

The Transcendental Necessity of Morality

JOSEPH HEATH
University of Toronto

David Gauthier tries to defend morality by showing that rational agents would choose to adopt a fundamental choice disposition that permits them to cooperate in prisoner's dilemmas. In this paper, I argue that Gauthier, rather than trying to work out a prudential justification for his favored choice disposition, should opt for a transcendental justification. I argue that the disposition in question is the product of socialization, not rational choice. However, only agents who are socialized in such a way that they acquire a disposition of this type could acquire the capacity to use language. Given the internal connection between language and thought, this means that no agent endowed with such a disposition could rationally choose to adopt another. Thus rational reflection by moral agents upon their own fundamental choice disposition will have no tendency to destabilize it.

"It is a necessary truth that people tend to do what they think they ought to do, for it is a necessary truth that people who occupy a linguistic position which means *I ought to do A now*, tend to do A. If they did not, the position they occupy could not mean *I ought to do A now*."

Wilfrid Sellars, "Some Reflections on Language Games."

One of the central tasks of the Western philosophical tradition since Socrates has been to provide a rational vindication of the claims of morality. The reason that morality has been felt to require some kind of vindication is that its demands are often in tension with those of self-interest. Under such circumstances, it is naturally tempting to disregard the former in favour of the latter, and it is very difficult to explain to someone who intends to do so precisely why he or she should not. Since just saying that it is "wrong" to disregard one's moral obligations is question-begging, philosophers have generally sought some more indirect justificatory strategy. The most common has been to attempt to show that it is, in some subtle sense, not in one's *interest* to act immorally. This can take two very general forms. The first is the route of straightforward reduction—to argue that, at the end of the day, one's interests will be better served if one acts morally. The second strategy is to

split the difference—to show that one's interest, properly understood, includes specifically moral concerns.

While no one has been able to execute either of these two argumentation strategies successfully, many people have raised even more fundamental doubts about the framework of the discussion. In particular, H.A. Prichard's famous essay "Does moral philosophy rest on a mistake?" suggested that the entire approach is self-defeating.¹ If anyone ever succeeded in showing that it was in one's interest to respect a certain class of moral obligations, then this would *undermine their status as moral obligations*. We would no longer commend people for performing these actions, once their ultimately self-serving character was revealed. Unfortunately, even those who feel the force of this observation have tended to shy away from its full implications. In part, this is because Prichard seems to empty the philosophical arsenal of all its weapons. If one cannot appeal to moral considerations in defense of duty without begging the question, and one cannot appeal to self-interest without undermining the moral status of these duties, then what exactly is left? The only response to moral skepticism seems to be the kind of complacent intuitionism that Prichard, as a matter of fact, endorsed.

One strategy, however, which has been largely overlooked in contemporary philosophical discussion is the transcendental justification. This basic Kantian argument form is widely used in attempts to defuse epistemological skepticism, but has not received widespread attention among moral philosophers.² The basic transcendental strategy is not to refute the skeptic directly, but rather to neutralize skeptical doubts by showing that they are cognitively inaccessible. This can be achieved by showing that while these doubts appear plausible at first glance, upon closer examination they turn out to violate certain conditions of possibility of thought. So while we have no *particular* reason for thinking the way that we do, the alternatives are all demonstrably incoherent. The equivalent argument in the case of *moral* skepticism would take a slightly different form. Since the moral skeptic is making a pragmatic, rather than a theoretical, recommendation, a transcendental justification of morality would show that the skeptic's recommendations are pragmatically inaccessible to us. We have no *particular* reason to be moral agents, but given that we are, we cannot coherently choose not to be.

¹ H.A. Prichard, *Moral Obligation* (London: Oxford University Press, 1968).

² The contemporary discussion in epistemology takes as its point of departure Barry Stroud's, "Transcendental Arguments," *Journal of Philosophy*, 65 (1968): 241-256. See also Peter Bieri and Rold P. Horstmann, eds. *Transcendental Arguments and Science* (Dordrecht: D. Reidel, 1979). On the use of transcendental arguments in moral philosophy, see A.J. Watt, "Transcendental Arguments and Moral Principles," *Philosophical Quarterly*, 25 (1975): 40-57. Watt surveys a sets of arguments, all of which fail because they attempt to provide a transcendental justification of substantive moral principles. The argument to be presented here is different, in that it attempts to justify a purely formal choice disposition.

The goal of this paper is to present an argument of this form. I begin by adopting the general framework for the analysis of practical rationality suggested by David Gauthier in his *Morals by Agreement*.³ I then try to show how Gauthier, rather than trying to work out a prudential justification for his favored "fundamental choice disposition," could opt for a transcendental justification. After a brief digression on the structure of transcendental philosophy, I present an argument of this form to support Gauthier's favored disposition. I argue that the disposition in question is the product of socialization, not rational choice. However, only agents who are socialized in such a way that they acquire a disposition of this type could acquire the capacity to use language. Given the internal connection between language and thought, this means that no agent endowed with such a disposition could rationally choose to adopt another. As a result, rational reflection by moral agents upon their own fundamental choice disposition will have no tendency to destabilize it, even though this disposition is not one that they have rationally chosen.

