Consequentialism, Cognitive Limitations, and Moral Theory[†]

Dale Dorsey

Department of Philosophy
University of Kansas
1445 Jayhawk Boulevard
Wescoe Hall, rm. 3090
Lawrence, KS 66045
ddorsey@ku.edu

ABSTRACT: In this paper I characterize a recent and penetrating objection to consequentialism on grounds of human cognitive limitations. Standard human agents, or so the objection claims, will rarely, if ever, conform to moral requirements identified by consequentialism: consequentialism will require acts that, though strictly possible for an agent to perform, will not be performed simply given the fact that humans are cognitively limited beings. I argue that this objection extends to virtually all plausible moral theories, and I propose a solution: moral theories should restrict their deontic evaluation of acts to those that could be performed *as an agent*.

KEYWORDS: consequentialism; agency; 'ought' implies 'can'; moral valence, moral requirability; contrastive reasons; objective reasons

Consider the moral theory called "objective act-consequentialism" (hereafter, "consequentialism"). This view holds that the moral quality of a given action ϕ in comparison to another action ψ is determined by the quality of ϕ's actual consequences in comparison to the quality of ψ's actual consequences. There are many reasons people have found to reject this sort of view. Some have claimed that it is too demanding.[1] Some have claimed that it disrupts a moral agent's integrity.[2] Some suggest that it inappropriately denies that there are no constraints on the pursuit of the overall

---

[†] Special thanks are due to Doug Portmore, whose generous comments radically improved an earlier draft of this paper. I would also like to thank John Bricke, Richard DeGeorge, Julia Driver, Ben Eggleston, Erin Frykholm, Scott Jenkins, Sean McKeever, Howard Nye, Mark van Roojen, Drew Schroeder, Dave Shoemaker, Daniel Star, and participants in the 2012 Arizona Normative Ethics Workshop for helpful conversations.

[1] Samuel Scheffler, "Morality's Demands and their Limits" in *The Journal of Philosophy* 83 (1986).

[2] Bernard Williams, "A Critique of Utilitarianism" in Smart and Williams, *Utilitarianism: For and Against* (Cambridge: Cambridge University Press, 1974).

good.[3]

Other objections to consequentialism focus on consequentialism's ill fit with the fact of human cognitive limitations. Traditionally, such objections focus on our general ignorance of the consequences of our actions, and hence their moral valence.[4] However, a recent and penetrating objection notes that the problem of cognitive limitations runs much deeper. Consequentialism seems to require us, at virtually all times, to perform actions that, though strictly possible for us to perform, we will not perform *simply* given our everyday limitations as cognitive agents. This is true even *if* we know that the action in question will have tremendously good consequences, better than any other alternatives.[5]

In this paper I argue that this problem is not isolated; it applies not just to consequentialism, but to virtually every plausible moral theory. I argue that the root of this problem is to be found in what might be called a *theory of moral requirability.* I propose an independently plausible alternative to the traditional theory of moral requirability, and argue that this alternative successfully avoids commitment to a view that renders morally acceptable behavior out of reach for virtually all normal human beings.[6]

1. Cognitive Limitations and Moral Requirements

It's well-known that consequentialism has some sort of problem with actions one can, strictly speaking, perform, but actions that one won't perform given normal cognitive limitations.[7] Examples of this kind abound. Take, for instance:

> *Wilda*: Wilda stands before a safe. Were she to open the safe, overwhelmingly good consequences would result. However, she

---

[3] Frances Kamm, "Non-consequentialism, the Person as an End-in-Itself, and the Significance of Status" in *Philosophy and Public Affairs* 21 (1992).

[4] Consider, for instance, James Lenman, "Consequentialism and Cluelessness" in *Philosophy and Public Affairs* 28 (2000).

[5] Eric Wiland, "Monkeys, Typewriters, and Objective Consequentialism" in *Ratio* 18 (2005).

[6] This moniker is imperfect for a number of reasons; I use it here simply as a placeholder for the sort of objection I explore in the next section.

[7] See, for instance, Francis Howard-Snyder, "The Rejection of Objective Consequentialism" in *Utilitas* 9 (1997); Erik Carlson, "The Oughts and Cans of Objective Consequentialism" in *Utilitas* 11 (1999); Mozaffar Quizilbash, "The Rejection of Objective Consequentialism: A Comment" in *Utilitas* 11 (1999); Dale Miller, "Actual-Consequence Act Utilitarianism and the Best Possible Humans" in *Ratio* 16 (2003), 50; Wiland, op. cit.

> has no idea what the combination of the safe is.

Despite the fact that Wilda doesn't know the combination of the safe, consequentialism would appear to require her to open the safe. After all, she *can* open the safe, i.e., by entering the right numbers in the right order. But it would seem odd to say that she behaved in a morally impermissible manner in failing to do so. After all, she doesn't know how.

Some consequentialists might feel comfortable simply biting this bullet. As Wiland notes, to say that Wilda is morally required to open her safe seems more or less innocuous, insofar as her case is pretty unusual, and, furthermore, consequentialism generally refuses to index an individual's moral requirements to his or her epistemic states.[8] Somewhere along the line, however, this sort of bullet-biting starts to sound a bit goofy. Take another example:

> *Jeffrey*: Jeffrey is spending a leisurely afternoon reading a titillating detective novel. He could, alternatively, sit down at his computer, type the cure for cancer, and send it to a top medical journal. The consequences of performing this action would be overwhelmingly good. It does not occur to Jeffrey to do this not because he believes that curing cancer would have suboptimal consequences, but rather because he has no clue what the cure for cancer is, or how to write it down.

