognize apparently non-egoistic reasons to do one's fair share in advancing the group's goals. This seems to make the appeal to the gas tax example and its kin otiose for purposes of refuting egoism. For if the existence of collective agents presupposes that at least some people have rejected egoism then *of course* egoism is not universalizable insofar as collectives are on the scene. This is an objection not to the soundness of the argument developed here but rather a worry that we do not need its elaborate machinery to see the problem collectives pose for universalizability. It would be strange if this were so since so far as I know nobody in the literature has appealed to these kinds of considerations to show that egoism is incompatible with univeralizability. If it really were that easy to refute egoism given the possibility of collective agents then it would be a little surprising that nobody has explicitly done so. Nonetheless, the worry is worth taking seriously.

The worry is reasonable, but unsound for two reasons. First, egoism is the view that there is reason for an agent to perform an action just insofar as the action promotes her welfare. This leaves open just how we should understand welfare. On some plausible accounts, an agent's welfare might be *partly* constituted by her goals (perhaps subject to some screening – blatantly self-destructive goals might not count, e.g.). If becoming a member of a collective itself involves adopting certain goals *qua* member of the collective then those considerations can provide perfectly respectable egoistic reasons after all. So a group of egoists can join together to form a collective where the collective

has some joint aim J. This will, let us suppose, entail that each member of the collective must also take him or herself to have reason to do his or her fair share in promoting J. However, this need not be seen as incompatible with egoism insofar as we allow that one's welfare is partly constituted by one's goals. For on this account, insofar as my joining a collective logically requires having the goal of doing my fair share to advance some joint goal G it will also be true that my joining the collective entails that my welfare is in part a function of my doing my fair share to advance J. So the mere existence of collectives is after all compatible with everyone's accepting egoism *if* we understand an agent's welfare at least partly in terms of her goals. One might reply that this makes the gas tax case incoherent, since it entails that the President will have reason to pass the gas tax after all. For by hypothesis the gas tax is in the national interest and on this sort of account the President's welfare is partly constituted by the national interest. However, this would undermine the coherence of the gas tax case only if being a member of a collective required that one took its reasons to trump any other reasons one might have. Gilbert does not argue for such a strong view and it would be very implausible in any event (I return to this issue in section II at some length). Surely one's membership in a collective does not require that one subordinates all of one's other concerns to the concerns of the collective. Human groups are in this way different from ant colonies and bee hives. So we can allow that the President has some reason to do his fair share in advancing the national interest but also allow that he also has most reason, all things considered, to advance his own welfare in other ways (e.g., by maintaining his public popularity). Unless we raise the bar for group membership absurdly high we will be able

to construct the sorts of cases on which the present argument depends simply by raising the stakes for the President high enough.

Second, recall that the argument developed here aims to undermine *all* purely agent-relative theories of practical reasons and not just egoism. So even if the machinery developed here were not necessary to refute egoism it might well be needed to refute *other* purely agent-relative views. For example, consider a close cousin of egoism which claims that there is reason for an agent to do something if either of two conditions is met: (a) it promotes the agent's welfare or (b) it advances one of the joint goals of a collective of which the agent is a member. This theory is fully agent-relative, since both (a) and (b) involve the relevant back-reference to the agent – it is only because it is *my* welfare or the goal of *my* group that I will have reason to do something. Clearly, though, an agent could accept these two principles and still belong to a collective in Gilbert's sense, since accepting (b) ensures that the agent does take the collective's aims as providing reasons for action. We could even characterize (b) in the explicitly moral terms of the agent's doing *her fair share* in advancing the group's goals and the principle would still be agent-relative. So the mere existence of collectives is not incompatible with the universalizability of this sort of agent-relative theory of practical reason. Furthermore, these principles are arguably more plausible and hence more worthy of arguing against than flat-out egoism since they make room for recognizably moral norms stemming from interpersonal relations on which groups are founded. The basic point is a simple one. We can agree with Gilbert that group membership is not normatively neutral but hold that the relevant norm(s) is (or are all) agent-relative. Indeed, an agent-relative construal of the relevant principle(s) has considerable intuitive plausibility. If those norms are agent-

relative, however, it follows that the existence of collectives is compatible with all of the members of the collective accepting an agent-relative (albeit non-egoistic) theory of practical reason. In which case, the mere existence of collectives is not enough to raise problems of universalizability for all agent-relative theories or even all prima facie plausible agent-relative theories. So the machinery developed here is necessary after all.

Admittedly, there may be interesting questions about how much weight reasons stemming from collective goals must have for an agent when they conflict with reasons of self-interest if the agent is to count as a member of a collective. However those questions are resolved it will still come out that a purely agent-relative theory is compatible with the existence of collective agents. Moreover, unless the reasons of the collective are given lexical priority over reasons of self-interest, this will also make it possible to construct cases like the gas tax case. All we have to do is raise the stakes (in terms of self-interest) for the President enough (suppose his life depends on it, or whatever) and the problem for universalizability. This is obviously analogous to the points made about a more expansive conception of welfare in terms of one's goals discussed above. We can understand one's reasons as a member of a collective as stemming either from a broad notion of welfare and the goals one adopts in joining a welfare or from a basic norm that claims one has reason to do one's fair share on behalf of one's own collectives. Either way, so long as the reasons in question are agent-relative, not the only reasons there are,[29] and not always overriding we can construct the relevant sorts of cases. I conclude, therefore that the fact (assuming it is a fact) that group membership is not "normatively neutral" does not show that the machinery developed here is unnecessary to argue against agent-relativism. Nor does it show that

the machinery is not sufficient, so long as we do not suppose that the reasons one must recognize to be a member of a collective must be taken as having lexical priority over one's other reasons. Again, I return to this point about the relative priority of group-based reasons and individual-based reasons in section II.