1. Definition of the problem

The primary advantage of Gauthier's conceptual framework is that he does not succumb to the temptation to fudge the distinction between interest and duty, or between *instrumental rationality* and *normative constraint*. He takes as his paradigmatic interaction form the classic prisoner's dilemma, which brings out this contrast quite sharply. In an interaction of this type, two individuals are in a position where they can engage in mutually beneficial cooperation, but where they can also exploit each other's willingness to cooperate. To take a typical sort of example: one farmer asks his neighbor to help him clear heavy rocks from his field, and offers, in return, to help drain a swamp that is encroaching upon the other's. Unfortunately, the two activities cannot be performed simultaneously, and so farmer two, in order to accept such an offer, must believe that, after all of the rocks have been cleared, farmer one will still be willing to come over and drain the swamp. The problem is that farmer one has an obvious incentive to renege on the agreement at this point. Thus "he which performeth first, does but betray himselfe to his enemy," as Hobbes put it.⁴ Since the second farmer has no reason to expect cooperation from the first, he in turn has no incentive to help clear the rocks. The result is a failure of cooperation.

What is so tantalizing about this situation, from the moral philosopher's perspective, is that it is obviously in the interest—broadly construed—of both farmers to be able to trust one another, because this is what would enable them to cooperate. But it is also in their interest to break this trust, whenever they are able to do so with impunity. So the only way that the two

³ David Gauthier, *Morals by Agreement* (Oxford: Clarendon Press, 1986).

⁴ *Leviathan*, ed. Richard Tuck (Cambridge: Cambridge University Press, 1991), p. 96 [68].

will be able to cooperate is if they are able to make a credible commitment that they will, under the correct circumstances, resist performing actions that it is in their interest to perform. It is when the time comes to make good on these commitments that the tension between morality and self-interest shows up especially clearly. If the two farmers enter into an agreement, and the second in fact helps to clear the rocks, this leaves the first with two options: "keep my word, and help drain the swamp," or "break my word, and don't help with the swamp." He is clearly under a moral obligation to keep his word, but it is in his interest to break it.

One of the things that examples of this type help illustrate is the fact that an integral component of moral agency is the ability to act in a way that is contrary to one's self-interest, narrowly construed. To use some modern jargon, we can call this the *capacity for counter-preferential choice*. The other thing that examples of this type help to bring out is that morality, so defined, is not empty "rule-worship." There is an important sense in which the farmer, faced with the choice of breaking or keeping his word, has no instrumental reason to keep it, since (*ex hypothesi*) no adverse consequences follow from breaking it. If he does keep his word, it will not be because doing so serves any further purpose. Thus the moral obligation, if it is to be respected, is to be respected for its own sake. At the same time, the capacity to act in this way—to make counter-preferential choice—is obviously *functional* for human societies, since it is a necessary feature of cooperative action. So while there are many things about our capacity for exercising moral restraint that are poorly understood, the simple fact that we possess such a capacity should not be entirely mysterious.

On the basis of such examples, Gauthier distinguishes between two different kinds of "choice disposition" that agents could have.⁵ The first is straightforward maximization. Agents of this type simply maximize expected utility in the usual fashion. Since these agents never make counter-preferential choices, they always defect in one-shot prisoner's dilemma-type interactions. The second type of choice disposition is constrained maximization. Agents of this type adopt a *plan* and, once they have done so, stick to it.⁶ This may lead them to make counter-preferential choices, because opportunities to re-optimize may arise as the plan unfolds which they ignore. So if the two farmers are both constrained maximizers, they can each adopt a plan that calls for helping the other, then proceed to carry out these plans. At the point where the rocks have been moved, and the first farmer is tempted to break his word, if he does show up to drain the swamp it will not be because he prefers

⁵ *Morals by Agreement*, pp. 157-189.

⁶ This way of formulating the position does not appear in *Morals by Agreement*, but in later work. In particular, see David Gauthier, "Assure and Threaten," *Ethics* 104 (1994): 690-721.

that outcome. It will be because he is "sticking to the plan." So what the constrained maximizer does, in effect, is accord priority to a certain kind of reason for action, one which prescribes actions directly, without direct reference to their outcome.

Once the problem has been set up in this way, Gauthier proceeds to relocate the issue of moral skepticism slightly. The central question, he claims, is not how one should act in any particular set of circumstances, but rather what *type of agent* one should want to be, what type of fundamental choice disposition one should adopt. Here he is formulating the problem in a way that has recognizably Kantian ancestry. For Kant, the agent can allow his will to be determined by either categorical imperatives or hypothetical imperatives. The central question is what priority should be assigned to these two "incentives."⁷ Kant conceives of the agent's *Gesinnung*, or character, as determined by a fundamental choice to assign priority to one of these two incentives. However, unlike Kant, who could see that a choice of this type has a very peculiar existential character, Gauthier treats it as a straightforward decision problem. Which fundamental choice disposition, he asks, would an instrumentally rational, i.e. self-interested, agent choose? He concludes that because being a constrained maximizer makes one trustworthy, it generates opportunities for cooperation that are unavailable to the straightforward maximizer. As a result, choosing this choice disposition would be utility-maximizing over the long run.⁸ Thus morality and self-interest are reconciled, he claims, because he has shown that it is in one's interest to *become a moral agent*, i.e. to adopt constrained maximization as one's fundamental choice disposition.