Jeffrey, in a perfectly sensible sense of "can", *can* widely disseminate the cure for cancer. All he needs to do is to sit down, type it, and disseminate it widely. Jeffrey can type. No one's stopping him. Assuming such a cure exists, if we also make the relatively innocuous assumption that curing cancer would produce overwhelmingly good consequences,[9] it would appear that he morally ought to do so.[10] Of course, there surely are additional senses of "can" that entail that Jeffrey cannot cure cancer.[11] However, the sense of "can" I work with is the sense of can that is implied by "did": if I ϕ'ed at t, it follows that I could have ϕ'ed at t. (In essence, we don't want to say that it could occur that someone performs an action at t that he or she *cannot* perform at t.[12]) Insofar as there is nothing stopping Jeffrey from having cured cancer at t—all he has to do is hit the right keys in the right order—it follows that he *can* cure cancer at t. Given the overwhelmingly good consequences of curing cancer, it would appear that simply continuing to read his detective novel is *far* below the moral ideal.

---

[8] Wiland, 355-356.

[9] For a more pessimistic view of the good generated by a cure for cancer, see Albert Brooks's novel *2030*.

[10] See Wiland, 356-7.

[11] See Howard-Snyder, 243-44.

[12] See Miller, 53-54.

But this is surely absurd. Though one might be willing to bite the bullet in Wilda's case (though this doesn't strike me as particularly plausible in any event), Jeffrey's case illustrates that the problem at hand is not confined to the stranger than fiction. In fact, as Wiland puts it: "once we become literate creatures, the number of action-options usually available to us is astronomically high. . . The chance you ever do what in fact has the best consequences is laughably small. If you act rightly only if you do what in fact has the best consequences, then you can be fairly certain that you never (or almost never) act rightly."[13] And, or so it would appear, there is good reason to reject (objective act-) consequentialism.

It is worth saying a little about what the objection in question amounts to. Wiland and Miller focus on the suggestion that "[objective consequentialism] implies that our actions are *almost always* wrong."[14] And this is bad enough. But I think the objection is broader than this. Recall that even if we focus strictly on Wilda's unusual case, the fact that she does not know how to open the safe seems to exempt doing so from the set of those actions that can be sensibly morally required of her. Thus the problem is not—or not only—that according to consequentialism we almost always act wrongly. Rather, the problem with consequentialism appears to be that it—however rarely or often—requires us to perform actions not that we cannot perform, but that we *will not* perform *simply* as a result of our limitations as cognitive agents (perhaps along with a failure to have cosmically good luck). This is true even if we have perfect knowledge of the relevant consequences. But this is absurd. Call this the "problem of cognitive limitations".

2. Cognitive Limitations and Non-Consequentialism

As stated, the objection from cognitive limitations is an objection to consequentialism. But the fact of cognitive limitations causes problems not just for consequentialism but, as I shall argue, virtually every moral theory. Take the following principle:

> *Minimal Consequentialism* (MC): that ϕ-ing will produce good consequences is a moral reason to ϕ, a reason that strengthens

---

[13] Wiland, 359.

[14] Wiland, 353; Miller, 55.

> as the quality of the consequences of ϕ-ing increases.

Minimal Consequentialism is shared by consequentialists and many non-consequentialists. For consequentialists, MC is sufficient to fully catalog moral reasons. However, a moral theory can accept *Minimal Consequentialism* without accepting consequentialism. After all, MC says only that good consequences of an action generate *a* moral reason for performance, not that this reason is overriding, or even comparatively very important. For instance, one could accept MC and a host of agent-centered restrictions, permissions, etc. However, any theory that accepts this general principle will face the problems noted in §1.[15] Take Jeffrey's case. Curing cancer by means of typing out the formula and widely disseminating it will produce overwhelmingly good consequences. But given that good consequences generate moral reasons, and given that the good consequences of curing cancer are *overwhelming*, one would expect that no matter how comparatively unimportant the reason to promote good consequences is, the sheer magnitude of good caused by such an action would be enough to entail that the reason to do so is sufficient to morally require doing so. To put this another way, it would appear that (a) the enormity of the good involved should, on any plausible view, morally override any non-consequentialist permissions or restrictions and (b) even if not, it is very unlikely that typing out the cure for cancer would be particularly burdensome for Jeffrey (beyond slightly delaying the *denoumént* of his gooseflesher), nor would it require him to commit an unpardonable moral sin. Even if the reason to promote goodness is very weak compared to all other moral reasons, any MC-accepting theory will face the problem of cognitive limitations.

One possible solution, whether for explicitly consequentialist theories or for theories that treat consequences as only one factor among many, is to reject the claim that moral reasons are a product of *actual* consequences. Rather, one might hold that moral reasons are generated by *expected* consequences, expected, presumably, by the agent in question. Correspondingly, such a view would accept:

> *Minimal Subjective Consequentialism* (MSC): that ϕ-ing will produce good expected consequences is a moral reason to ϕ, a reason

---

[15] An anonymous reviewer notes that some moral theories will hold that moral reasons must also be paired with, say, additional "requirement generating" factors to generate a moral requirement, which, in Jeffrey's case, say, won't be present. But this suggestion is just another way of getting at the central problem of this paper: which acts are morally *requirable*, and what makes them so?

> that strengthens as the quality of the expected consequences of ϕ-ing increases.[16]

MSC suggests that the moral valence of action is, at least in part, determined by the action's expected, rather than actual, consequences. MSC permits of varied interpretation; some will say that expected consequences are those that can be expected given the agent's actual epistemic states. Others will say that expected consequences are determined by what an agent would expect under certain idealized or counterfactual epistemic conditions.

Offhand, however, it is a little difficult to see why any view that accepts MSC rather than MC could avoid the problem of cognitive limitations.[17] After all, the expected consequences of my curing cancer are certainly extremely good; the expected consequences of any other action I could perform at the time pale in comparison. If so, MSC is no fix: curing cancer not only has the best *actual* consequences, but also the best *expected* consequences. This is true even if we hold that expected consequences are determined from the perspective of the agent's actual epistemic states. As I sit, right now, I certainly have sufficient information to determine that the expected consequences of curing cancer are extraordinarily good. If so, or so it would seem, I am morally required to do so.