At the risk of belaboring the point, it is important to emphasize that present argument is not limited to egoism, but generalizes to any (non-inter-defining) agent-relative theory. For so long as a collective rational agent and its members must promote different ends, the relevant sort of antagonism can emerge. Consider, for example, the agent-relative view that one ought to promote the welfare of one's allies (a close cousin of C.D Broad's "self-referential altruism"). It is easy to see how the present argument could be extended to cover a theory which had this as its only axiom. For we need only to add to the case that it is in the interest of The President's allies (his friends in the oil industry, e.g.) for the U.S. not to impose a gas tax while it is in the interest of U.S.'s allies (other nation-states who would be harmed by global warming) for the U.S. to impose such a tax. To take a more controversial example, it has been alleged by some that Winston Churchill knew about the bombing of Coventry during World War II before it happened because of the code-breaking work of ULTRA but chose not to warn the people of Coventry to avoid letting the Germans know that their code had been broken.[30] This may well not be historically accurate, but it is at least possible. Assume for the sake of argument that this popular account is accurate. Let us suppose also suppose, for the sake of argument, that Churchill had personal friends in Coventry. Now consider the following agent-relative principle:

Each agent A ought to promote the welfare of A's allies.

It seems plausible to suppose in the Churchill case as we are imagining it this principle would recommend that Britain not warn the people of Coventry, assuming that it was very important to the war effort not to let the Germans know that their code had been broken at this point. For it is clear that Britain's allies had a very strong interest in maximizing their prospects of defeating Germany. On the other hand, that very same principle would seem to recommend that Churchill warn his friends in Coventry that they were about to be bombed, assuming this would increase their chances of survival. Since Churchill's leaking that information would constitute Britain's leaking the information, it seems that if Churchill complies with the agent-relative theory in question then Britain does not. So it is impossible for all agents concerned to comply with the theory, again flouting CU. In each case, it is the agent-relative structure of these theories that makes them flout universalizability.

We are now in a position to see more clearly why the problem sketched here arises only when we consider the possibility of collective rational agents. Return to the case of individuals who are competing for a scarce resource. Recall that agent-relativism only seemed to be incompatible with CU in that case if one mischaracterized the reasons in question. Once it was clear that to respect the constraint that 'ought' implies 'can' that we must characterize the reasons in such cases as reasons to try one's best to x rather than reasons to x, it became clear that there was no problem with universalizability after all.

The crucial contrast is that in the relevant cases involving a collective and an individual, making this move is of no help precisely because the individual's action constitutes the collective's action even if we understand the collective's reason as a reason to "try its best." In particular, it seems quite plausible to suppose both that (i) the

US will not count as trying its best to enact a gas tax at time t if the President vetoes it at time t, and that (ii) the President will not count as trying his best to veto the legislation at time t if the US passes it at time t. After all, it is plausible to suppose that an agent tries *his best* to do something only if his will is completely committed to it. However, it is also plausible to suppose that for a collective agent like the US, the collective's will is at least partially constituted by those in certain positions of authority. In particular, it is very plausible to suppose that the President would count as partly constituting the US's will. The President is, after all, in charge of the Executive branch of the government, and very many of the decisions made by the U.S. are largely in the President's hands. Indeed, if any individual has a claim to be a constitutive part of the will of the U.S. it seems that the President has the best such claim. In which case, if the President vetoes the legislation, it follows that the US's will was not *completely* committed to passing the legislation, in which case the US did not try *its best* to pass the legislation after all. Nor dos this implausibly entail that whenever the US tries to do something that it will succeed – even if it tries its best to win the war, it might still lose. Its failure in such cases will be due to external factors, though. If this account of trying is at all on the right track, then (i) is correct. Furthermore, if this account of trying is roughly right then a very plausible argument can be given for (ii). For if the US passes the legislation then given the facts of the case it follows that the President did not exercise his veto power. Given the description of the case, though, this could not be because he was deprived of the opportunity to veto the legislation or ignorant of the facts, etc. The case is meant to be understood as stipulating that the President need only exercise his will in the appropriate way to count as vetoing the legislation, and that he knows this. In which case, it would

seem to follow that if the US passed the gas tax that the President did not try his best to veto it. For given the facts of the case, if the President tried his best to veto the legislation then he would succeed. So both (i) and (ii) are correct. This means, however, that we cannot invoke the possibility of both agent's trying their best, as we did in the individual case. For given (i) and (ii), it is *not possible* for both agents to try their best because one is a proper part of the will of the other.

By contrast, these sorts of cases cannot be constructed for plausible agent-neutral theories, for the relevant sort of antagonism simply cannot emerge. Consider, for example, act-utilitarianism as a comprehensive theory of practical reason. Given that theory, there will be no cases in which a collective has decisive reason to do something incompatible with its members doing what they have decisive reason to do. For if the collective has decisive reason to do it, then it must maximize aggregate utility. However, if one of its members had decisive reason to do something incompatible with the collective's putatively rationally required action, this could be the case only because that action would produce more utility than whatever actions are consistent with the collective's putatively required action. In that case, though, the collective's putatively required action cannot be required after all, for if it instead performed the action in which the relevant member acted differently then it would follow that even more utility would be produced; otherwise, the individual would lack decisive reason to perform her action.[31] Because agent-neutralism entails that a collective and its members must share a common fundamental aim, the relevant antagonism cannot emerge. So agent-neutralism can satisfy CU while admitting that collective agents are possible.[32]

Up until this point, my argument has simply assumed that collective and individual rational agents are subject to the same principles of practical reason, but one might reasonably contest this assumption. That collectives and individuals are subject to the same principles of practical reason is at least a reasonable if defeasible initial assumption. For if collectives have different kinds of reasons from individuals then this presumably is not simply a "brute fact" - there should be some explanation of why they differ in this way. Let us suppose that some such defense can be given, and that collectives and individuals are bound by different theories of practical reason. So long as there are at least some non-trivial principles to which collectives are subject, the present argument still goes through, *even if* collectives and individuals are subject to different principles. Call the theory which applies to collectives, "C" and the theory which applies to individuals "I," where C≠I. Given that each of these theories must either be agent-relativist, agent-neutralist, or a hybrid view, then there are nine possible cases:

(1) C and I are both agent-neutralist.
(2) C and I are both agent-relativist.
(3) C and I are both hybrid views.
(4) C is agent-neutralist and I is agent-relativist.
(5) C is agent-neutralist and I is a hybrid view.
(6) C is agent-relativist and I is agent-neutralist.
(7) C is agent-relativist and I is a hybrid view.
(8) C is a hybrid view and I is agent-relativist.
(9) C is a hybrid view and I is agent-neutralist.