There are a couple of peculiar things about this argument. The first, and perhaps less interesting, is that it fails. Gauthier simply is not able to show that it is in one's interest to become a constrained maximizer. The most he is able to show is that it is in one's interest to *appear* to have such a disposition, in order to induce others to cooperate. This then leaves open the option of defecting when this can be done with impunity.⁹ But the second unusual thing about the argument is the underlying assumption that one can *choose* a fundamental choice disposition. In particular, Gauthier takes it as unproblematic that one can suspend one's capacity for rational reoptimization at will. This is highly suspect. For instance, a person who was able to do

⁷ Immanuel Kant, *Religion within the Limits of Reasons Alone*, trans. Theodore M. Greene and Hoyt H. Hudson (New York: Harper & Row, 1960), p. 32.

⁸ *Morals by Agreement*, pp. 170-4.

⁹ See Joseph Heath, "A Multi-Stage Game Model of *Morals by Agreement*," *Dialogue* (1996): 529-52. At the risk of making left-handed compliments, I would say that this failure is Gauthier's greatest achievement. In my view, Gauthier gets as close as it is possible to get to reducing morality to self-interest. The fact that his argument does not work is grounds for abandoning the entire approach. In this respect, *Morals by Agreement* is the ultimate *reductio* of the philosophical tradition from which it springs.

this could never be successfully threatened, blackmailed, or extorted. The fact that most of us are vulnerable to these measures suggests that we cannot just "turn off" our capacity for case-by-case rational choice because it is utility-maximizing in the long run to do so. Thus the fact that one might want to be a constrained maximizer does not mean that one can simply choose to become one.

This problem has been widely overlooked in the discussion of Gauthier's project. The reason for this, I suspect, is that we already recognize ourselves in Gauthier's characterization of constrained maximization. The choice disposition that he sketches is not an abstract possibility, it is a reconstruction of our existing inclinations. In this respect, Gauthier's disposition looks like a plausible option because it articulates an important element of our moral phenomenology. We often keep promises, just because we have made them, and we often tell the truth, just because we are supposed to. We already are constrained maximizers. Thus there is an element of "thrownness" in the problem of fundamental disposition that Gauthier overlooks. The question is not whether we should become moral agents, but whether we should want to change our existing choice disposition. These two questions may sound similar, but as we shall see, there are important differences between them.

2. Morality

Morality has been characterized so far very broadly as the capacity to assign reasons that prescribe actions directly (or "categorically") deliberative priority over reasons that prescribe actions as a means to some further end. It has been suggested that, as a matter of fact, we do possess such a capacity. Setting aside the deeper question of *whether* we should have it, we can begin by addressing the empirical question of *why* we have it.

This question can be approached by means of an analogy. In *The Possibility of Altruism*, Thomas Nagel makes the important observation that not just moral reasoning, but also prudential reasoning may lead the agent to make counter-preferential choices.¹⁰ Acting prudentially requires that we exercise foresight. Often this will mean anticipating desires that we do not yet have, and taking actions necessary to ensure that these desires can be satisfied at the time when they do arise. A prudent agent is therefore one who has the capacity to assign reasons that arise from the anticipation of future desires priority over reasons that arise from present desires. (For instance, an agent may override her current desire to sleep in order to go search for food, not because she is currently hungry, but because she foresees that eventually she will be.)

It is not hard to imagine an explanation for the fact that we have such a capacity. The sort of planning horizon that we have is clearly influenced by aspects of our biology, and is the focus of extensive socialization. First of

¹⁰ Thomas Nagel, *The Possibility of Altruism* (Oxford: Clarendon Press, 1970), pp. 33-46.

all, on the biological front, our capacity to feel the force of future events shows clear signs of evolutionary adaptation. We have a good intuitive grasp of "medium-sized" chunks of time—we can easily imagine what it is like for a week to pass, we can imagine with some difficulty what it is like for a decade to pass, and we have no intuitive capacity to grasp the passage of an eon.¹¹ Similar constraints apply to our attention span, our patience, and our capacity to defer gratification. On top of these general parameters established by our biology, socialization plays an important role in determining the specific planning horizon that we settle upon (usually to extend it—we regard the inability to defer gratification as a typical sign of immaturity). The important point is that our capacity to defer gratification, although subject to some self-control, is largely something that we experience as given. It is, in any case, not something that we can just choose to alter at will (or in the more precise German terminology, it is not something over which we exercise *willkürliche Freiheit*.) Most of us cannot blithely ignore the future and live in the moment. This is why we require extensive "technologies of the self" and mind-altering substances in order to relax.

It is not hard to see why this should be so. Humans need more than one decade to achieve sexual maturity, and traditionally did not live for more than around three decades. Individuals with extremely short planning horizons, relative to this frame, would have perished long ago, as would any individuals inclined to defer satisfaction of their desires much beyond actuarially justifiable limits.