A response is worth considering. One might say that though the expected consequences of *curing cancer* are good, the expected consequences of taking the steps necessary to cure cancer, i.e., typing a set of random characters on one's keyboard, are not particularly good. After all, the chances that I will actually cure cancer by sitting down and typing out a bunch of random words or characters are pretty slim. And if so, I can't be required to do so: the expected consequences are of very low comparative quality. But this response doesn't solve the problem. While I may not be required to sit down at my computer and punch keys randomly—which is a means to curing cancer—I am nevertheless required, given its foreseeable consequences, to *cure cancer*. To claim that I am not required to cure cancer given that I am not required to sit down at my computer and type a bunch of characters randomly is to treat as identical two distinct acts: the act of curing cancer and the act of typing a bunch of random keys.[18] And insofar as we can distinguish these actions, as we surely can, any view that accepts MSC will generate a moral reason to cure cancer. Thus MSC does not avoid the problems cited here.

---

[16] Wiland, for instance, explicitly focuses on what he calls "objective" consequentialism—a view that accepts MC rather than MSC, though he does not consider whether subjective consequentialism would be able to avoid the problem at hand.

[17] Dale Miller, in proposing a version of the problem discussed here, suggests that "even some utilitarian or non-utilitarian forms of probable-consequence act consequentialism" may succumb to it (Miller, 50).

Here's a promissory note: I argue in the next section that *any* halfway plausible moral theory will face the problem of cognitive limitations. However, even if we allow that entirely non-consequentialist moral theories can avoid these problems, MC and MSC are critical features of a very wide range of moral theories. Any moral theory that cares about the consequences of our actions will accept one or the other. Insofar as there appears to be very good moral reason to care about consequences (even if just a little bit), a fix for the problem of moral obligation in light of cognitive limitation is thus of the first importance not just for consequentialism, but for moral theory more generally.

3. Moral Requirability

The reflections just offered seem to indicate that a solution to the problem at hand cannot be found simply by subbing-out one moral theory for another. Rather, insofar as any theory that accepts MC or MSC seems to succumb to the problem of cognitive limitations, it would appear that the solution must be found elsewhere.

To see this in more detail, I want to introduce a bit of terminology. Call an action ϕ "requirable" for a person x at time t if and only if x's ϕ-ing is suitable to be assigned a deontic valence at t. Every moral theory must offer what might be called a "theory of moral requirability", i.e., a method by which to determine, for any person at any time, what set of actions are eligible to be assigned a deontic valence. But, in addition, every moral theory must offer what might be called a "theory of moral valence". A theory of moral valence takes the requirable acts and determines their deontic valence; some will be required, the others will have some different valence (permissible, impermissible, etc.). As a theory of moral valence, consequentialism will assign deontic categories to requirable actions as a result of the quality of their actual consequences. Other theories of moral valence will assess requirable actions differently, according to a set of specified moral factors.

---

[18] See, for instance, Alvin Goldman, *A Theory of Human Action* (Englewood Cliffs, NJ: Prentice Hall, 1970), chs. 1 and 2. I should note here that I will assume a fine-grained theory of action-individuation. This is, of course, controversial. But nothing substantive rides on this. One could say also that Jeffrey expects very good consequences to result from his action of curing cancer *by* typing the right characters in the right order, which is surely correct. I will assume a fine-grained scheme in this paper, and will put arguments in terms of such a scheme, but the arguments can be translated to a coarse-grained scheme *mutatis mutandis.*

Notice, however, that substantive theories of moral valence need not differ on the proper theory of moral requirability. A strict Kantianism and a strict act-utilitarianism will certainly differ concerning what morally requirable action, for x at t, is morally required, etc. But they needn't differ about which actions are *requirable*, or which actions can be properly assessed for a particular comparative deontic valence.

By far the most popular theory of moral requirability is:

> *The Traditional View* (TRAD): ϕ is requirable for x at t if and only if x can ϕ at t.

TRAD is compatible with any theory of moral valence. TRAD does not say how one ought to evaluate the actions it identifies as morally requirable. Rather, it simply restricts the set of actions that can properly be assigned a moral deontic valence to those that can be performed by the agent.

The root of the problem of cognitive limitations lies in accepting TRAD. TRAD holds that any action one can perform is in principle eligible to be assigned a deontic valence. But this yields that, for Jeffrey, curing cancer will be among those actions that will be evaluated by a theory of moral valence. But no halfway plausible theory of moral valence, whether consequentialist or non-consequentialist, will hold that *assuming that curing cancer is morally requirable for Jeffrey*, Jeffrey is not morally required to cure cancer. To say so is to render one's theory of moral valence too absurd for consideration *if* we assume that to cure cancer is an action that can be assigned a moral valence. But if this is correct, the problem of cognitive limitations lies in TRAD, not in any particular theory of moral valence. Hence it would seem sensible, if we are concerned about the puzzle of cognitive limitations, to narrow the range of morally requirable actions whatever one's preferred theory of moral valence.

4. Alternatives

But how? Examining the cases of Wilda and Jeffrey might lead to an obvious possibility:

> *The Knowledge View* (KNOW): ϕ is morally requirable for x at t if and only if x knows how to ϕ at t.

Insofar as Wilda does not know how to open the safe, and Jeffrey does not know how to cure cancer, it would appear that KNOW entails that they cannot be morally required to do so even given the quality of the resulting consequences. So far so good.

