

ARTICLES

The Universalizability Trap: Rational Agency, the Authority of Ought-Claims, and the Limits of Utilitarian Justifications for Killing

Daniel Teshman^a

Keywords: analytic ethics, ethics, normative contradiction, utilitarianism

<https://doi.org/10.35297/001c.169490>

Journal of Libertarian Studies

Vol. 30, Issue 2, 2026

This article defends a formal argument concerning whether a group of rational agents (A , B , and C) can be rationally justified in killing a fellow rational agent (D) on utilitarian grounds. I do not contest the utilitarian calculus, and I do not claim that utilitarians derive morality from instrumental rationality. I argue instead for a general condition on when a normative claim binds an agent. A claim of the form " S ought to ϕ " gives S a reason that S cannot rationally reject only if S already shares the end from which the claim is generated or if S 's acting otherwise is demonstrably incoherent. Utilitarian prescriptions to eliminate a nonconsenting agent satisfy neither condition with respect to that agent. There is no contradiction in D 's continuing to pursue survival, and D is under no rational compulsion to adopt the end of aggregate welfare maximization. The utilitarian conclusion, however sound within its own framework, therefore carries no authority over D . Since the lethal coercion of a rational agent requires a reason that binds that agent, and the utilitarian prescription supplies none, A , B , and C are not rationally justified in killing D . I call this result the universalizability trap, and I distinguish it carefully from Kantian universalization and from Alan Gewirth's argument to the principle of generic consistency.

One of the most persistent questions in moral philosophy concerns what is owed to an individual who stands in the way of an aggregate good. Classical utilitarianism, in its most austere form, holds that the right action is the one that maximizes net welfare. If a sincere application of that standard yields the conclusion that an innocent person ought to be killed for the benefit of a sufficiently large group, then, on the austere view, the conclusion must be accepted. A familiar body of work resists this implication by importing external constraints (e.g., rights, agent-relative restrictions, threshold deontology, and contractualist side-constraints), all of which have been deployed for the purpose.

This article takes a different route, and it is important to state at the outset what that route is not. I do not argue that utilitarians derive their moral theory from instrumental rationality alone; no serious utilitarian has held such a view, and I do not attribute it to utilitarians. I do not argue that D 's sincere assertion "I ought to live" makes it true that D ought to live; the capacity to voice a norm



This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-4.0). View this license's legal deed at <http://creativecommons.org/licenses/by/4.0> and legal code at <http://creativecommons.org/licenses/by/4.0/legalcode> for more information.

^a Daniel Teshman (teshmandan@gmail.com) is an independent researcher.

is not a warrant for its truth. And I do not rest the argument on a symmetric clash between two equally valid ought-claims, since, as a careful critic will observe, such a clash can always be dissolved by denying that both claims were valid to begin with.

The argument rests instead on a question that is prior to the utilitarian calculus and that the calculus cannot settle: When does a normative claim issued by one agent bind another? My contention is that an ought-claim binds an agent only if that agent either shares the end from which the claim is generated or cannot, without incoherence, act otherwise. A claim that meets neither condition is not thereby false; it is simply not authoritative for the agent in question. It retains whatever force it has for those who share its end but no more. Utilitarian prescriptions to kill a nonconsenting rational agent, I argue, are claims of exactly this kind with respect to the victim. They are binding on those who have adopted the end of maximizing aggregate welfare. They are not binding on the one person whose life they propose to end, because that person need not have adopted the end, and his declining to be killed involves no contradiction.

If this contention is right, the implication is not that *D*'s welfare outweighs the welfare of *A*, *B*, and *C* in the utilitarian ledger. The consequence is that the prescription never possessed the kind of authority over *D* that would be needed to justify killing him, as opposed to merely overpowering him. The party who proposes to kill bears the burden of producing a reason that binds the person to be killed. Utilitarianism, applied to a rational agent who does not share its end, cannot discharge that burden. I call the resulting structural limit the universalizability trap.

The article proceeds as follows. It first situates the argument in the utilitarian, agency-based, and libertarian literatures, then states the formal argument, then defends its premises: the generation of ought-claims by rational agents, the condition under which such claims bind others, the failure of the utilitarian prescription to meet that condition for *D*, and the burden of justification that consequently defeats the killing. It then states the universalizability trap as a general principle. The article next answers objections, including the three most likely to be pressed against it: that the argument covertly derives morality from instrumental rationality, that it begs the question by assuming an asymmetry between the parties, and that it merely restates Kantian or Gewirthian commitments under a new name. It closes by drawing implications for self-defense and for the authority of the state.

Situating the Argument

Because the argument turns on a claim about normative authority rather than on the contents of any particular calculus, it is worth locating it among the positions it engages.