The analogy with morality is, I would argue, quite close. Regardless of whether we are naturally disposed toward cooperative action, it is no doubt an important element of our socialization that we acquire the capacity to subordinate our more immediate inclinations to the general set of social norms that govern our interactions. Furthermore, most of us experience this disposition, once acquired, as involuntary. We are not able to disregard moral considerations at will, and when we do, it is usually because we have found some way to "rationalize" our conduct—hence the phrase "hypocrisy is the homage that vice pays to virtue." And again, it is not hard to see why this should be so. Gauthier's focus on prisoner's dilemma-type interactions is salutary in that it reminds us of the evolutionary reasons why humans might be disposed to subordinate satisfaction of their immediate preferences to rules that prescribe specific actions. This capacity creates the possibility of beneficial forms of

¹¹ Richard Dawkins notes the irony that our "natural" incredulity toward evolutionary theory is itself plausibly explained as a product of evolution. We are not equipped to imagine the passage of billions of years of time, and so have difficulty determining whether the evolutionary account of our origins is probable or not. *The Blind Watchmaker* (New York: Norton, 1986).

cooperation.¹² His mistake is to think that these benefits provide a reason for choosing the disposition, instead of just an explanation for the fact that we have it.

It is important at this point, however, not to fall into the opposite sort of error. Evolutionary theory and developmental psychology might give us an explanation for the fact that we are constrained maximizers, but it does not offer any sort of justification for this disposition. Even if evolution does account for the brute fact that we are moral agents, this remains, from the standpoint of rationality, neither here nor there. Just as we might try to overcome our tendency to worry about the future, we might also choose to deploy "technologies of the self" to overcome our subservience to moral imperatives (read a lot of Nietzsche, etc.). If we take moral skepticism to be a hyperbolic version of this option, then the problem might be formulated as follows: "What would you do if you could take a pill that would undo your socialization, and turn you into a straightforward maximizer?"¹³

This might seem to take us back to where we started, with Gauthier's original choice problem. But it does not. What this analysis of the problem invites us to ask is not whether constrained maximization is choiceworthy, but whether it is *reflectively stable*. The evolutionary perspective is needed simply to set up the question in this, the correct manner. If we were in a position where we had to choose between two alternatives in the abstract, then it would be question-begging to proceed from the assumption that we are constrained maximizers. But evolutionary theory justifies this assumption, allowing us to take as our point of departure the perspective of an agent who already has a constrained choice disposition. This difference in starting-point will turn out to have significant philosophical implications.

3. Transcendental philosophy

This is the point at which the transcendental argument can be initiated. To see how this will go, it is helpful to recall the circumstances under which Kant pioneered this argumentation form. Kant's most significant argument of this type concerned our conception of the physical world as a causal nexus. Hume, it will be recalled, pointed out that observation alone is insufficient to give us a very rich conception of causality. All that we ever see, he argued, is a series of discrete events. The idea that there could be any underlying connec-

¹² For a nice defense of this argument, see the first half of Elliott Sober and David Sloan Wilson, *Unto Others* (Cambridge, MA: Harvard University Press, 1998). Here and in the discussion that follows, I ignore the question of whether the rules that secure cooperation take the form of "plans" that have been chosen by individuals, or "social norms" that are merely adopted from the social environment.

¹³ For a sharp formulation of this problem in these terms, see Geoffrey Sayre-McCord, "Deception and Reasons to be Moral," *American Philosophical Quarterly* 26 (1989): 113-22.

tion between them, much less one which would allow us to project the outcome of future interactions, is not something that experience alone can furnish. He concluded, on these grounds, that our idea of a causal connection arises only from a certain habit of mind. Having seen events unfold in a certain sequence, he argued, we develop a tendency to expect the same sequence again under similar circumstances. This is the way we are inclined to think, and there is no reason that other people should not think differently. And if we encountered someone who didn't have this particular habit of mind, there isn't much that we could do to recommend it to them.

Kant responds to this argument by first granting the core of the "psychologist's" thesis. Causal relations are not something that, strictly speaking, we perceive; they are something that we "read into" experience. This does not entitle us, however, to regard them as arbitrary, or as merely a habit of mind. This is because, Kant claims, we would not be able to have a perceptual experience of an object if we did not also conceptualize it as something that fits into a causal nexus. So while we "happen" to treat objects as though they were causally connected, there is nothing arbitrary about this, since we would not be able to perceive them at all if we did not do so.

The argument that purports to establish this conclusion is the notoriously obscure transcendental deduction.¹⁴ The details of this particular argument are not especially important here, it is the form that is of interest. The transcendental deduction does not attempt to justify directly our imputation of a causal ordering to events (i.e. it does not provide us with a reason why we should do so), and it is certainly not designed to convince someone who doesn't have this structure of mind that he should acquire it. In this respect, the transcendental deduction is not really a justification of our claims about causality.¹⁵ The way that Kant develops it, it is simply a way of disarming a certain sort of philosophical anxiety. He is claiming, in effect, that even if we can't justify the way things are, the alternative cannot be coherently conceptualized, and so we don't have to worry about it. Thus the task of philosophical justification is supplanted by the critique of metaphysics—"metaphysics" here denoting the temptation to speculate about what might happen under inconceivable circumstances.

The conclusion of Kant's argument can be clarified by reconstructing it within the framework of contemporary modal semantics. It is common these days to understand modal operators—necessity, possibility, impossibility—as

¹⁴ Immanuel Kant, *Critique of Pure Reason*, trans. Norman Kemp Smith (New York: St. Martin's Press, 1929), pp. 151-60 [B129-43].