However, there are good reasons to reject KNOW. KNOW is both under-inclusive and over-inclusive. First, it is under-inclusive. Consider:

> *Sondra*: Sondra stands before a safe that, were she to open it, cancer would be cured. Also imagine that Sondra believes that entering the numbers 867-5309 will open the safe. Furthermore, her belief is justified and true—this number is, in fact, the safe's combination. However, imagine that Sondra's justified true belief was a product of a Gettier-like scenario,[19] and hence that this particular number will open the safe does not count, for Sondra, as knowledge.

Whatever else is true about the relationship between knowledge-that and knowledge-how, it is *surely* the case that to know how to open the safe requires one to know it's combination. But in this case, Sondra does not know the combination. Her justified true belief is not knowledge. Hence, she fails to know how to open the safe. Nevertheless, it would seem absurd to say that she could not be required to open the safe. After all, she believes she knows how to do so, that belief is justified, and, were she to actually enter the number she believes will open the safe, cancer would be cured. Not only can she be required to open the safe, her failure to do so is a grave moral wrong.

Second, KNOW is over-inclusive. Some actions we know how to perform are inappropriate targets of moral obligation even if we know how to perform them. Assume, for instance, that the correct combination to Wilda's safe is 867-5309. Wilda certainly *knows how to enter* 867-5309. Insofar as she knows how to work a safe, she knows how to enter any particular number into that safe.[20] And insofar as performing this action entails *curing cancer*, many theories of moral valence will, if doing so is requirable, require it.[21]

---

[19] Imagine, for instance, that Sondra overheard the owner of the safe declare that anytime he had a safe, he would set the combination to the title of his favorite Tommy Tutone song. But also imagine that the owner never actually set the combination, and it just happened to have been set to that number randomly in the factory.

[20] Cf. Carlson, 92.

[21] This might not be a problem for KNOW on the assumption of MSC. It is certainly the case that entering 867-5309 does not have very good expected consequences. Nevertheless, KNOW seems to me to fail for two reasons, and the former is applicable even if we accept MSC rather than MC.

But requiring Wilda to enter 867-5309 is certainly no more plausible than requiring Wilda to *open the safe*.[22] After all, if the problem of cognitive limitations is that people, like Wilda, will fail to perform morally required actions *simply* given the standard cognitive limitations of human beings, this problem surely applies to a moral requirement to enter 867-5309 no less than to a requirement to open the safe. Parallel reasoning applies in Jeffrey's case. We can assume that there is some string of characters, S, such that were Jeffrey to type S, he would cure cancer. Insofar as Jeffrey knows how to type, he knows how to type S. KNOW implies that Wilda can be morally required to enter 867-5309, and Jeffrey can be morally required to type S. But it seems implausible to say that, e.g., Wilda should be required to do so; the reason surely is, at least in part, that she fails to recognize entering 867-5309 as a method by which to open the safe, just as Jeffrey fails to recognize S as a method by which to cure cancer.

I think we should reject KNOW. It seems wrong to say that the central cognitive limitation involved in the problem of cognitive limitations is knowledge of how to perform the action in question, for the reasons just cited. But we might do better by indexing requirability not to one's epistemic stance *vis-á-vis* how to ϕ, but rather to one's ability to successfully *intend* to ϕ. Take:

> *The Intention View* (INTEND): ϕ is morally requirable for x at t if and only if x can intend to ϕ at t, and were x to intend to ϕ at t, x would ϕ at t.[23]

This view seems to avoid holding Wilda and Jeffrey morally accountable for failing to open the safe, and failing to cure cancer, respectively. Given their cognitive limitations, we can assume that, were they to intend to cure cancer or open the safe, they would (under most normal conditions) fail to do so. INTEND also avoids the problematic suggestion that Sondra cannot be morally required to open the safe. However, INTEND also seems to imply that Wilda can be morally required to enter 867-5309, and Jeffrey can be morally required to type the string of characters designated here as "S". Were they to intend to perform those actions, they would do so.

---

[22] An anonymous reviewer disputes this claim. The contrary proposal is that to require Wilda to enter 867-5309 is *much* more plausible, insofar as this is an open option for Wilda. After all, *she knows how to do it* and *to do it would be to produce the best consequences*. But this seems to me not enough; this reasoning begs the question in favor of KNOW. A fan of TRAD might say that to open the safe is also an open option. She *can* do it. And so we need some way to settle what it means for a particular action to be an open option. This can only be settled, as far as I can tell, by figuring out which is the most plausible theory of moral requirability. For my money, it simply isn't plausible to say that Wilda should be required to enter 867-5309, for every reason we find it implausible to say that Wilda should be required to open the safe. And hence we should look beyond KNOW. I admit, however, that this is simply my considered judgment, which just like all the others is subject to dissent.

[23] See Howard-Snyder, 244.

An important response arises for INTEND that is not available for KNOW: if we conjoin INTEND to MSC rather than MC, we can hold that, because the expected consequences of entering 867-5309 are not particularly good, INTEND+MSC avoids the problematic claims we've been discussing so far. However, I remain unconvinced that INTEND+MSC can solve the problem in a satisfactory way. It seems right to say that Wilda could intend to open the safe *by* entering 867-5309.[24] Of course, the expected consequences of opening the safe by entering any number at all are extraordinarily good. And were it the case that Wilda were to intend to open the safe by entering 867-5309, she would succeed at so doing.[25]

I regret not having the space to consider a number of other potential theories of moral requirability; I have surely not yet scratched the surface of the sheer volume of potential iterations of the various views.[26] Many alternatives, some perhaps plausible, may remain, and so I leave open that any or all of the previous views might be tweaked here or there to arrive at better results. However, I believe that there are special reasons to accept the proposal I advocate here (even leaving aside the problem of cognitive limitations), and so I will focus on reasons for and against my preferred alternative from here on out.