Since my primary aim is to refute agent-relativism as a global theory about *all* practical reasons (both for collectives and individuals), strictly speaking I only need to deal with (2). In fact, the argument applies straightforwardly to (2). For if it can be shown that a collective and one of its individuals can be pitted against one another in the relevant way when they accept the *same* principle, it will be all the easier to show how this could

happen when they embrace *different* principles. Returning to the gas tax case, let us suppose collectives have reason to promote the satisfaction of their preferences, and individuals have reason to promote their hedonistic well-being. Then it might well be the case that enacting a gas tax is what the U.S. ought to do, all things considered, while the President must veto the gas tax if he is to maximize his hedonistic well-being. In fact, the argument can be extended quite easily to agent-relativism for collectives only [(6) and (7)] as well as agent-relativism for individuals only [(4) and (8)]. For in each case, if one party is subject to agent-neutral reasons and the other party is not subject to such reasons, then there will be cases in which the one party's agent-neutral reason outweighs any relevant agent-relative reasons she might have. In which case, the one party will be required to act on the basis of an agent-neutral reason and the other party will not, for *ex hypothesi* the other party simply is not bound by any agent-neutral principles. In that case, though, it could be that the collective has decisive reason to do something which is incompatible with the relevant individual doing the same thing. In the gas tax case, if we suppose that collectives have agent-neutral reasons but individuals do not, then the U.S. might have decisive reason to pass the gas tax for such reasons. Still, the President might have decisive agent-relative reason to veto it. This eliminates (4) and (8), and the argument can just as easily be run in the other direction, in the cases in which individuals, but not collectives have agent-neutral reasons). For the same antagonism can arise with the roles reversed, and this eliminates (6) and (7). In fact, the argument can even be extended to some versions of (1), in which all reasons are agent-neutral, but collectives are bound by different agent-neutral principles than those binding individuals. Suppose that collectives have agent-neutral reason to maximize the prospects of the least well-off,

whereas individuals have reason to maximize aggregate utility. It is not hard to see how a case could be generated for a collective like a nation-state and the relevant individual (e.g., the President), involving legislation which would maximize the least well-off at the expense of aggregate utility, again producing the relevant sort of antagonism and thereby flouting universalizability. Surprisingly, the present argument therefore also seems to refute even fully agent-neutral theories according to which collectives and individuals are subject to different principles, though I shall not here try to prove this decisively.[33] So the possibility of collectives and individuals being bound by different principles is a red herring as an objection to the present argument, though it is an instructive one.

Recall that the present argument applies only to theories of insistent reasons. Consider, e.g., the following theory of practical reason:

(1) There is insistent reason to maximize utility (agent-neutral).
(2) There is *non-insistent* reason for an agent to maximize *her own* utility (agent-relative).
(3) The reasons of (2) trump the reasons of (1) unless the agent's promoting her own utility involves a disproportionate sacrifice of aggregate utility (meta-principle).

The notion of "trumping" as deployed in (3) is not meant to be inconsistent with the non-insistence of the reasons characterized by (2). For a non-insistent reason to trump an insistent reason (as I am using the term 'trump') is for it to be *permissible* for the agent to act on the trumping reason in spite of the opposing reason(s). Whereas an insistent reason's being trumping (again, as I am somewhat arbitrarily using the term 'trumping') is for it to be *required* that the agent act on the trumping reason in spite of the opposing reason(s). With the notion of trumping so understood, I take it that the above theory (better: theory schema, for the notions of utility and disproportion must be spelled out for purposes of a full theory) represents a view much like the one defended by Samuel

Scheffler in *The Rejection of Consequentialism*, though Scheffler himself does not explicitly characterize his theory in terms of insistence versus non-insistence.[34] The theories differ in a number of other important respects. For example, Scheffler argues for a distribution-sensitive version of (1). Also, Scheffler's theory is meant only to be an account of moral reasons, whereas the above account is meant to be a global theory of practical reason. Still, the above theory is similar in spirit to Scheffler's. For the idea behind the above theory is that while it is always permissible (and sometimes required) to promote aggregate utility, it is also very often permissible but not mandatory for an agent to promote her own utility *even when doing so is incompatible with maximizing aggregate utility*. Such a theory incorporates what Scheffler calls an "agent-centered prerogative,"[35] but does not include any deontological restrictions.[36]

The crucial point for present purposes is that the above theory holds that all *non-insistent* reasons are agent-relative and yet is not undermined by the present argument. The reason it falls outside the scope of the present argument is precisely because of the non-insistence of the agent-relative reasons. Return to the gas tax case. That case, and others like it, present no problem for a theory like this one. For while it may be impossible for the President and the U.S. both to maximize their own welfare, this does not mean that universal full compliance with the above theory is impossible. For neither agent is under any *requirement* to maximize his own utility - that is the point of the non-insistence of the agent-centered prerogative. So if the U.S. passes the gas tax, and the President therefore does not veto it, each will have acted permissibly. It is only if the agent-relative reasons in question are insistent that the present argument would come into play, as in that case it would not always be permissible for the President to forgo his own

interests for the greater good. That such prerogatives are compatible with the present argument is significant, as they are *prima facie* plausible, and if the present argument ruled out such prerogatives then perhaps it would "prove too much."[37]

**II. Interdefining Theories.**

A tempting response to the argument of section one is to invoke the possibility of an agent-relative theory giving priority either to the reasons of the collective or to the reasons of the individual in cases in which they seem to come into conflict. It might be argued that this is both independently plausible and analogous to invoking a meta-principle to adjudicate between an individual's conflicting *prima facie* duties. For it might seem that the case of collectives and individuals is just an interpersonal version of the intrapersonal conflict found in the case of conflicting *prima facie* duties. In thinking about the objection it is helpful to distinguish between two sorts of agent-relative theories. Some agent-relative theories may make what an individual agent has reason to do a strict function of what the collective(s) to which she belongs has reason to do. Going in the other direction, some agent-relative theories may make what a collective agent has reason to do a strict function of what all the individuals currently constituting that agent have reason to do. Each of these kinds of theories distinguishes between individual and collective agents, and defines what the one has to do in terms of what the other has reason to do. In effect, such “interdefining” theories define away the conflicts otherwise endemic to agent-relativism.