¹⁵ Kant writes, "This peculiarity of our understanding, that it can produce *a priori* unity of apperception solely by means of the categories, and only by such and so many, is as little capable of further explanation as why we have just these and no other functions of judgment, or why space and time are the only forms of our possible intuition." *Critique of Pure Reason*, p. 161 [B 146].

a set of restricted quantifiers over possible worlds. They are restricted by an implicit accessibility relation. Thus to say that *p* is necessary is to say that *p* is true at all possible worlds accessible to our own. Different accessibility relations then produce different concepts of necessity. If all worlds with the same laws of logic as our own are considered accessible, then this provides the notion of logical necessity. If all worlds with the same laws of physics as our own are considered accessible, then this provides the notion of physical necessity. Within this framework, transcendental necessity can be introduced simply by defining a new accessibility relation. According to this view, a proposition is transcendently necessary if it is true at all possible worlds cognitively accessible to our own.

If we think that the limits of what can be coherently conceptualized are determined only by the laws of logic (i.e. anything non-contradictory can be conceived), then this transcendental accessibility relation will be redundant. But for Kant, this would be true only of a purely "discursive" intellect (i.e. God). As corporeal beings, we are restricted in what we can perceive. This imposes a broadly verificationist constraint on what we can conceive, which in turn makes the notion of cognitive accessibility much narrower than that of logical accessibility. Thus the set of cognitively accessible possible worlds are those containing states of affairs which could be objects of possible intuition (i.e. which could be perceived). The transcendental deduction attempts to show that a world in which there are no causal connections between events, while logically possible, is not transcendently possible (because states of affairs in it could not be perceived, given the kind of mental equipment we have). Because our system of perception requires us to conceive of objects as causally linked, the existence of such connections is true at all possible worlds cognitively accessible to our own, and so it is transcendently necessary.

While Kant was primarily interested in the constraints that the structure of perception imposes on conceptualization, the "linguistic turn" drew attention to the role that language plays in constraining the range of conceptualizable states of affairs. With Wittgenstein came the recognition that in order for a state of affairs to be cognitively accessible to us, it must be possible for us to say what that state of affairs consists in. This is the idea underlying his claim that "the limits of *language* are the limits of *my world*."¹⁶ But not just

¹⁶ *Tractatus Logico-Philosophicus*, trans. D.F. Pears and B.F. McGuinness (London: Routledge, 1961), p. 57 [§5.62]. For both Kant and Wittgenstein, this distinction is the key to the critique of metaphysics. For both theorists, metaphysics starts when we attempt to make claims about what happens at possible worlds that are not cognitively accessible to our own. For Kant, this takes the form of treating noumena as phenomena. For Wittgenstein, it means trying to say something about that which can only *show* itself. In both cases, the key is to restrict one's speculations to the set of cognitively accessible possible worlds (for Kant, those which satisfy the conditions of possible experience, i.e. phenomena, and for Wittgenstein, those which can form the contents of propositions).

anything can be said. Certain constraints must be satisfied in order to make an intelligible statement. As a result, many philosophers began to suspect that the question of which possible worlds are cognitively accessible to our own would be best answered by developing a theory of meaning.

One immediate consequence of this view is that any conditions which must be satisfied in order for language to function correctly will be transcendently necessary. To take one example, Donald Davidson has argued that the interpretations we give to one another's linguistic behavior are severely underdetermined by the evidence available to us.¹⁷ Any particular utterance can be interpreted in a variety of different ways, simply by varying the beliefs that we ascribe to the person who uttered it. And since these beliefs are propositional attitudes, the content of these beliefs can be varied by changing the interpretations that we give to these sentences, and so on. As a result, the only way that we can possibly understand one another is if we privilege one of these interpretations. Davidson argued that we do so by selecting the ascription of meaning and belief that maximizes the number of true beliefs held by that individual—this is the famous "principle of charity."

It is a consequence of the principle of charity that belief is intrinsically veridical. In order to ascribe a set of predominantly false beliefs to an individual, one would have to interpret this person uncharitably (since it is always possible to make more of these beliefs come out true by changing one's assumptions about what the person means by what she says). But once the principle of charity is abrogated, there is no longer much left to go on in constructing an interpretation. People can be interpreted as saying or believing pretty much anything at all. This makes it impossible to figure out what the contents of these beliefs are, and as a result, gives us no reason to ascribe contents to them in the first place. Thus a world in which people have predominantly false beliefs is not cognitively accessible to us.

This consequence is what underlies the Davidsonian response to Cartesian skepticism. "Evil-demon" thought-experiments, in which one has been tricked into developing beliefs about the world that are systematically false, describe a state of affairs that is not logically contradictory, but is at the same time not conceivable. Under such circumstances, we would have to reinterpret these beliefs so that they come out predominantly true. As a result, such skeptical thought-experiments are metaphysical in the Kantian sense—they ask us to speculate about events that occur in possible worlds that are not cognitively accessible to our own.