5. The Agency View[27]

---

[24] Two notes. First, the fine-grained theory of action will refer to this as a "longish" action; see next section for more discussion of this idea. Second, this claim depends on a controversial view of intention, as any interpretation of INTEND surely does. Some hold that one cannot intend to ϕ unless one believes one will be successful in ϕ-ing. But there are well-rehearsed reasons to doubt this claim. As Bratman notes, I can intend to rescue someone, despite the fact that I have very little confidence that I will actually succeed at so doing. See Michael Bratman, *Intention, Plans, and Practical Reason* (Cambridge, MA: Harvard University Press, 1987), 38. See also Hugh J. McCann, "Settled Objectives and Rational Constraints" in *American Philosophical Quarterly* 28 (1991), 27.

[25] Even if we reject the suggestion that Wilda could intend to open the safe by entering 867-5309, there remains a problem: INTEND+MSC seems to get the wrong answer in Joey's case, which I explore in §6.1.

[26] In particular, I'd like to than Sean McKeever and Mark van Roojen for proposing a variant of INTEND that I simply don't have space to discuss.

[27] This section was improved by helpful conversations with Julia Driver.

To introduce my proposal, let's consider just what it is that Sondra maintains, but that Wilda lacks. What *can't* Wilda do, that Sondra can? Though there may be many such things, one thing that speaks to me is that Wilda cannot *reason her way* to opening the safe. Sondra can, on the basis of reasons she recognizes, open the safe, or enter the right numbers, etc. For Wilda, whether she opens the safe or enters the right numbers is the result, quite simply, of a lucky guess.

If this is correct, and I think it is, one might consider the following alternative theory of moral requirability:

> *The Agency View* (AGENT): ϕ is morally requirable for x at t if and only if x can ϕ at t as an agent.

To see AGENT in more detail, and why it offers a solution to the problems I investigate, I must first explore the notion of "agency" with which I seek to work. My intention is not to provide a full theory of agency or any cognate concepts. I remain neutral on such topics as whether moral deliberation is a result of cognitive, conative, or affective pro-attitudes, the nature of autonomy or autonomous action, whether autonomous agents act under the guise of the good, or whether normative judgments are necessarily motivating, etc. Rather, I seek a minimal theory of what it means to perform a particular act at a particular time *as an agent*, one that I hope will be broadly ecumenical when it comes to substantive controversies such as those noted here.

In assessing whether a particular action can be performed as an agent in the sense I mean, it is insufficient to say that the action was (a) performed by x and (b) x is a moral agent. Moral agents can sometimes fail to perform actions *as* agents. For instance, moral agents can decide based on whim or caprice, they might on occasion govern their actions by the simple flip of a coin. Doing so does not entail that they are not moral agents. But doing so does entail that the resulting actions are not performed *as agents.* To perform an action as an agent is to perform that action on the basis of one's own deliberative agency. It is, in other words, to *see a reason to perform that action*, and to *perform the action on the basis of that reason.*

Two qualifications should be made clear immediately. To perform a particular action as an agent, it is insufficient simply to be responsive to reasons generally, or (in other words) to see *some reason or other* to perform that action. To perform an action as an agent, the reason one sees, or is responsive to, must be of the right *kind.* To see what I mean, take the following case.

> *Roulette*: I stand at a roulette table, and I see reasons to play roulette: it will be thrilling, it will satisfy a desire of mine, etc. However, I see no reason to bet on any *particular* number. I bet on 22.

In *Roulette* I have performed a number of different actions. I have played roulette. I have bet on 22. I have bet on a black number, an even number, and a number in the second third of the table. However, if we ask: "have I bet on 22 as an agent?" it seems to me the answer is clearly *no.* Of course, I have played roulette as an agent, insofar as I deliberated on the basis of reasons I recognize to play roulette. But I haven't bet on 22 on the basis of my deliberative agency. This is true despite the fact that I see clear reasons to bet on 22. Insofar as betting on 22 is a way to play roulette, this is certainly a reason I will recognize to bet on 22. Insofar as betting on 22 is one way to not cut my own legs off, which I'd prefer to avoid doing, I certainly recognize a further reason to bet on 22. But importantly none of these considerations will tell in favor of betting on 22 rather than placing any *other* bet on the roulette table. The reasons I recognize do not support betting on 22 *per se*, but rather a disjunction of potential actions I might perform.

Thus to perform an action as an agent it is not sufficient to see some reason or other to perform the action. Rather, one must see what I shall dub a "contrastive" reason. A contrastive reason to ϕ is a reason to ϕ *rather than* any other action ψ one might perform. The reason(s) one sees cannot simply support a disjunction of potential actions. They must support *that action in particular.*[28]

The second qualification runs as follows. To ϕ as an agent, it is not sufficient to ϕ on the basis of some contrastive reason I see to ϕ. Rather, it must be the case that the difference between ϕ-ing and merely *trying* to ϕ is under one's deliberative control. To see why this is important, take the following example:

> *Sarah*: Sarah, a scientist, believes that there are good reasons to produce organism X in a certain petri dish. However, though

---

[28] One might say that this radically restricts the range of actions that I might perform as an agent. Imagine, for instance, that I recognize a reason that tells in favor only of *two* potential actions ϕ and ψ; though it rules out all others I might perform. On this view, if I see no reason to ϕ rather than ψ I cannot ϕ as an agent, despite the fact that I might deliberate, and on the basis of my deliberation, decide to either ϕ or ψ. But it seems to me that this verdict is exactly the correct one. If I cannot see a reason to ϕ rather than ψ, we might say I "decided, as an agent, to ϕ or ψ". If we allow the possibility of disjunctive acts, we may say I ϕ-or-ψ'ed as an agent. But we would not say that I ϕ'ed as an agent.