I take the argument of section one to have shown that all non-interdefining versions of agent-relativism are unsound (given CU and the possibility of collective rational agents), and this would be of substantial interest in its own right. After all,

defenders of agent-relativism have not recognized that they must adopt an interdefining theory simply in virtue of universalizability, and this is an interesting result. Still, my larger aim here is to refute agent-relativism across the board, and at this point its seems that the agent-relativist could avoid the force of my argument by adopting an *inter-defining* version of agent-relativism. I have until now implicitly put such theories to one side. In this section I present independent arguments against interdefining theories. Interdefining theories can come in two varieties. First, the agent-relativist might define an individual's reasons in terms of what each and every collective to which she belongs has reason to do, so that the individual's reasons and the reasons of the collective(s) to which she belongs could never come into conflict. On this account, a well-functioning termite colony might provide an apt metaphor for the seemingly appropriate relationship between an individual and her collective. Call this strategy "collectivism."[38] Second, the agent-relativist might adopt the converse strategy, and define a collective agent's reasons in terms of what each and every one of its current members has reason for it to do, so that unless there is a genuine consensus of reasons among its constituents, it will not have reason to do anything. Call this strategy the "individualist" one, since it makes what any collective has reason to do completely dependent on what each and every one of its members converges on having reason to promote. The main virtue of each of these accounts is that by inter-defining what an individual and her collective ought do, the relevant antagonism seems no longer to be possible. Nonetheless, such theories flout CU.

***Collectivism.***

In some ways, it would be ironic if the agent-relativist were forced to adopt the collectivist position. For the clearest paradigm of agent-relativism is rational egoism,

which is reasonably understood to be an individualist account. An arch-egoist like Thrasymachus, after all, would hardly have advocated the individual sublimating her own individual desires and needs for the greater good of the collective(s) to which she belongs. More generally, those of us with liberal individualist sympathies will be *prima facie* quite suspicious of any view that so completely subordinates the reasons of the individual to her collective(s). Historically, accounts which give such pride of place to a person's role as a member of some larger collective have paved the way for ugly forms of totalitarianism and fascism. So the agent-relativist might purchase CU via collectivism at the price of giving up its original plausibility.

However, collectivism actually flouts CU. By defining the reasons of the individual in terms of the reasons of the collective(s) to which she belongs, the collectivist account does avoid cases in which it is impossible for both the individual and a collective to which she belongs to act as they ought. It is for this reason that the collectivist account seems tailor-made to avoid the main line of the present argument. However, there is another sort of problem involving collective agents that even collectivism cannot avoid. The relevant cases are certain of those in which two collectives share a common member who stands in a privileged position in each collective. The problem with CU arises, *not* because it is impossible for an individual and a collective both to act as they ought, but because it is impossible for *two* collectives both to act as they ought, given their overlap. Defining an individual's reason in terms of the reasons of the collective(s) to which she belongs is of no help here. Again, an example helps make the point.

Steve Jobs is CEO for both Pixar Animation and Apple Computers. I assume that such corporations count as collective rational agents. Take one's favorite agent-relative theory, and let us suppose that the facts are such that according to that theory Pixar and Apple both ought to merge with other companies (Pixar with 20th Century Fox, and Apple with Time Warner, say). For in each case, let us suppose, such a merger would dramatically increase the profits of each corporation, and our agent-relative theory entails that a corporation always ought to maximize its profits. Further, let us suppose that in each case time is of the essence, for one reason or another - if the merger does not happen today then it either will not happen or will no longer be profitable. Perhaps this is because the Congress is about to pass a law (tomorrow, say) that will have a substantial bearing on mergers but which will not apply retroactively to mergers enacted prior to its passage. Now, finally, let us suppose that each of these mergers can be brought about only if Steve Jobs attends a crucial meeting, in one case with Fox and in the other case with Time Warner. In each case, we might suppose, either no other representative from Pixar and Apple could attend or if they could attend that they would not have sufficient credibility to close the deal. However, the meetings are going to happen simultaneously and in distant locations, so that there is no way for Jobs to attend both. Given our agent-relative theory, Pixar ought to close its deal *and* Apple ought to close its deal, but at this point in time it is impossible for both of them to do so. If Pixar acts as it ought, then Jobs will not attend the relevant meeting on behalf of Apple and Apple will not have acted as it ought, and vice-versa. CU is flouted once again, as it is impossible for both agents (Pixar and Apple) to act as they ought, according to the theory. The interdefining collectivist account is of no help in such cases. For the collectivist account helps only to

rule out cases in which an individual like Jobs and a collective like Pixar cannot both act as they ought, and that is simply not the problem in this case. Rather, the problem here is that it is impossible for both collectives to act as they ought in virtue of their sharing a common member who stands in a critical position in each collective. The argument easily generalizes. For any agent-relative theory, construct a possible case in which two collectives share a common member who must act in one way if one collective is to count as acting advisably and must act in another, incompatible way if the other collective is to count as acting advisably.

Agent-neutralist accounts avoid this problem by not pitting collectives against one another in the relevant way. It is again instructive to consider act-utilitarianism. Cases like the Pixar/Apple case pose no problem for act-utilitarianism. For each of the mergers produce either the same amount of utility as the other, or they do not. Either way, it is possible for both corporations to act rightly, given act-utilitarianism. If each merger produces the same amount of utility, but neither is optimal, then neither should be performed, so there is no problem. If each merger produces the same amount of utility, and each is optimal, then it is still possible for both to act rightly. For it is possible for either one of them to enact the relevant merger, in which case, the one that does enact the merger obviously has acted rightly. Not quite as obviously, though, the other corporation has also acted rightly in that case. For in that case, a consequence of the other corporation's not merging is that the other corporation did merge, and the other corporation thereby still acted optimally. Returning to our case, if Pixar did not merge with Fox, then Pixar can claim that if they had merged with Fox that Apple would not have merged with Time Warner, and by hypothesis Apple's merging with Time Warner

was equally optimal. Admittedly, this does require that we adopt a broad conception of what it is for a state of affairs to be a consequence of an agent's action (or omission) but that does not pose any obviously insurmountable problems. A similar argument works in the case in which the mergers are not equally good. To be specific, let us suppose that Apple's merger would produce more utility, and would be optimal. In that case, Apple is required to merge with Time Warner. If Pixar refrains from merging and Apple then merges instead, then Pixar's action was permissible (in fact, it was required) according to act-utilitarianism, given that if Pixar had not refrained from merging then Apple could not have merged and Apple's merging by hypothesis would produce more aggregate utility. So again, it is possible for both collectives to act rightly, given the circumstances of the case. Cases with this structure do not pose a challenge to agent-neutral theories like act utilitarianism simply because such theories do not involve the relevant sorts of antagonism between agents.