Again, it is important to note that Davidson's transcendental argument in defense of the intrinsic veridicality of belief does not provide a positive justification for their having this status. What he says is something more like,

¹⁷ Donald Davidson, "Radical Interpretation," in his *Inquiries into Truth and Interpretation* (Oxford: Clarendon Press, 1984).

"well if they didn't, we wouldn't be having this conversation." It is a brute fact about us that we interpret one another charitably. But since we wouldn't be able to interpret one another at all without doing so, given that this principle provides the central criterion of the intelligibility of our utterances, any speculation about suspending it, or doing things some other way, is cognitively idle. And if we did happen to meet someone who didn't interpret utterances charitably, then we would not be able to persuade her that she should, simply because we would not be able to understand what she was doing at all.

4. The argument

With these preliminaries out of the way, the transcendental argument for constrained maximization can now proceed. It has four steps. The general goal is to show that the disposition to accord normative reasons for action deliberative priority is a precondition of all rational thought, and is therefore not something that we can coherently opt to change. (Several of the substantive claims made during the course of the argument are controversial, and cannot be adequately defended here. The goal is simply to get a clear presentation of the structure of the transcendental argument.)

1. *Primary socialization is moral socialization.* This is a slightly exaggerated way of stating the point that training children to act morally is not simply a *component* of the socialization process, something that could be omitted or modified (like teaching them to swim). Socializing children *consists in* cultivating a moral choice disposition. This is an insight that we owe to Emile Durkheim, who observed that there is an element of moral constraint in all normatively regulated social interactions. Society is, at its basis, "*une oeuvre morale.*"¹⁸ Routine social interactions—having a dinner conversation, buying groceries, sitting on the bus, canceling a reservation—are all governed by an intricate set of rules that constrain our conduct. The ability to function in adult society requires a mastery of these rules, along with the capacity to adhere to them. Primary socialization involves the acquisition of this capacity—a "normative control system." (This is usually distinguished from secondary socialization, in which the *content* of the rule system is learned.) The key point is that the same control system is at work when we conform to explicitly "moral" rules—not lying, cheating, murdering, etc.—as when we respect the rules that regulate the details of everyday interaction.

One of the strengths of Gauthier's account is his tacit recognition of this fact. He does not treat "morality" as a special kind of motivation or knowledge. Moral philosophy, for Gauthier, is not distinct from general action theory. What interests him is our ability to make commitments and counter-

¹⁸ Emile Durkheim, *Leçons de sociologie* (Paris: PUF, 1950), p. 51. See also Durkheim, *L'éducation morale* (Paris: PUF, 1963), pp. 25-26.

preferential choices. Whether this capacity is exercised in the form of “politeness” (e.g. not raising one’s voice in a conversation, but waiting one’s turn to be heard), or in the form of “morality” (e.g. not pushing through to the fire exit, but queuing up like everyone else), the interaction has the same action-theoretic structure. There is a tendency to use the term “morality” to refer only to the most important set of social norms governing our conduct, or only the norms that we would endorse, all things considered. This use of the term distracts from the fact that the basic moral phenomenon—the capacity for counter-preferential choice—may be present whenever we conform to any social norm.

One way of stating the thesis is to say that primary socialization involves acquisition of the capacity to respond to one’s *deontic status*. Social norms specify what it is that we are obliged to do under specific circumstances. Making a commitment generates an obligation to perform some particular action. In either case, the agent acquires a deontic status. Socialization involves development of the capacity to assign reasons for action that arise from this status deliberative and motivational priority.

2. *Language is a social practice.* It is a commonplace view among those impressed by the work of the later Wittgenstein that the meaning of linguistic expressions is determined by their use. The key point for the purposes at hand is that the use of such expressions is determined not just by conventions, in the game-theoretic sense, but by social norms. The most highly articulated version of this claim can be found in the work of Robert Brandom.¹⁹ In Brandom’s view, the meaning of linguistic expressions is determined by their inferential role. This role is to be understood as a kind of deontic status in the language-game of assertion. Producing an utterance with assertoric force *commits* one to a series of further utterances. The utterance is permissible only if one has an *entitlement* to make it, on the basis of some other set of utterances. For Brandom, to understand the meaning of an expression is to grasp this set of commitments and entitlements.

It is a consequence of this view that the capacity to produce a meaningful utterance, i.e. an utterance that will be understood by others, requires the capacity to take on and discharge such commitments. (Consider the case of a speaker who says “I’m going for a walk” but then does not acknowledge any of the inferential consequences of this utterance, or perform any of the actions that would be consistent with the kind of commitment undertaken. If one grants the speaker’s sincerity, one has no choice but to suspect that he either meant something else by what he said, or else simply didn’t understand what the words he used meant.) If language is grounded in a normatively regulated social practice, then it follows that only agents capable of making counter-

preferential choices, i.e. constrained maximizers, should be able to produce meaningful utterances. (Note the parallel to the Davidsonian position outlined above. Straightforward maximizers could produce utterances that sound exactly like the utterances produced by constrained maximizers. It is just that we would have no grounds for ascribing *content* to these utterances.)