> Sarah knows that she must, at the very least, leave open the petri dish, she doesn't know what to add to it, where to place the petri dish, at what temperature the petri dish must be kept, or any other details about the process by which organism X is grown. On a whim, she places it at a certain spot and temperature, adds random chemicals, and leaves for the night. When she returns in the morning, she has grown organism X.

We would say that Sarah grew organism X, surely. But would we say that she grew organism X *as an agent*, as a result of her deliberative agency? Surely not. This is true despite the fact that she sees contrastive reason to grow organism X, and, one can assume, threw a bunch of chemicals together partly on the basis of this contrastive reason. How, then, to explain what Sarah lacks?

To explore this point in more detail, it is helpful to make use of the idea of a "longish" action.[29] Some actions unfold over time, and are made up of individual sub-acts. Sarah has deliberative control over some of the individual sub-acts and not others. Hence we can make a distinction between the sub-acts of growing organism X that Sarah can, and those that Sarah cannot, perform as an agent. Some of these sub-acts will spell the difference between trying to grow organism X and actually succeeding at doing so (which she does, but not as an agent). Sarah has deliberative control over whether she washes up, uses a petri dish rather than a dirty lunchbox, places it in a temperature controlled incubator rather than in a hot iron skillet, etc. These are all sub-acts for which Sarah sees contrastive reason. Nevertheless, these sub-acts are not sufficient to succeed at growing organism X. The petri dish must be placed at the *right* temperature, with the *right* chemicals. But Sarah can see no reason to select the right chemicals rather that any other set of (wrong) chemicals. To be clear: I don't mean to say that any action one could perform as an act must be "longish" in this sense, or that for every such action there is a distinction between trying on the basis of contrastive reasons and succeeding on the basis of contrastive reasons. Sometimes one has deliberative control over whether one ϕs (rather than merely tries to ϕ) in simply seeing contrastive reason to ϕ; if we assume that there is such a thing as a "basic action" it could be that seeing contrastive reason to perform it may be enough to perform it as an agent. But sometimes one can see contrastive reason to perform an action without having success rather than a simple try under one's deliberative control. This is illustrated in Sarah's case. And in such cases, to perform the action in question as an agent, one must see contrastive reasons to perform the individual sub-acts that spell the difference between success and failure.

---

[29] Goldman, ch. 2.

Thus my account of what it means to perform an action as an agent runs as follows. One ϕs as an agent to the extent that (a) one ϕs, (b) one sees contrastive reason to ϕ, and (c) one sees contrastive reason to perform any individual sub-acts that constitute the difference between merely *trying* to ϕ and ϕ-ing (if ϕ-ing is a "longish" act, *a lá* Sarah's case). Once these conditions are in place, one ϕs as an agent.

One note before I move on. Some will complain that my account of what it means to ϕ as an agent does not match their considered judgments. Fair enough. However, I argue in the next section that there is independent reason to believe that morality should not treat as requirable actions that could not be performed as an agent *in my sense.* If that's right, then not much will ride on whether my account of the performance of an action on the basis of deliberative agency is the true account. Readers can feel free to treat my account as stipulative, as an account of "schmagency", as it were, if they so choose. The reasons in favor of AGENT (understood in my way) will remain the same.

To begin, note that AGENT—understood in my way—avoids the problem of cognitive limitations. Recall that it seems implausible to say, of Wilda, that (a) she can be morally required to enter 867-5309 and (b) she can be morally required to open the safe. Take, first, entering 867-5309. In this case it is relatively clear that Wilda cannot see any contrastive reason to enter that number. Were she actually to enter the right number, she would not do so on the basis of deliberation, but simply as the result of a guess, a random selection, etc. This is because, as noted before, she does not recognize 867-5309 *as* the safe's combination. Parallel reasoning applies in the case of Jeffrey's moral requirement to type S.

Take now Jeffrey's purported moral requirement to cure cancer. Though Jeffrey can see contrastive reasons to cure cancer, he cannot do so as an agent. Recall that, as in Sarah's case, to perform the action as an agent, one must see a contrastive reason to perform the individual sub-acts that spell the difference between trying to ϕ and ϕ-ing. But insofar as Jeffrey is unaware of the proper procedure to cure cancer, he is unable to see reasons to take the *proper* procedure rather than some alternative *improper* procedure. Because Jeffrey hasn't the foggiest idea how to cure cancer, he cannot *cure* cancer as an agent insofar as he sees no contrastive reason to take the necessary steps rather than any other set of steps. If, against all odds, he actually cures cancer he will not have done so as an agent. He will have *tried*, perhaps as an agent, to cure cancer but—as in Sarah's case—he will not have seen contrastive reason to perform the crucial sub-acts that constitute *curing* (rather than merely trying to cure) cancer. His performance of them would have been the result of mere guesses. Parallel reasoning applies in the case of Wilda's moral requirement to open the safe.

I think a general conclusion is worth stating here. There are many cognitive limitations that we may possess. Some of these limitations will, if we accept TRAD or other theories of moral requirability, render it the case that we will not perform actions we are morally required to perform. But if we accept AGENT, the problem of cognitive limitations is eliminated; given that morally requirable actions are those that an individual can perform as an agent, i.e., as a result of deliberative agency, cognitive limitations *of themselves* will not determine whether we do or do not conform to our moral obligations. The set of actions I can perform as an agent is determined at least in part by what actions I will not perform *given* my cognitive limitations. This set of actions surely excludes any actions that can be performed by me simply as a matter of luck *given* my cognitive limitations. And thus the problem of cognitive limitations—whatever the relevant cognitive limitations are—is dissipated. Under AGENT, for any morally required action I fail to perform, it remains the case that I *could have* performed that action *as a deliberative agent.*

5. Objections

In the final section of the paper, I consider three important objections to AGENT. The first argues that AGENT is *ad hoc*, especially in light of the specialized theory of what it means to act as an agent that I offer in the previous section. The second argues that my view unduly restricts the range of objective moral reasons. The third argues that AGENT is under-inclusive: it rules out the possibility of morally requiring particular actions that really should be requirable.