***Individualism.***

So collectivist versions of interdefining agent-relativism therefore are not universalizable. Perhaps individualist accounts fare better. It is, after all, intuitively much more plausible to suppose that a collective's reasons are subordinated to the reasons of the individuals which make it up, rather than vice-versa. For the individualist account to avoid the present argument, however, it must hold that a collective has decisive reason to act in a certain way only if its acting in that way is supported by the reasons of all the agents currently constituting it. Without such a unanimity condition, the possibility of the relevant sort of antagonism remains. Such unanimity is *extremely* unlikely to be forthcoming in a huge range of real cases. Note that even in the case of a collective of

only three persons that only one out of eight possible patterns of reasons for acting would be the one with unanimity on a particular action by the collective. As we expand the number of agents, this problem becomes more and more pronounced, since for any collective with n members there will be $2^n$-1 non-unanimous patterns.[39] Needless to say, with collectives as large as modern nation-states such unanimity is almost never going to be realized, even putting aside complications due to the fact that most individuals belong to multiple collectives. Indeed, for all practical purposes such an account would mean that most collectives never have any reason to do anything. Perhaps the individualist will be happy with this result but it seems implausible when one contemplates real cases. If, we allow the US is a rational agent, is it plausible to suppose that it has no reason to impose a gas tax because a few Texan oil barons would be harmed by it?

A deeper problem faces the individualist account, though. To be plausible, it seems that the individualist view should incorporate the following requirement - of two actions available to a collective, A and B, if B would be inferior to A for *all* of the individuals who constitute the collective, then the collective has more reason to perform A than B. Call this a mutual advantage requirement. For the idea behind the individualist account is that a collective's reasons are a direct function of the reasons of the individuals who constitute it. It would therefore be perverse if the individualist theory allowed a collective to forgo an opportunity to make *all* its members better off, as each member has at least some reason to make themselves as well-off as possible. However, if we incorporate such a requirement then the individualist account fails CU.

Here we must turn to certain kinds of prisoners' dilemmas. As noted earlier, the general phenomenon of prisoners' dilemmas seems to pose no threat to agent-relativism.

However, certain particular kinds of prisoners' dilemmas involving agents who occupy a special role within a collective rational agent do present a problem for agent-relativism, given CU. To ease exposition, I again assume rational egoism provides all individual reasons for acting, but the argument easily generalizes. Let me begin with an example. In the real world, a law faces judicial review only *after* it has been officially passed into law, and has been challenged by someone with standing. Let us imagine a world in which the U.S. is just the same, save that in this world, a bill is not officially law until it has undergone successful judicial review. Let us suppose that in this world, the Congress has passed conservative legislation outlawing flag-burning. It would be in the interest of the President if the law did not pass, for the President has deeply held liberal values. However, it is also in the President's interest not to exercise his veto power, for Congress would definitely overrule his veto and he would then look very weak and have accomplished nothing. So, all things considered, he rationally prefers not vetoing the legislation to vetoing it. Whether he vetoes it or not, though, the legislation will then go on to the Supreme Court. Naturally, the President would prefer that the Court strike down the legislation. In fact, killing the legislation is more important to the President than not looking weak. So the President has the following ranking of outcomes:

1. President does not veto, Court strikes down.
2. President vetoes, Court strikes down.
3. President does not veto, Court does not strike down.
4. President vetoes, Court does not strike down.

The Court has a different agenda. The Court rationally (given their interests) would prefer the President's being weakened. Further, they would prefer that the legislation not

be struck down. Suppose, however, that it is more important to the members of the Court that the President be weakened than it is that the legislation be passed. The Court, therefore, has the following ranking of states of affairs:

1. President vetoes, Court does not strike down.
2. President vetoes, Court does strike down.
3. President does not veto, Court does not strike down.
4. President does not veto, Court does strike down.

Finally, let us suppose that whether the President vetoes the legislation will not influence whether the Court upholds it. The President and the Court are in a Prisoner's dilemma.[40] For we have the following matrix of rankings:

| | | Court | |
|---|---|---|---|
| | | Strikes Down | Does not strike down |
| President | Vetoes | 2, 2 | 4,1 |
| | Does not Veto | 1,4 | 3,3 |

Since the President and the Court are in a Prisoners' Dilemma, if each acts rationally then they will both end up in the lower right-hand cell, with their third-ranked options. Finally, though let us suppose that everybody else in the U.S. would prefer that the legislation not be passed than that it be passed.[41] Assuming our mutual advantage requirement, though, this means that it is impossible for the U.S. *and* all of its constituent

individuals to accept and follow the theory in question. For if the President and the Court act rationally given their commitment to egoism, then the U.S. will have passed the legislation in a way that is worse for everyone in the U.S. than another option available to the U.S. - *not* enacting the legislation in a particular way (via a veto followed by the Court's striking down the legislation). Since the upper left-hand cell is better for everyone than the lower right-hand one for the individuals constituting the U.S., the U.S. is obligated by the theory's mutual advantage requirement to perform the action represented by the upper left-hand cell. Again, what the U.S. does in this situation is constituted by the actions of the President, the Congress, and the Court, so that if the President, Court and Congress act in certain ways then the U.S. will count as having passed the legislation, and if they act in other ways, then the U.S. will not count as having passed the legislation. So if the President, Congress and Court all comply with the theory then the U.S. necessarily does not. CU is flouted again. So long as the individualist theory includes a mutual advantage requirement, it fails CU. If individualism rejects this requirement, then it contradicts the only intuitions from which it might draw support.

**Conclusion.**

If the present argument succeeds, then we have sufficient reason to reject agent-relativism about insistent reasons in all its forms. However, even if we should not reject agent-relativism on the strength of the present argument, there would still be a fairly interesting argument in this general neighborhood. For at the very least, I hope to have shown that the following propositions give rise to a contradiction:

(1) Compliance Universalizability.
(2) The thesis that certain kinds of collective rational agents are possible.
(3) Non-interdefining agent-relativism.