Utilitarianism and the Separateness of Persons

The utilitarianism at issue is the agent-regarding tradition that runs from Jeremy Bentham (1789) and John Stuart Mill (1863) through the preference- and interest-based formulations of R. M. Hare (1981), John Harsanyi (1977), and Peter Singer (1993). What these formulations share, and what matters here, is that they count an individual's welfare because the individual is the kind of being whose ends can be achieved or frustrated and whose interests can be advanced or impaired. The most influential structural objection to the tradition is John Rawls's (1971, 27) charge that utilitarianism does not take seriously the distinction between persons, conflating the aggregation of satisfactions across distinct lives with the rational balancing of satisfactions within a single life. Robert Nozick (1974, 32–33, 41) sharpens the point with the side-constraint view and the figure of the utility monster, arguing that persons are not resources whose separate existences may be summed and traded against one another. My argument can be read as supplying a reason for the separateness intuition from a direction Rawls and Nozick did not take: Rather than starting from the moral standing of persons asserted at the outset, I start from the conditions under which one agent's normative conclusion can bind another.

Agency, Normativity, and the Bindingness of Obligations

Several traditions try to locate the source of an obligation's grip on us in the nature of rational agency. Immanuel Kant ([1785] 1998) distinguishes hypothetical imperatives, which bind only relative to an end the agent has, from categorical imperatives, which purport to bind regardless of the agent's ends, and he ties the categorical case to the impossibility of consistently willing the maxim's universalization. Christine Korsgaard (1996) reframes the issue as the normative question: not whether obligations exist, but what gives them authority over the agent they obligate. T. M. Scanlon (1998) answers a neighboring question for the interpersonal case, holding that a principle binds others only if it could not reasonably be rejected by those subject to it. The present argument draws on this family of views for a single, deliberately minimal premise, defended in the section "When an Ought-Claim Binds Another Agent": that to bind an agent who does not share one's end, one must show that the agent's acting otherwise is incoherent. I take no stand on the larger constructivist program, and the section "This Is Just Kantianism, or a Smuggled Constructivist Theory of Bindingness" explains why my argument does not depend on it.

The closest precedent, and the one a reader is most likely to mistake the argument for, is Alan Gewirth (1978). Gewirth holds that any agent, simply in virtue of pursuing purposes, must claim rights to the freedom and well-being that are the necessary conditions of agency, and must on pain of contradiction extend those rights to all other agents, yielding the principle of generic consistency. My argument is leaner and, I will suggest, less exposed. It does not

require the contested step from “I need the conditions of agency” to “I have a right to them,” the step that others have pressed Gewirth on. It requires only the converse and weaker observation that a party seeking to bind another agent must meet the bindingness condition, and that utilitarian elimination cannot meet it. The section “This Is Just Gewirth’s Argument” develops the contrast with Gewirth.

The Libertarian Literature

For readers of this journal, the relevant background is the libertarian treatment of when force against a person can be justified. The self-ownership tradition of Murray Rothbard (1982) and the nonaggression principle place the burden of justification squarely on the party who initiates force, not on the person who merely persists in living and acting. Jan Narveson (1988) reaches a kindred conclusion by a contractarian route, asking what rules agents could agree to and finding no agreement that licenses one party to dispose of another for aggregate gain. Loren Lomasky (1987) is especially apposite: His account of persons as project pursuers, each with ends that structure a life from the inside, is close in spirit to the conception of the agent as an end setter that does the work in the section “Rational Agency and the Generation of Ought-Claims.” The argument of this article can be understood as giving the nonaggression intuition an internal rationale. The reason the initiator of lethal force bears the burden of justification is that force against an agent requires a reason that binds the agent, and an appeal to an end the agent does not share cannot be such a reason. Where the libertarian tradition often begins from self-ownership as a premise, the present argument arrives near it as a conclusion.

The Formal Argument

The argument is stated in full before its premises are defended. It proceeds through seven premises and three conclusions.

P1. A rational agent is one who acts on the basis of reasons, selecting means in pursuit of ends. This is instrumental, means-ends rationality, and nothing more is assumed by it.

P2. In acting for reasons, a rational agent generates normative claims of the form “I ought to ϕ ,” which are practical conclusions relative to the agent’s ends. Generating such a claim does not yet establish that any other agent is bound by it.

P3 (the bindingness condition). A normative claim “ S ought to ϕ ” binds an agent S , in the sense of giving S a reason S cannot rationally reject, only if either (a) S shares the end from which the claim is generated or (b) S ’s acting otherwise than ϕ is demonstrably incoherent. A claim meeting neither condition has authority only over those who already hold its end.

P4. *A*, *B*, and *C* are rational agents who have adopted utilitarianism and have concluded, by a correct application of it, that they ought to kill *D* for the sake of aggregate welfare.

P5. The utilitarian conclusion is generated from the end of maximizing aggregate welfare, an end that *A*, *B*, and *C* have adopted. There is no demonstrable contradiction in an agent's declining to adopt that end, and no axiom yields the prohibition on *D*'s survival unless utilitarianism is already presupposed.