5.1. Ad Hoc

Perhaps the most important objection to AGENT runs like this: why believe that this account of agency has anything to say about what is or is not morally requirable? Even if, or so it may be claimed, AGENT is successful at turning back the problem of cognitive limitations, what independent reason is there to believe it? Isn't this simply an *ad hoc* addition to our understanding of the boundaries of moral assessment?

I don't think so. Indeed, to recognize contrastive reason to perform an action seems a perfectly sensible constraint on moral obligation independently of the problem of cognitive limitations. Consider the following case:

> *Joey*: Joey is the victim of a dastardly neuroscientist, who has implanted a device in his brain. This device shuts off Joey's ability to see contrastive reasons for or against committing certain acts of aggression for a limited period of time (about a half-hour) on Thursday evenings. Before and after this period, his deliberative capacity operates in perfect working order.

Imagine that Joey kicks someone in the shins at the specified time on Thursday evening. Would we be tempted to say that Joey behaved in a way he morally ought not to have? If Joey has the ability to, say, perform some good deed on Thursday evening, but does not, would we say that Joey has failed in his moral obligations? These suggestions sound wrong to me.[30] And the reason stems from a range of quite general assumptions we make about the role of moral requirement and the nature of moral assessment. Plausibly, moral requirements apply to individuals who have some degree of second-order reasons-based control of their first-order motivations. I take this thought to be roughly platitudinous. But if this is the case, it is incongruous to say that moral requirements cannot apply to individuals who have no reasons-based control of their motivations, but that individuals *can* have a moral requirement to ϕ, even though they have no *reasons-based* control over whether they ϕ rather than not ϕ-ing (given that they do not recognize contrastive reason to ϕ rather than to $\neg\phi$). But if this is correct, AGENT—in my sense—is a natural result. After all, if Joey cannot see contrastive reason to ϕ rather than ψ, he cannot control whether he kicks or doesn't kick *on the basis of reasons.* And hence whether he kicks rather than performing the good deed in question will be the product, at most, of a bare first-order motivation, whim, caprice, or happenstance. And hence AGENT is not *ad hoc.* It is a natural result of a plausible thought concerning the application of moral concepts.

5.2. Objective Moral Reasons[31]

---

[30] Notice that these cases also demonstrate the problem with INTEND+MSC (whether in its revised iteration or not). In this case, Joey could certainly intend to kick someone in the shins (perhaps as a means to a relevant secondary act), and could certainly see that this had worse expected consequences, etc. But it remains implausible to say that he is morally required to refrain from so doing, simply because in this case *whether* he will do so rather than not is not under his deliberative control.

[31] Thanks to Doug Portmore for Roger's case, and a very helpful exchange on this topic.

To see the second objection, take the following case:

> *Roger*: Roger is a member of a crack bomb squad and is in the middle of defusing a bomb that, if left undefused, will destroy a large office building, killing dozens in the process (as well as himself). Roger knows that to defuse the bomb he must either cut the red wire or the green wire. But he has no idea which wire he should cut. He can see no reason to cut the red wire rather than the green wire, or *vice versa.* As it happens, the red wire defuses the bomb.

Several things seem natural to say about this case. First, given that the red wire defuses the bomb, Roger morally ought to cut the red wire. In addition, it seems natural to say that, despite his lack of evidence, there is an *objective* moral reason for Roger to cut the red wire: doing so will save dozens of lives.

But there are two problems for AGENT. First, it would appear that AGENT holds that Roger has no objective moral reason to cut the red wire, insofar as he sees no reason to cut the red wire rather than the green wire. Second, and perhaps more importantly, AGENT seems unable to accommodate our commonsense idea of what it means for a person to have an *objective* moral reason rather than a *subjective* moral reason. To see this, consider the following account of objective moral reasons:

> *Objective Moral Reasons* (OMR): if r is an objective moral reason reason for x to ϕ at t, the true moral theory holds that r is a reason for x to ϕ at t irrespective of x's epistemic states at t.

According to the objection at hand, AGENT treats the class of objective moral reasons as empty. Insofar as AGENT treats moral reasons as indexed to a person's recognition of reasons, they cannot fail to be indexed—at least to some degree—to an individual's epistemic states; whether x recognizes a reason to ϕ will necessarily depend on such states. And if this is the case, there will be no actions that will satisfy the consequent, and hence no actions that will satisfy the antecedent.

I'll take these in reverse order. To hold that AGENT cannot accommodate the everyday conception of objective moral reasons relies on one way of reading a scope ambiguity in OMR. One might ask: what does "irrespective of x's epistemic states" mean? On the wider scope interpretation, for a reason to be objective, this reason can make no mention of x's epistemic states at all. In other words, if the fact or proposition that constitutes this particular reason includes reference to anything about x's epistemic states, this is not an objective reason. But there is another reading. Call the "narrow scope" reading the suggestion that objective moral reasons can, in fact, include reference to epistemic states, but that the *force* of these reasons are not further indexed to x's recognition, or understanding, or knowledge of, these reasons. In other words, objective reasons are determined by "what the reason-constituting facts about [x's] choice happen[s] to be, and so irrespective both of what [x] *take[s]* those facts to be and of what [x's] evidence suggests that those facts might be."[32] This reading takes the true reason-constituting facts as primitive, and hence allows that the "reason-constituting facts" could be determined, in part, by facts about x's mental states, including x's epistemic states. This reading notes only that objective moral reasons are not indexed to what x takes these reasons to be, or what x's evidence suggests these reasons are. Put this another way, the narrow reading holds that r: "x believes q", could be (part of) an objective reason for x to ϕ. But it would not, then, further index the weight of r to x's recognition of r *as* a reason to ϕ.