I have taken the way in which these theses give rise to a contradiction to show that we should abandon (3) on the strength of (1) and (2). However, even if we should hold onto (3) we need to avoid contradiction in some way. Agent-relativists might try to argue from (2) and (3) to the conclusion that universalizability must be rejected or at least qualified in its application to collectives. Alternatively, one might argue from (1) and (3) to the conclusion that the relevant sorts of collective rational agents are not possible. Any of these results would be of substantial interest. I therefore conclude that however we avoid this looming contradiction we are likely to learn something interesting. [42]

---

[1] For discussion of these different ways of interpreting the relevant notion(s) of a contradiction with respect to the categorical imperative, see Christine Korsgaard, "Kant's Formula of Universal Law," in her *Creating the Kingdom of Ends*. (Cambridge: Cambridge University Press, 1996): 77-105.

[2] It is, moreover, an assumption which has been given considerable defense. See, for example, Peter French, "The Corporation as Moral Person." *American Philosophical Quarterly* 16 (1979): 207-215, Joel Feinberg, "Collective Responsibility," *Journal of Philosophy* 65 (1968): 674-688, Margaret Gilbert, "Modelling Collective Belief," *Synthese* 73 (1987): 185-204, Margaret Gilbert, *On Social Facts* (London: Routledge, 1989), Rolf Gruner, "On the Actions of Social Groups," *Inquiry* 19 (1976): 443-454, and Larry May, "Collective Intention and Shared Responsibility," *Nous* 24 (1990): 269-277.

[3] The agent-relative/agent-neutral distinction finds its classic discussion in Thomas Nagel, *The Possibility of Altruism* (Princeton: Princeton University Press, 1970) though in that work Nagel uses the overworked terms 'objective' and 'subjective'. He later takes over the terms 'agent-relative' and 'agent-neutral' from Derek Parfit, who introduced them in *Reasons and Persons* (Oxford: Oxford University Press: 1984).

[4] Some philosophers argue that Kantianism must be understood in an agent-relative way, but that is much less obvious and I shall not assume it here (in fact, I think it is false).

[5] See Frances Kamm, *Morality, Mortality*, vol. 2 (New York: Oxford University Press, 1996), p. 231.

[6] Shelly Kagan suggests that such reasons are impossible, and tries to show that prima facie plausible instances of non-insistence are best reinterpreted as slightly different insistent reasons). See Shelly Kagan, *The Limits of Morality* (Oxford: Oxford University Press, 1989), pp. 378-381.

[7] See Samuel Scheffler, *The Rejection of Consequentialism* (Oxford: Clarendon Press, 1982). I discuss Scheffler's account more below.

[8] It might even be held that David Hume took the radical view that *all* reasons for acting are non-insistent, given that he seems to think both (a) that there are reasons for and against actions that can be appreciated from a "general point of view," *and* that (b) no action, as such, is really irrational (though it may be called "irrational" in an extended sense if it is based upon an irrational belief). One nice way of reconciling these prima facie conflicting claims would be to suppose that Hume in effect thought all reasons for acting are non-insistent.

[9] Immanuel Kant, *Groundwork of the Metaphysics of Morals* , trans. Mary Gregor, in *The Cambridge Edition of the Works of Immanuel Kant: Practical Philosophy* (Cambridge: Cambridge University Press, 1996), p. 73 (421).

[10] There is room for some dispute as to whether this is what Kant actually thought. For support for my interpretation, see Onora Nell (now O'Neill), *Acting on Principle* (New York, Columbia University Press, 1975). Some instead hold that on Kant's view, maxims are simply intentions, where intentions are understood in a way that involves no normative judgment whatsoever. For present purposes I need not

delve into this exegetical question, since the question here is not what Kant thought, but what follows from one philosophically interesting interpretation of Kant's universalizability constraint.

[11] Trans. By Theodore Greene and Hoyt Hudson (New York: Harper & Row, 1960).

[12] Barbara Herman claims that there is "now fairly general agreement" on this reading. See Herman, *The Practice of Moral Judgment*. Cambridge, MA: Harvard University Press (1993): 46-47.

[13] For an argument in a similar vein, but with an emphasis on the way in which a law of nature must itself hold with necessity, see Ted McNair, "Universal Necessity and Contradictions in Conception," *Kant-Studien* 91 (2000): 25-43.

[14] See Kant's the *Groundwork of the Metaphysics of Morals*, previously cited, p. 83 (433-434).

[15] Though, as I shall argue (in section II), certain particular kinds of prisoners' dilemmas involving collective rational agents do present a problem for the agent-relativist if we accept universalizability.

[16] It might be objected that one cannot try when one knows one's efforts will fail. A couple of points are worth making in response to this objection. First, if this were true it would not actually help to show agent-relativity flouts universalizability in such cases. For *if* it were true that one of the agents simply cannot try to get the resource because she knows she would fail then given that 'ought' implies 'can' it is no longer the case that she should even try to get it in such a case. So again there is no problem with universalizability; the case would collapse into the previously considered case in which universalizability is satisfied in virtue of the possibility of one agent's acquiescing and the other agent's getting the resource. Second, it is not obvious that one cannot try even when one strongly believes or knows that one will fail if one tries. So long as there is some chance in the agent's mind, however slim, that she might succeed if she tried then it seems like trying is a possibility. I can try to make a basketball shot that I think I have almost no chance of making. Indeed, it seems that I can try even when I am quite sure I will fail. I am absolutely positive that I cannot swim across the Atlantic all by myself from New York to England. Pretty clearly, though, I could try to do it anyway - I could get in the water and start swimming in what I take to be the correct direction.

[17] It is worth noting that in assuming that nation-states are rational agents I may be committed to supposing that they are moral agents as well, but it would *not* follow that they must also be moral *patients*. It has long been recognized that there can be moral patients that are not moral agents - nonhuman animals, infants, and other sentient but non-rational beings are all plausible cases (Kant's view of nonhuman animals notwithstanding). The question of whether the converse is possible - whether there might be moral agents who are not also moral patients - is rarely discussed. Collective agents may be instances of this category. For while it is very plausible to suppose that collectives intentionally perform actions for which they are responsible, it is not so plausible to suppose that collectives (with which we are familiar, anyway) have moral rights or interests which transcend the interests of the individuals constituting them. A failure to note that the fact that collectives like nation-states are rational and moral agents does not entail that they are also moral patients may be indirectly responsible for the influence of a certain kind of scepticism about the possibility of moral and political theory in international relations and/or an overblown scepticism about the possibility of legitimate intervention by one state in the affairs of other states. For such scepticism arguably relies on a strained analogy between individual human beings *qua moral patients* (in particular, rights-bearers) and nation-states. For interesting discussion of this point (though not in these terms), see Charles Beitz, *Political Theory and International Relations* (Princeton, NJ: Princeton University Press, 1979). The crucial point, for present purposes, is that scepticism about collectives as moral patients should not undermine the idea that collectives can be moral agents.