C1. Therefore the claim "*D* ought to be killed" does not bind *D*. It fails condition (b) of P3, since *D*'s continuing to pursue survival is not incoherent, and it fails condition (a) of P3, since *D* need not share the end of aggregate welfare maximization. The claim is binding only on those who hold that end.

P6. *D* is a rational agent of the same kind as *A*, *B*, and *C*. Moreover, *D*'s standing as an end-having agent is the very feature in virtue of which *D*'s welfare entered the utilitarian calculus at all.

P7 (the burden of justification). To be rationally justified in killing a rational agent, as distinct from merely exercising power over him, the killer must possess a reason that binds the one to be killed. The party who initiates lethal force bears this burden.

C2. Since, by C1, *A*, *B*, and *C* possess no reason that binds *D*, they cannot discharge the burden identified in P7. They are therefore not rationally justified in killing *D* on the basis of the utilitarian prescription.

C3 (the universalizability trap). More generally, no normative framework that derives its prescriptions from the adoption of an end, rather than from the incoherence of that end's negation, can generate an obligation binding on an agent who does not share the end. Utilitarian aggregate-welfare maximization is such a framework. Its prescriptions to eliminate nonconsenting rational agents never bind those agents, and so can never justify their elimination.

Rational Agency and the Generation of Ought-Claims (P1, P2)

The first task is to be precise about a distinction on which the argument depends, and on which a careful reader is right to insist: the distinction between two things, both loosely called "normative."

P1 defines rational agency in minimal, means-ends terms. A rational agent chooses an end, surveys available means, and selects among them on the basis of reasons. The definition is deliberately thin. It presupposes no account of which ends are worth having and commits us to no ethical theory. It is satisfied by the prudent thief no less than by the saint.

P2 records what such an agent does in acting. To deliberate toward an end is to reach conclusions of the form “Given this end, I ought to employ this means.” These are genuine ought-claims, but they are practical and end-relative. They are the oughts of instrumental reason, what Kant ([1785] 1998) called hypothetical imperatives. When I judge that I ought to take the earlier train, the ought reports a requirement that holds because I aim to arrive on time; it makes no claim on anyone who does not share that aim, and it makes no claim on me once I abandon the aim.

The normative distinction the argument must keep in view throughout is that to say that a rational agent generates ought-claims, in the sense just described, is to say that the agent reaches end-relative practical conclusions. It is not yet to say that the agent has identified a moral reason that obligates anyone, least of all anyone who does not share the agent’s end. The capacity to reason practically is the capacity to form judgments about what to do to reach one’s ends. It is not, by itself, the capacity to bind others. I grant this without reservation, and I want to be explicit that the argument does not need the stronger claim. Nothing in what follows derives a moral theory from instrumental rationality. The argument concerns the much narrower question of when an agent’s practical conclusion acquires authority over a second agent, and the answer to that question is given not by P1 or P2 but by P3.

It is equally important to register the converse, which an objector may rightly press. The fact that *D* can articulate the judgment “I ought to live” does not establish that *D* ought to live. Articulation is not justification. If the present argument depended on treating *D*’s assertion as self-warranting, it would deserve the objection. It does not. *D*’s assertion plays no justificatory role in the argument at all. What does the work is not anything *D* says, but the failure of *A*, *B*, and *C* to produce a claim that binds *D*. The asymmetry the argument relies on, defended in the section “Parity and the Burden of Justification,” is not an asymmetry of assertion. It is an asymmetry of justificatory burden.

When an Ought-Claim Binds Another Agent (P3)

P3 is the load-bearing premise, and it deserves a careful statement and defense. The question it answers is this. An agent has reached a normative conclusion. Under what conditions does that conclusion bind a second agent, in the strong sense of giving the second agent a reason he cannot rationally reject?

Begin with the uncontroversial half of the answer. A claim binds an agent if the agent shares the end from which the claim is generated. If you and I both aim to finish the proof, and taking a certain lemma is necessary to finish it, then “you ought to take the lemma” binds you, because you have the end relative to which the ought holds. This is condition (a). It explains the ordinary force of instrumental oughts among agents who share purposes, and it explains why

such oughts lapse for an agent who abandons the purpose. Condition (a) is just the observation that hypothetical imperatives bind those who hold the relevant hypothesis.

The interesting half is condition (b). Suppose the second agent does not share the end. Can the claim still bind him? It can, but only in one way: by showing that his acting otherwise is not merely contrary to some end he happens to lack, but incoherent, something he cannot consistently will. This is the structure of a genuinely categorical requirement. What makes such a requirement binding on an agent who does not antecedently accept it is precisely that its denial is not a coherent option for him. He is bound not because he has been told to comply, and not because compliance would serve some aim of his, but because the alternative is closed to him as a rational matter. The law of noncontradiction is binding in this way; so, if Kant is right, is the prohibition on maxims that cannot be universalized without contradiction. The general principle is that an ought binds an agent independently of his ends only where the contrary is rationally foreclosed.