3. *Intentional states are deontic statuses.* There is an internal connection between belief and assertion. A belief appears to achieve *in foro interno* what an assertion achieves *in foro externo*. While one tradition in philosophy of language seeks to explain assertion as the expression of a belief, the “social practice” perspective outlined above suggests that beliefs are best understood as a kind of deontic status. To say that someone believes that *p* is to say that this person is committed to the claim that *p* (and so could, for instance, be called upon to display her entitlement to it, or acknowledge some of the further commitments that follow from it). In order to counteract the tendency to hypostatize these sort of intentional states (i.e. to think of them as something inside the agent’s head), Brandom suggests that we should use the term “doxastic commitment” instead of belief.²⁰

From this, it follows that only constrained maximizers can have contentful beliefs, because only constrained maximizers are able to respond to deontic statuses, which include doxastic commitments. (This will sound odd to people who are inclined to ascribe beliefs to, say, animals. But there are familiar Davidsonian reasons for not doing so. It is misleadingly metaphoric to say that an organism has a “belief” when it lacks the capacity to acknowledge any of the inferential consequences or entitlements associated with the relevant propositional content.)

4. *To decide that one should be a straightforward maximizer, then do it, would be to respond to a deontic status, and so would be an exercise of constrained maximization.* This puts things a little bit too neatly, but it conveys the general idea. In order to rationally reflect upon one’s fundamental choice disposition, one must acquire a series of doxastic commitments, and then track their consequences. This is what makes the process recognizable as *reasoning*, rather than just behaviour. But as a result, constrained maximization is a capacity that we must *exercise* in order to carry out the process of rational reflection. Questioning our disposition to accord normative reasons for action deliberative priority is therefore cognitively idle, because the very intelligibility of the question depends upon a background exercise of precisely such a prioritization. Otherwise put, from the standpoint of rationality, constrained maximization is transcendentally necessary. A rational agent could

¹⁹ Robert Brandom, *Making It Explicit* (Cambridge, MA: Harvard University Press, 1994).

²⁰ *Making It Explicit*, p. 157-9.

never choose to become a straightforward maximizer, because constrained maximization is a condition of possibility of thought.

It is always possible that we might opt for straightforward maximization through some kind of radical existential choice. The point is simply that this is not something that could rationally recommended itself to us. In the same way, we could cease to interpret people charitably. But in so doing, we would cease to interpret them at all, and so would be opting out of rational agency.

The key idea in this argument is that because rationality involves the use of language, and learning language requires mastery of a normatively regulated social practice, moral agency is a precondition of rational agency. This is not to deny that people can act immorally, or that they can rationally choose to do so. The claim is simply that because of the internal connection between moral constraint and rationality, it is impossible to argue oneself out of having a higher-order moral choice disposition. By the time that one has the capacity to engage in this sort of deliberation, it's too late. Immorality, in this context, is like false belief for Davidson. It is possible to persuade oneself that one has a few false beliefs. But one cannot persuade oneself that one's beliefs are systematically false, because this is inconsistent with the presuppositions of rational agency. Similarly, it is possible to persuade oneself to perform specific actions that are wrong, but it is impossible to argue oneself into a state where one no longer experiences the force of moral constraints.

5. Objections

Once the position has been laid out in this way, one can see how important it is that the argument be addressed to agents who are already constrained maximizers. The argument does nothing to recommend a moral choice disposition to someone who does not already have it. It demonstrates only the reflective stability of this disposition. The evolutionary story was introduced in order to render plausible the initial presumption that the relevant audience consisted only of those with the appropriate disposition. However, it is important to note that the argument itself retroactively justifies this presumption. If only constrained maximizers are capable of language mastery, then it is natural that no justification of this disposition could be aimed at those who do not have it. Constrained maximizers are the only sort of agents for whom the question "What kind of person should I be?" makes sense.

This argument gives rise to a number of fairly obvious objections. Two of them are worth discussing in detail:

1. Even if all this is correct, couldn't we be disposed to respect only those deontic statuses that are needed to participate in the practices that consti-

tute our linguistic abilities—and thus our capacity for rational deliberation?

Both Kant and Gauthier assume that the fundamental choice disposition is completely general—one chooses to be a constrained maximizer, or to assign categorical imperatives priority, in all cases. The suggestion here is that the agent might adopt a more flexible disposition, and respect only the set of "cognitive norms," while systematically violating all other, moral or social norms.

However, there are good Sellarsian reasons to believe that such a compartmentalization of the set of cognitive norms is not possible. According to Wilfrid Sellars (and Brandom), empirical content enters into the "game of giving and asking for reasons" through both language-entry and language-exit moves.²¹ Language-entry moves typically consist of observations, and provide the representational dimension of linguistic meaning. Language-exit moves consist of actions, and provide the pragmatic content of meaning. Mastery of an expression involves familiarity with both the "upstream" reasons that entitle one to it, as well as the "downstream" consequences of accepting it. Producing intelligible utterances involves not only responding to one's environment in a certain way, but also acting in a way that is consistent with these utterances (this is the point of the quotation from Sellars at the start of the paper). Since these downstream consequences may involve any sort of action that falls under an intentional description, there is no discrete set of norms that constitute "the" practice of assertion. Language is woven into the fabric of all social actions, and so requires the capacity to respond to one's deontic status generally, not some specific set of such statuses. Thus the agent cannot adopt a disposition to respect only "cognitive" norms.