[18] Carol Rovane's recent account of agents as aiming at rational unity is a good case in point. While Rovane argues at some length that what she calls "group persons" are possible, it is not at all clear that she would allow that collective rational agents, in the sense I have in mind, are possible. Unfortunately, I lack the space here to give Rovane's account the detailed discussion it deserves. See Carol Rovane, *The Bounds of Agency* (Princeton: Princeton University Press, 1999). For present purposes, I shall just indicate that my assumed conception of collective rational agents is closer to the sorts of accounts developed by Peter French and David Copp (references above and below, respectively).

[19] For a useful discussion of the way in which an individual's action can constitute the action of a collective to which she belongs, and more generally of how one person's action can constitute the action of another (as in fiduciary relationships, e.g., where no collective agency is involved), see David Copp, "Collective Actions and Secondary Actions" *American Philosophical Quarterly*. 16 (1979): 177-186.

[20] For present purposes, I overlook the possibility of the veto's being overridden, though it would be simple enough to take that possibility into account. We would simply need to add that if everyone in Congress

acts rationally according to egoism, the veto will not be overridden, for reasons similar to those that made it irrational for the president to refrain from vetoing it. Or we could imagine an alternative political system in which the President's veto cannot be overridden.

[21] It might be objected that whenever we consider whether it is possible for everyone to comply with a theory in a world in which they do not actually comply with it that we must consider possible worlds which differ from the actual world in ways that go beyond their complying. For it seems that they would not have complied with the principle if their histories had not been different in various ways – this is just to register the point made familiar by David Lewis, that such counterfactuals are "backtracking." In which case, one might hold that it is possible for the President and the U.S. both to comply with the theory, in that there is a possible world in which the President's situation was very different when the time came to pass the gas tax (perhaps in the relevant possible world, passing a gas tax would not have been unpopular). The objection merits two replies. The first is that the objection that the question of whether an agent could have acted differently is settled by considering such back-tracking counter-factuals. Here it is important to register that the 'could' in CU is meant in roughly the same sense as the 'could' in the 'could have done otherwise' is best understood in the debate over free will and responsibility (which is not to suggest that there is agreement over what is the best understanding of 'could' in that context) This means that if we are incompatibalists, though, we will hold that the possibility of the Lewis-style back-tracking counterfactuals are irrelevant to whether the agent "could have done otherwise" at the time of action, and so those will not establish the possibility of joint compliance in the relevant sense. Rather, what is relevant is whether both agents could, by an act of will, simultaneously both act appropriately at the time of action even if we keep their histories completely fixed. Moreover, insofar as some sort of incompatibalism is often (though not universally) thought to be presupposed by Kantian moral and political philosophy, it is not unreasonable to help myself to that assumption in exploring the ramifications of that theory against the background of the possibility of collective rational agents.

Second, any change in the history that would make it possible for both agents to comply with the theory in the gas tax case (and analogous cases) would involve a change in the features of the situation which are themselves relevant to the agent's choice according to the theory of practical reason in play (egoism, in this case). This, however, is not the sort of possibility which is relevant to universalizability. Clearly, the fact that there is a possible world in which the President's action would not be unpopular is no more relevant to whether he and the U.S. can both comply with egoism than is the fact that there is a possible world in which I am wearing a parachute is relevant to whether I could successfully jump out of a plane here in the actual world in which I do not have a parachute. Note that compatibalists more generally had better be able to rule out some nearby possible worlds as irrelevant in this way in any event. For it is surely not the case that Henry now ought to type 70 words per minute (with no errors) even though he is illiterate and cannot type. Nonetheless, there is a perhaps nearby possible world in which at this point in time Henry is not illiterate and can type, in virtue of the right sort of "back-tracking" modifications. The compatibalist needs some story as to why this possible world does not block the 'ought' implies 'can' objection to the claim that Henry ought to type 70 wpm with no errors. Thanks to Michael Smith for useful discussion of this point.

[22] David Copp, "Collective Actions and Secondary Actions" *American Philosophical Quarterly*. 16 (1979): 177-186, p. 180.

[23] Copp, p. 180.

[24] Copp, p. 178.

[25] Margaret Gilbert, *On Social Facts*, 1990 (Routledge: London), p.206.

[26] See Gilbert, p. 289 and pp. 293-294.

[27] She notes that, "…a plausible account of group action through the acts of special representatives will not generate any objections to my general account of social groups." (Gilbert, p. 207)

[28] Christine Korsgaard has a similar view, maintaining that members of collectives must take themselves to have certain distinctive kinds of reasons; see Christine Korsgaard, "Self-Constitution in the Ethics of Plato and Kant," *Journal of Ethics*, 1999.

[29] In fact, we can construct cases that handle agent-relative theories that understand all reasons in terms of one's membership in a collective, but a discussion of this point must await the discussion of so-called "interdefining theories" in section II.

[30] This account has been defended by Group Captain F.W. Winterbotham in his, *The Ultra Secret* (New York: Dell, 1982, first published by Harper & Row in 1974). For a recent criticism of the view that

Churchill knew in advance, see Allan W. Kurki, *Operation Moonlight Sonata: The German Raid on Coventry* (London: Praeger, 1995), chapter 11.

[31] Donald Regan has independently given a virtually identical argument to the one given in the text. See Donald Regan, *Utilitarianism and Cooperation* (Oxford: Clarendon Press, 1980), pp. 54-55.