Why accept that these are the only two routes? Because any third route would have to bind an agent neither through an end he holds nor through the incoherence of the alternative, and it is obscure what such a route could be. A bare assertion that “*S* ought to ϕ ” does not bind *S*; assertion is not authority. An appeal to a fact that *S* has no rational stake in does not bind *S* either, for we may always ask why that fact gives *S* a reason, and unless the answer cites an end of *S*’s or a contradiction in *S*’s alternative, the question simply recurs. This is a descendant of David Hume’s ([1739] 1978, 469–70) observation that no description of how things are, by itself, entails what any particular agent ought to do; the gap is closed only by an end the agent has or by a requirement of consistency the agent cannot escape. Korsgaard’s (1996) normative question presses the same point from the side of the obligated agent: An obligation has authority over me only if I cannot coherently distance myself from it—that is, only if rejecting it is not a live option for me.

It is worth emphasizing how modest P3 is. It is not a moral principle. It does not say that one ought never coerce, or that aggregate welfare does not matter, or that any particular thing is valuable. It is a condition on a relation, the relation of binding, that any party who claims that his conclusion obligates a second party must satisfy. A utilitarian can accept P3, just as he can accept the law of noncontradiction, while remaining a utilitarian. What he cannot do, having accepted P3, is maintain that his verdict obligates an agent who neither shares his end nor faces incoherence in declining it. The section “This Is Just Kantianism, or a Smuggled Constructivist Theory of Bindingness” takes up the objection of the reader who wishes to reject P3 outright by holding that moral facts bind regardless of the agent’s rational constitution; I argue there that even this reader inherits the difficulty in a new form.

The Utilitarian Prescription Does Not Bind *D* (P4, P5, C1)

With P3 in hand, the central step is short. We grant *A*, *B*, and *C* everything they ask for short of the conclusion. We grant that they are rational, that they sincerely hold utilitarianism, and that they have applied it correctly to reach the verdict that they ought to kill *D*. The question is whether that verdict binds *D*.

Consider the source of the verdict. It is generated from the end of maximizing aggregate welfare. That end is one *A*, *B*, and *C* have adopted; for purposes of the present argument, its adoption is a substantive commitment rather than something required by logic alone. And here is the decisive observation, the one that does the work. There is no axiom available independent of utilitarianism itself from which a contradiction follows upon *D*'s continuing to live and to pursue his survival. *D*, who declines to be killed, is not thereby committed to anything inconsistent. He has not contradicted the law of noncontradiction, nor any principle he cannot coherently reject. He has simply declined to adopt an end, the maximization of aggregate welfare across all persons, including himself as a mere summand, that he was never under rational compulsion to adopt. The only way to manufacture a contradiction in *D*'s survival is to feed the utilitarian end back in as a premise, which is to assume precisely what is in question.

So the verdict fails condition (b) of P3. *D*'s acting otherwise than the verdict prescribes is not incoherent. Does it satisfy condition (a)? Only if *D* shares the end. But *D* need not and, in the case before us, plainly does not. A coherent rational agent can decline to value the aggregate over his own continued existence without any failure of rationality; nonutilitarian agents are not thereby irrational, only nonutilitarian. The verdict therefore fails condition (a) as well.

The conclusion, C1, follows. "*D* ought to be killed" does not bind *D*. This is not a claim that the verdict is false, nor a claim that utilitarianism is incoherent. The verdict may be perfectly correct as an output of the framework, and binding on every agent who has adopted that framework's end. What it lacks is authority over the one agent who has not, and who faces no incoherence in withholding his assent. The verdict is, in the end, an ought-claim that *A*, *B*, and *C* have generated in light of the end they hold and the experience and values that led them to hold it. As such, it governs their own conduct and the conduct of their fellow utilitarians. It does not reach across the framework's boundary to an agent who stands outside the end, for the simple reason that they have not shown, and cannot show without circularity, that his standing outside it is a rational mistake.

Parity and the Burden of Justification (P6, P7, C2)

The reader who has followed the argument to C1 may still press an objection that in its sharpest form runs as follows. Granting that the utilitarian verdict does not bind *D*, why should we not say with equal warrant that *D* has no

binding claim to live, so that the parties are simply at an impasse, each generating a verdict that fails to bind the other? Why does the failure of bindingness of the killing defeat the killing rather than leaving the matter open? This is the question on which the argument turns, and answering it requires P6 and P7.

Take P6 first. *D* is a rational agent of the very same kind as *A*, *B*, and *C*. This is not a moral premise smuggled in; it is forced on the utilitarian by his own theory. The reason *D*'s welfare figures in the calculus at all is that *D* is an end-having being for whom things can go well or badly (i.e., a rational agent in the sense of P1). The utilitarian cannot treat *D* as a full agent for the purpose of counting his welfare but as a mere obstacle for the purpose of dismissing his standing. The same feature that admits *D* into the calculus secures his parity with those who perform it.