This does not mean that a person must actually endorse the prevailing set of "moral" norms in order to use language. It means only that a person must have a generalized disposition to assign normative reasons for action deliberative priority. It is still possible for an agent who has such a disposition to reject the entire set of "moral" norms, and accept only "cognitive" ones (the same way a person might refuse to accept some significant segment of, e.g., our everyday empirical beliefs). But such a person poses a different sort of problem from the one posed by the straightforward maximizer. The problem here is one of content skepticism (to use Christine Korsgaard's terminology), not motivational skepticism.²² The straightforward maximizer is someone who simply does not respond to his or her deontic status, and so will not feel the force of any moral argument. The constrained maximizer who simply

²¹ Wilfrid Sellars, "Some Reflections on Language Games," in *Science, Perception and Reality* (London: Routledge & Kegan Paul, 1963): 321-358.

²² Christine Korsgaard, "Skepticism about Practice Reason," *Journal of Philosophy*, 83 (1986): 5-25.

rejects "moral" norms feels the force of such arguments, but simply denies that they are sound. That such people exist does not constitute an objection to the transcendental argument presented above, because this argument is intended only to preclude the former.

2. *Does this argument not run the danger of proving too much? Is it not the case that we sometimes meet rational people who genuinely lack a moral choice disposition?*

The short answer to this question is "yes, but..." In other words, a sufficiently nuanced analysis of the question reveals that it does not create a serious problem for the view advanced.

The first thing to note is that Gauthier treats constrained maximization as a choice disposition that generates a lexical ordering of types of reasons for action. Reasons that arise from the "plan" simply trump reasons that arise from one's preferences in the immediate choice situation. I am inclined rather to regard the fundamental choice disposition as one which assigns a certain weight to normative and instrumental considerations.²³ Social norms are more or less important, just as preferences can be more or less intense. The choice disposition simply assigns a weight to these normative constraints relative to one's preferences (much as one's prudential constraint, expressed in the form of a discount factor, assigns a weight to future preferences relative to present ones). As a result, it is possible to submit to temptation, without having this reflect any disturbance of the underlying choice disposition. In the same way that short-term satisfaction can simply overshadow the long-term consequences of an act, it can also overshadow concerns over its impropriety.

One no doubt encounters an enormous amount of variation in the strength of this moral disposition from person to person, ranging from hyperconformism on one side to moral laxity on the other. Most of the "everyday immorality" that one encounters is, I would argue, the result of laxity, and not a fundamentally different choice disposition. The second important point to note, however, is that one does not respond to cases of serious moral laxity with argumentation alone. We generally respond to moral intransigence by *sanctioning* the offender. Although this sometimes takes the form of outright punishment, it is often more subtle. We cease to trust persons with poor moral character, we refuse to cooperate with them, we disassociate ourselves from them, and we symbolically censure their actions. This is just the mechanism of socialization at work. It is precisely through the internalization of these sanctions that the moral choice disposition is acquired. Thus one does not argue someone into assigning moral considerations greater deliberative weight, one socializes them into doing so.

²³ See Joseph Heath, "The Structure of Normative Control," *Law and Philosophy*, 17 (1998): 419-442.

An adult who genuinely assigns no weight to moral considerations is not someone who lacks a particular sort of sensitivity, it is a person who has suffered a more general failure of socialization. Such persons no doubt exist, but the point to note is that this failure of socialization, which impairs their ability to feel the force of moral constraints, also impairs their ability to respond to rational argumentation. Argumentation is ultimately a form of moral suasion. A certain threshold level of moral constraint is needed in order to function as a fully rational agent. It is a mistake to assume that moral commitments, in order to be justified, need to be justifiable to those who fall below this threshold, since only those above the threshold are able to act as full participants in the "game of giving and asking for reasons."

Finally, it should be noted that a lot of behavior which we are inclined to classify as evil is in fact rationalized. People who do bad things usually have some kind of story that purports to justify the conduct from the moral point of view—often grounded in a sense of grievance (e.g. punishing others for harms that the agent has suffered), or a failure of reciprocity (i.e. preemptively or defensively defecting from cooperative arrangements—the "everyone else is doing it" line familiar to any parent). The key is that in all of these cases, the agent is responding by providing reasons (good or bad) that have force within the existing normative framework. As a result, they are not genuinely opting out of morality, they are just behaving badly.

The more general point is that we need not worry about how to justify our moral commitments to individuals who do not feel the force of moral constraints. As a result, the argument presented here wouldn't convince an alien or a sociopath. The purpose of the transcendental argument is to show that we don't need to convince these sorts of people in order to allay our doubts about the defensibility of our own moral commitments. It is intended to show that the person who asks "What if you could take a pill that would make you a straightforward maximizer?" is similar to the person who asks "How do you know that your beliefs aren't all false?" The transcendental argument shows that we do not require a step-by-step argument that excludes this claim. The question is intelligible only under assumptions that are metaphysical in the pejorative sense. Morality is therefore not given a foundation, it is simply shown to be (again in the more precise German terminology) *nicht-hintergebar*.²⁴

²⁴ Thanks to David Davies and Christine Tappolet, along with audiences at McGill University and Université de Montréal for comments on earlier versions of this paper. Financial assistance provided by the Social Sciences and Humanities Research Council of Canada.