[32] However, there are cases involving the sorities paradox and the apparent non-transitivity of the "is perceptibly different from" relation, which seem to pose a problem for agent-neutralists in spite of their theory not involving the relevant antagonism. The classic discussion of the sorts of cases I have in mind is Parfit, *Reasons and Persons*, Chapter 3. Parfit discusses a case inspired by Jonathan Glover in which each person's contribution of water to a collective effort will make no perceptible difference to anyone though the aggregate of their contributions would alleviate the thirst of those suffering from dehydration. See also Jonathan Glover, "It Makes No Difference Whether Or Not I Do It," *Proceedings of the Aristotelian Society*, *Supp. Vol*. 49 (1975). Parift's discussion of such cases does not suggest that the groups of people involved form a collective agent but this could easily be added, and then even agent-neutral theories might look like they cannot accommodate the possibility of collective rational agents and universalizability. For in Glover-type cases, such theories would seem to require each individual to drink her water herself (say) since giving it away makes *no* perceptible difference (for each individual's small contribution will be divided equally amongst thousands, say) to anyone, but she would get at least some enjoyment from drinking it herself. Whereas it would be wrong for the collective not to give away any of its water, as that would forgo the opportunity to make an enormous perceptible difference to those in need.

At best, though, this is true of *some* but not all versions of agent-neutralism. For versions of agent-neutralism that define an individual's right action in terms of the right action of others (or, at least, like-minded others who also aim to cooperate for the sake of utility maximization - see Regan, *Utilitarianism and Cooperation* for a plausible view of this sort) can avoid these sorts of problems.

Further, it is not at all clear that these cases really even undermine more straightforwardly act-utilitarian kinds of views. We might, for example, follow Parfit and hold that there are imperceptible harms of the relevant kind, and that would seem to avoid the problem. For further defense of this claim, as well as the more general claim that the utility of the whole can be no greater than the utility of the sum of its parts, see Michael Otuska, "The Paradox of Group Beneficence," *Philosophy and Public Affairs* 20 (1991): 132-149, esp. pp. 145-148. I suspect that the counter-intuitiveness of imperceptible harms relies on a failure to distinguish a perception of a difference from a difference in one's perceptions, but I lack the space here to elaborate on that suspicion. In any event, this problem has its source in the sorities paradox, which is a genuine paradox and a problem for everyone. It is therefore likely that the correct solution to the sorities paradox (whatever it is) would provide the act utilitarian with an adequate solution. Many thanks to Folke Tersman for useful discussion here.

[33] In fact, the argument might even extend to (3), (5) and (9), and show that collectives and individuals must be bound by exactly the same principles, period. Whether this is so depends upon how the present argument bears upon the hybrid view more generally, though, and I must put that very complicated issue to one side for present purposes.

[34] See Scheffler, *The Rejection of Consequentialism*, especially p. 20, for Scheffler's official characterization of such prerogatives. Interestingly, Bernard Gert's view, as I understand it, takes the opposite view of which reasons are insistent and which are not insistent, though he does not use the terms 'insistent' and 'non-insistent'. For Gert holds that there are certain agent-relative considerations (one's own pain, for example) which are such that it is always irrational not to act upon them unless one has some other reason that one takes to outweigh those agent-relative considerations. So agent-relative reasons are insistent - when they require a particular action, one is irrational not to act upon them. By contrast, on Gert's view, it is never irrational not to act on what one correctly takes to be one's agent-neutral reasons. Agent-neutral reasons, while perfectly real, are non-insistent. Since Gert's view, as I understand it, is committed to agent-relativism about insistent reasons, his view should fall within the scope of the present argument, and fail to satisfy universalizability given the possibility of collective rational agents. See Bernard Gert, *Morality: Its Nature and Justification* (New York and Oxford: Oxford University Press, 1998).

[35] Frances Kamm employs the same idea but refers to them as "options."

[36] In this regard the theory is similar in spirit to the one Scheffler defends in *The Rejection of Consequentialism*. More recently, Scheffler has warmed to the idea that at least some agent-centered

restrictions may be defensible. See Samuel Scheffler, "Relationships and Responsibilities," *Philosophy and Public Affairs* 26 (1997): 189-209, esp. p. 209.

[37] Since deontological reasons are plausibly thought to be insistent (they are meant to be requirements, after all), one might wonder whether the present argument would present at least a prima facie problem for deontological restrictions. I must, however, put this question to one side here because answering it would require settling the logically prior issue of whether such restrictions are agent-centered. They are generally thought to be so, but in my view this should be seen as a controversial and perhaps mistaken view. For an argument that deontological restrictions are not agent-centered, see Eric Mack, Deontic Restrictions are Not Agent-Relative Restrictions" *Social Philosophy and Policy* 15 (1998): 61-83. If such restrictions are agent-centered, however, and if the present argument were to show that such restrictions could not be sound then the argument in the text would not only be compatible with Scheffler's view (in *The Rejection of Consequentialism*, anyway) that in addition to an agent-neutral reason to promote the good, there are agent-centered prerogatives but no agent-centered restrictions, it would provide an independent argument for that view.

[38] It might be argued that Hegel held a view in this neighborhood, given his claim that the "supreme duty" of the individual is to be a member of the state. See Hegel's *The Philosophy of Right*, trans. T.M. Knox (London: Oxford University Press, 1952), section 258.

[39] A point also made by James Coleman. See *Power and the Structure of Society*, (W.W. Norton and Company: New York, 1974), p. 39.

[40] For simplicity, I assume that though the President and the Court will continue to interact with one another, that this does not undermine the rationality of dominance reasoning in the case at hand. Whether this is generally true or not is controversial, since many argue that it is rational to adopt and follow through on a "tit-for-tat" strategy in iterated prisoners' dilemmas. Regardless of how this controversy is settled, though, there can be mitigating factors in a given case that would speak against the adoption of such a strategy (the political fallout of such a strategy, e.g., might be serious), and we might suppose that the President in any case has no effective direct means of retaliating against the Court, since they have the final say over whether the law is passed or not.

[41] It might then seem that Congress must have acted irrationally in voting for the legislation in the first place, but we can assume that they did so simply in response to bribes from external agents.

[42] Many thanks to Simon Blackburn, Zena Childs, Stephen Darwall, Robin Flaig, Robert E. Goodin, Thomas E. Hill, Jr., Keith Horton, Karen Jones, Chandran Kukathas, William G. Lycan, Sean McKeever, Philip Pettit, Gerald Postema, Daniel Ryder, Michael Smith, Geoffery Sayre-McCord and Folke Tersman for helpful comments, discussion, and encouragement. Thanks also for useful discussion of an earlier version of the paper given at the *Paton Colloquium* in St. Andrews in 2002 and in particular thanks to Christopher Taylor and Jens Timmermann for their incisive commentary on the paper at that event.