Now take P7, which breaks the supposed impasse. The parties are not symmetrically situated with respect to justification, because they are not proposing symmetrical acts. *A*, *B*, and *C* propose to kill. *D* proposes only to go on living. These are not on a par, and the difference is not moral favoritism but a structural feature of justification itself. To go on living and pursuing one's ends is the default condition of any agent; it requires no license because it imposes no requirement on anyone else that stands in need of vindication. To kill an agent is to override his agency by force, and an act of that kind requires a justification addressed to the one who suffers it. The burden of justification falls on the party who initiates force, a principle that the libertarian tradition (Rothbard 1982, chap. 12, esp. 78) states as the presumption against aggression and that is in fact a quite general feature of practical reason: The one who would constrain or destroy another's agency must be able to give that other a reason, whereas the one who merely continues to act within his own sphere need give no one any reason.

This is the asymmetry the argument needs, and it is earned rather than assumed. *D* does not have to produce a verdict that binds *A*, *B*, and *C*. *D* is not trying to bind them; he is declining to be bound by them. The question is never whether *D* can justify his survival to *A*, *B*, and *C*. The question is whether *A*, *B*, and *C* can justify killing *D* to *D*, since they are the ones who would act on him. By C1 they cannot, because the only reason they have to offer, the utilitarian verdict, does not bind him. The impasse the objector imagines is illusory because it pictures two parties each owing the other a justification. In truth only one party owes a justification, the party who would kill, and that party cannot pay.

C2 follows. *A*, *B*, and *C*, lacking any reason that binds *D*, cannot discharge the burden of justification that the killing of a rational agent imposes. They remain free, of course, to kill *D* as a matter of brute power. What they cannot do is represent the killing as rationally justified, for to do so is to claim an authority over *D* that they have been shown not to possess. The killing, if it occurs, is

not the conclusion of a valid practical inference binding on all parties. It is the imposition of the ends of some agents upon another agent whose equal standing they are committed, by their own theory, to acknowledging.

The Universalizability Trap: A Unified Statement (C3)

The general principle can now be stated. No normative framework that derives its prescriptions from the adoption of an end, rather than from the incoherence of that end's negation, can generate an obligation binding on an agent who does not share the end. Where such a framework prescribes the elimination of a rational agent who does not share its end, the prescription fails to bind that agent and so cannot justify the elimination.

The trap has four conditions. First, the framework must derive its prescriptions from an adopted end rather than from a requirement of consistency, so that an agent can decline the end without incoherence. Utilitarianism satisfies this condition, since the maximization of aggregate welfare is one end among possible ends and its rejection is not a contradiction. Second, the prescription must be directed at a rational agent, one capable of holding or declining ends, rather than at a being with no standing to be bound or unbound. Third, the agent at whom the prescription is directed must not in fact share the end, for if he shares it the prescription binds him by condition (a) of P3. Fourth, the prescription must be one whose execution overrides the agent's agency by force, so that the burden of justification in P7 is triggered.

When the four conditions are met, the trap closes. The utilitarian has invoked a framework whose authority extends only to those who hold its end. He has directed its lethal prescription at an agent who does not hold the end and who can decline it without contradiction. He has thereby undertaken to justify killing that agent, and he finds he has no reason that reaches the agent. The framework's prescriptive power, which is real for its adherents, simply does not extend across the boundary it is being asked to cross. It is not that *D*'s welfare wins the calculation. It is that the calculation was never addressed to anyone who could bind *D*, and so *D* is not bound.

Objections and Replies

The argument invites several objections concerning its account of normative authority, the role of instrumental rationality, and its relation to competing approaches in moral philosophy. This section addresses the principal objections and clarifies the argument's scope, assumptions, and implications.

You Still Derive Morality from Instrumental Rationality

The first objection holds that the argument, despite its disclaimers, must be deriving moral conclusions from the bare capacity for means-ends reasoning, since that capacity is all P1 grants the parties. This would indeed be an extraordinary claim, and one no utilitarian need accept.

The reply is that the argument derives no morality from instrumental rationality, and does not need to. P1 and P2 establish only that agents form end-relative practical conclusions. The moral work, such as it is, is done by P3, which is not a thesis about instrumental rationality at all but a condition on the relation of binding, and by P7, which is a condition on the justification of the initiation of force. Neither premise says that an instrumentally rational agent is thereby moral, or that morality follows from agency. Indeed, the argument is officially neutral on where utilitarian or any other moral commitments come from. It grants the objector's own point: Utilitarians ground their view in substantive considerations about welfare and interests, not in instrumental rationality. The argument simply observes that whatever those considerations are, and however well they support the verdict for those who accept them, they bind an outside agent only if that agent shares the relevant end or faces incoherence in declining it. The objection, in short, attacks a derivation the argument does not attempt.

Articulating a Norm Is Not the Same as Its Being True

A second objection, closely related, observes that *D*'s ability to say "I ought to live" does not make it true that he ought to live, and warns against any argument that treats the victim's assertion as self-justifying.

This is correct, and the argument honors it. *D*'s assertion is given no justificatory weight. The argument does not say that *D* ought to live because *D* says so, nor that *D*'s counterclaim is true. It says something that does not concern the truth of *D*'s assertion at all: that *A*, *B*, and *C* have failed to produce a claim that binds *D*, and that the burden of producing such a claim is theirs because they are the ones who propose to kill. *D*'s correctness in asserting that he ought to live is simply not a premise. The pivot is the silence on the other side, the absence of any reason that reaches *D*, not the content of anything *D* announces.

You Have Begged the Question; Why Not Say D Lacks a Justification to Live?

The most serious objection grants that the utilitarian verdict does not bind *D* but contends that this yields only an impasse. For one might argue with equal right that, given the utilitarian verdict, *D* lacks any justification for continuing to live. The argument is then accused of helping itself to an asymmetry, treating the utilitarian derivation as defective while treating *D*'s position as sound, when nothing has earned that difference.

The section "Parity and the Burden of Justification" is the reply, and it can be put compactly here. The asymmetry is not assumed; it follows from P7, which locates the burden of justification on the party who initiates lethal force. The objection pictures two claims, "*D* ought to die" and "*D* ought to live," competing for the same office, each failing to bind the other side—hence a

standoff. But the two claims are not competing for the same office, because the parties are not proposing the same kind of act. *A*, *B*, and *C* propose to kill, and a killing must be justified to the one killed. *D* proposes only to persist, and persistence requires no justification addressed to anyone, since it overrides no one's agency. So the question is never whether *D* can justify his life to *A*, *B*, and *C*. It is whether they can justify his death to him. The demand for a binding reason falls on them alone, and they cannot meet it. There is no symmetric counterpart demand on *D*, because *D* is not the one wielding force. Far from being assumed, the asymmetry is the direct consequence of the most ordinary principle governing the justification of harm: The agent who would constrain another by force owes that other a reason; the agent who merely goes on living owes none. This asymmetry reflects the ordinary principle that using lethal force against a person requires greater justification than leaving that person alone.

Why Does D's Standing Outweigh a Correct Calculation?

A reader sympathetic to the argument may still find its crucial step too quick, asking why, if *A*, *B*, and *C* have correctly calculated that *D* ought to be killed, *D*'s standing as a rational agent should outweigh that result. The objection deserves a fuller answer than the bare observation that *D* is a rational agent with equal standing, since that observation does not yet explain why the utilitarian calculation lacks authority over *D*. Here is the explanation.

D's standing does not outweigh the calculation, and the argument does not claim that it does. To say that *D*'s standing outweighs the calculation would be to enter the utilitarian ledger and contest the sums, which the argument explicitly declines to do. The point is that the calculation, however correct, was never authoritative for *D* in the first place. A correct calculation is a correct derivation of what follows from the utilitarian end. Its correctness is internal to the framework and binds those who hold the end. It does not, by being correct, acquire reach over an agent who does not hold the end and who can decline it without contradiction any more than a correct derivation of what I ought to do given my aim of winning a game binds a person who is not playing. So the relation between *D*'s standing and the calculation is not a contest of weights at all. The calculation produces an end-relative ought; *D* stands outside the end; the ought does not bind him. The reason this looked as though it needed a weighing is that one naturally assumes the verdict already obligates everyone and then asks what could be heavy enough to override it. But the verdict never obligated *D*, so there is nothing to override. The expansion this calls for is, at bottom, the replacement of the picture of weighing by the picture of authority and its limits.

This Is Just Kantianism, or a Smuggled Constructivist Theory of Bindingness

A reader may grant the structure of the argument but object that P3, the bindingness condition, is itself a substantive and contestable thesis, recognizably Kantian or constructivist, so that the argument is not the internal result it advertises itself to be but the imposition of a rival metaethics. A robust moral realist, in particular, may hold that moral facts bind agents, regardless of whether the agents share any end or face any incoherence, simply because the facts are there and are about what those agents have reason to do.

There are two replies. The first concedes the diagnostic observation and denies it is fatal. P3 is indeed a thesis in the theory of normative authority, and it has affinities with Kant and Korsgaard. But it is a minimal thesis, far weaker than the categorical imperative or any constructivist account of the content of morality. It says only that to bind an agent who does not share one's end, one must show that the alternative is incoherent. A utilitarian can hold this thesis and remain a utilitarian; it constrains how he may bind others, not what he may value. To reject even this thesis is to claim that one's conclusions obligate other agents neither through any end those agents have nor through any incoherence in their dissent nor through anything they could be brought to see, but simply because one has reached the conclusion. But merely reaching a conclusion does not explain why that conclusion has authority over another agent. It may explain why *A*, *B*, and *C* act on their judgment, but it does not yet explain why they are justified in imposing that judgment on *D* by lethal force. Without some account of how the conclusion binds *D*, the appeal to it supplies only a motive or rationale for the killers' conduct, not a justification addressed to the person they propose to kill.

The second reply meets the realist on his own ground. Suppose there are mind-independent moral facts that bind regardless of any agent's rational constitution, and that one of them is the utilitarian verdict against *D*. Even so, *A*, *B*, and *C* face a difficulty that is now epistemic and dialectical rather than constitutive. They claim to perceive a binding fact that obligates *D* to accept his own death. *D*, an equally competent rational agent by P6, perceives no such fact and perceives instead that he is to resist. The realist owes an account of why *A*, *B*, and *C*'s claim to discern the fact carries authority to act lethally on *D* when *D*'s symmetric claim not to discern it is, by hypothesis, the judgment of an equally reliable perceiver. The burden of justification from P7 does not dissolve under realism; it reappears as the burden of showing that the killers, rather than the killed, have read the moral facts aright, and mere confidence in one's own perception does not discharge the burden of justification owed to the person one proposes to kill. Realism changes the source of the alleged obligation; it does not supply the killers with the authority over *D* that the act requires.

This Is Just Gewirth's Argument

Because the argument moves from features of agency to a constraint on how agents may be treated, a reader acquainted with Gewirth (1978) may take it to be a restatement of the argument to the principle of generic consistency. The resemblance is real and worth marking precisely, since the differences are what protect the present argument from Gewirth's well-known vulnerabilities.

Gewirth argues that an agent, reflecting on the necessary conditions of his own action, must claim rights to freedom and well-being, and must then, on pain of contradiction, grant the same rights to all other agents. The most pressed objection targets the first move: The step from the agent's recognizing that he needs freedom and well-being to his claiming a right to them looks like a slide from a prudential "I need" to a normative "I am owed." The present argument makes no such move. It does not have *D* claim a right, and it does not derive any positive entitlement from *D*'s needs as an agent. It does not require *D* to generate a universal principle at all. The only positive premise about *D* is P6, that *D* is the same kind of agent the utilitarian already counts, which the utilitarian concedes. Everything else is negative: a condition that the killers' reason must meet in order to bind *D* (P3) and a burden the killers must discharge in order to be justified (P7). Where Gewirth builds a substantive morality of rights out of the structure of agency and pays the price of the contested ascent to rights, the present argument builds nothing positive and only denies that a particular reason has a particular reach. It is, one might say, the conservative half of Gewirth's ambition: not that agency generates rights binding on all, but only that an appeal to an unshared end cannot bind the agent it is turned against.

Why Should the One Prevail Against the Many?

A final objection notes that utilitarianism is built to aggregate, and asks why the single agent *D* should prevail against the combined verdict of three agents, or three thousand.

The reply is that aggregation operates inside the framework and cannot manufacture authority outside it. Adding more agents who share the utilitarian end produces a verdict with more adherents; it does not produce a verdict that binds an agent who is not among them. Number bears on the weight of considerations within the calculus, and the argument does not dispute those weights. It bears not at all on the prior question of whether the calculus reaches *D*, because that question turns on whether *D* shares the end or faces incoherence in declining it, and neither of these changes when the number of utilitarians grows. A unanimous community of welfare maximizers still owes the one dissenter a reason he can be bound by, and the size of the community does nothing to supply it. This is, in a different key, why the separateness of persons that Rawls (1971) and Nozick (1974) insisted on

cannot be aggregated away: The boundary between the agent who holds an end and the agent who does not is not the kind of thing that more agents on one side can erase.

Broader Implications

If the argument succeeds, several consequences follow for ethical and political theory.

The first concerns the relation between consequentialist and deontological thought. The constraint the argument reaches, that a rational agent may not be eliminated by an appeal to an end he does not share, is one deontologists have long defended on independent grounds, as a prohibition on treating persons as mere means. The present route to it is neither deontological nor a contest within consequentialism. It is a result about normative authority that a consequentialist and a deontologist can both accept, since it follows from the conditions of binding rather than from any ranking of outcomes or any catalogue of duties. To that extent it offers a point of convergence of the kind Derek Parfit (2011) hoped to locate, though by a more minimal path than his.

The second consequence concerns self-defense. Traditional accounts have grounded the permissibility of self-defense in the aggressor's wrongdoing, in the defender's interest in his own survival, or in contractualist principles about justifiable rules. The present argument adds a further and more structural ground. *D*'s resistance does not require a prior demonstration that *D*'s life outweighs the attackers' ends. It requires only the observation that the attackers have produced no reason that binds *D*, so their action against him is, with respect to him, the unjustified initiation of force—exactly the condition under which resistance has always been thought permissible. Whether this grounds every form of self-defense is a further question; that it grounds a presumptive right to resist an unbinding lethal prescription seems clear.

The third consequence, and for this journal the most pointed, concerns the authority of the state. Let *A*, *B*, and *C* be not three persons but a collective: a majority, a legislature, a state. Liberal and democratic theories standardly ground the authority of such a collective in the rational agency of the citizens who compose and consent to it. But a collective that grounds its authority in the agency of its members cannot coherently invoke that authority to eliminate a member on the strength of an end he does not share, for the member's standing as a rational agent, the very thing that grounded the collective's authority, is what places him beyond the reach of any merely end-relative prescription directed against his life. Applied at this level, the argument yields a structural constraint on any practice in which a collective claims the authority to remove an individual agent from existence for the sake of an aggregate end the individual need not endorse (e.g., capital punishment or the liquidation of

political opponents). The constraint is not that such killings always produce worse outcomes. It is that the authority invoked to perform them does not extend to them.

Conclusion

I have argued that a group of rational agents cannot be rationally justified in killing a fellow rational agent on utilitarian grounds, and that the reason is not a defect in the utilitarian calculus but a limit on the authority of any prescription generated from an adopted end. A normative claim binds an agent only if he shares the end from which it springs or cannot coherently act otherwise. The utilitarian prescription to kill *D* meets neither condition for *D*, who need not share the end of aggregate welfare and whose survival involves no contradiction. Since the killing of a rational agent must be justified to the one killed, and the only reason on offer does not bind him, the prescription cannot justify the killing. It can compel *D* in the sense that *A*, *B*, and *C* may physically force their verdict upon him, but such coercive power does not establish that the verdict has normative authority over him.

This is the universalizability trap. A framework that derives its force from an end has authority only over those who hold the end. When it is turned against an agent who stands outside the end, to destroy him, it reaches for an authority it does not have, and the reach fails not because some external morality forbids it but because the prescription generated by the framework was never shown to have authority over the one it would now bind. Humans *A*, *B*, and *C*, precisely as rational agents who must justify what they do to the agents they do it to, cannot on utilitarian grounds justify killing human *D*. Whether the result generalizes to every elaboration of utilitarianism and whether the burden-shifting account of self-defense put forward in the section “Broader Implications” can bear the full weight of a positive theory of rights remain open questions. What the argument establishes, I submit, is narrow and firm, and for that reason secure: The authority of an end-relative ought, however correct for those who hold the end, stops at the agent who does not.

Submitted: May 03, 2026 CDT. Accepted: July 28, 2026 CDT. Published: September 21, 2026 CDT.

REFERENCES

- Bentham, Jeremy. 1789. *An Introduction to the Principles of Morals and Legislation*. London. https://archive.org/details/bim_eighteenth-century_an-introduction-to-the-p_bentham-jeremy_1789/page/n1/mode/2up.
- Gewirth, Alan. 1978. *Reason and Morality*. University of Chicago Press.
- Hare, R. M. 1981. *Moral Thinking: Its Levels, Method, and Point*. Clarendon Press. <https://doi.org/10.1093/0198246609.001.0001>.
- Harsanyi, John C. 1977. "Morality and the Theory of Rational Behavior." *Social Research* 44 (4): 623–56. <https://www.jstor.org/stable/40971169>.
- Hume, David. (1739) 1978. *A Treatise of Human Nature*. Edited by L. A. Selby-Bigge. 2nd ed. Revised by P. H. Nidditch. Clarendon Press. <https://doi.org/10.1093/oseo/instance.00046221>.
- Kant, Immanuel. (1785) 1998. *Groundwork of the Metaphysics of Morals*. Translated by Mary Gregor. Cambridge University Press. <https://doi.org/10.1017/CBO9780511809590>.
- Korsgaard, Christine M. 1996. *The Sources of Normativity*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511554476>.
- Lomasky, Loren E. 1987. *Persons, Rights, and the Moral Community*. Oxford University Press. <https://archive.org/details/personsrightsrmor0000loma>.
- Mill, John Stuart. 1863. *Utilitarianism*. London. <https://archive.org/details/a592840000milluoft>.
- Narveson, Jan. 1988. *The Libertarian Idea*. Temple University Press. <https://archive.org/details/libertarianidea0000nary>.
- Nozick, Robert. 1974. *Anarchy, State, and Utopia*. Basic Books.
- Parfit, Derek. 2011. *On What Matters*. 2 vols. Oxford University Press. <https://doi.org/10.1093/acprof:osobl/9780199572809.001.0001>.
- Rawls, John. 1971. *A Theory of Justice*. Harvard University Press. <https://doi.org/10.2307/j.ctvjf9z6v>.
- Rothbard, Murray N. 1982. *The Ethics of Liberty*. Humanities Press.
- Scanlon, T. M. 1998. *What We Owe to Each Other*. Harvard University Press. <https://doi.org/10.2307/j.ctv134vmrn>.
- Singer, Peter. 1993. *Practical Ethics*. 2nd ed. Cambridge University Press.