AGENT can accommodate the existence of objective reasons if read in the latter way. For instance, it could be that x sees a reason to ϕ, but only a prudential reason to ϕ. x's evidence suggests no moral reason to ϕ, and yet AGENT allows that there could be an objective moral reason, even a decisive objective moral reason, to ϕ. Furthermore, the subjective moral significance of ϕ-ing—which is certainly determined in part by an individual's actual or counterfactual epistemic states—could be substantially different than the objective moral significance of ϕ-ing. x's evidence might suggest that r is a very strong moral reason to ϕ, despite the fact that the "true reason-constituting facts" present a weaker moral reason to ϕ. In this way, we can preserve the distinction between *reasons the theory says there are* and *reasons in light of a person's evidence or epistemic states* that is crucial to the distinction between objective and subjective moral reasons at least on the narrow reading.[33] Of course, my view cannot preserve a distinction between objective and subjective moral reasons if OMR is read in the wider scope way. But it seems to me that there is little reason to believe that the conceptual distinction between objective and subjective moral reasons must be read in this way.

---

[32] Douglas Portmore, *Commonsense Consequentialism* (Oxford: Oxford University Press, 2011), 12. My emphasis.

[33] For more on this topic, see Dale Dorsey, "Objective Morality, Subjective Morality, and the Explanatory Question", in *Journal of Ethics and Social Philosophy* 6 (2012).

Even if AGENT can make a sensible distinction between objective and subjective moral reasons, however, we would like to say that Roger has objective moral reason to cut the red wire. I humbly admit that this seems, at first glance, like the right answer. But should we treat this as dispositive reason to accept TRAD rather than AGENT? Consider the following. Imagine that Roger faced not only two wires, but 1000. It seems to me that there is no principled reason to say that Roger has objective reason to cut the red wire in the case above, but doesn't have objective reason to cut the 456th wire (which is the correct one) in the current case. But if this is correct, Roger's case simply becomes (more or less) a version of Wilda's, and hence in suggesting that Roger has objective moral reason to cut the red wire one seems committed to the claim that Wilda has objective moral reason (and hence objective moral requirement) to enter 867-5309. Hence it seems to me that to accept Roger's objective reason to cut the red wire is to commit to a theory of moral requirability (TRAD, or perhaps INTEND) that we should reject for reasons already explored. This is, of course, not to gainsay the contrary intuition. But I think we should regard this considered judgment as a remnant of a theory of moral requirability that, on reflection, we do better to reject.

5.3. Bob

Finally, AGENT might be under-inclusive for the following reason. Consider:

> *Bob*: Bob is a member of a hard-core Satanic cult. Bob believes that drowning children is morally required, and seeks to subject children to drowning whenever possible. In fact, Bob's Satanic indoctrination is such that Bob can see no reason not to drown children.

Now imagine that Bob is faced with the choice of costlessly saving a child from drowning, or letting that child drown. Most would say that Bob is morally required to save the child. But the question now arises: given the structure of Bob's attitudes, could he save the child *as an agent*? Not on my account. After all, *given* his attitudes, if he sees a reason to save the child at all, it will certainly not be a *contrastive reason* (i.e., a reason to save rather than not save). If he ends up saving the child, he will not be doing this as a result of deliberation—given his psychology—but as a matter of mere accident. And if Bob cannot save this child as an agent, surely AGENT is under-inclusive. We certainly wouldn't want to say that Bob can only be morally required to perform actions he sees reason to perform, given his wacko normative beliefs.

But I think this verdict is not as implausible as might first be thought. First, Bob's status *as* an individual who has deliberative control over thought and action is severely compromised given his indoctrination. One might think, even in Bob's case, that moral categories are somewhat less than appropriate given his utter failure as a deliberative agent. In this way, we might compare Bob to Joey: Bob, like Joey at the relevant time on Thursdays, has no capacity to perform the action of saving the child *by means of reasons*. Though Bob can see contrastive reasons to perform some actions (i.e., to drown the relevant child rather than not doing so), a range of actions is simply barred to him *as* a deliberative agent, and this includes the action of saving the child. And hence, like Joey at the relevant time on Thursdays, it seems implausible to say that he could be morally required to save the drowning children, given that *whether* he does so is beyond Bob's capacity to control by means of reasons. Of course, morality will require Bob to perform the morally best of those actions Bob *can* perform as an agent. Nothing about AGENT requires us to assign the moral valence of morally requirable actions for Bob on the *basis* of Bob's chilling normative beliefs.

Of course, we certainly want to say that there is something very morally wrong with Satanists who refuse to save children, even if their beliefs are so entrenched that they cannot do so as agents. But there is a very important sense in which we can accept this verdict even if we accept AGENT. We would want to say that Satanists behaved badly, for instance, if they *developed* Satanic beliefs—beliefs they could have refrained, as agents, from developing. We can morally criticize them for developing their Satanic principles had they the deliberative option to choose other principles, principles the choice of which on their parts would have led to better consequences, or a greater achievement of whatever other moral factor the substantive theory in question deems relevant. Furthermore, if it is within Bob's deliberative control to develop *new* attitudes, he seems morally required to do so. But assume that the fact that these individuals are Satanists is simply beyond their deliberative control. I find it difficult to criticize such individuals morally speaking, so long as they perform actions that are morally best among those they can actually direct themselves to perform as a result of deliberative action rather than luck or mere happenstance.

Of course, we can still make a number of plausible claims about Bob's refusal. We can say it is bad, regrettable, that Bob did not save the children in question. We will certainly seek to retrain Bob or to get him to change his normative beliefs. We will even *despise* Bob, or seek to *blame* him. It may be that any or all of the *reactive attitudes* are appropriate to direct toward Bob. But I see little ground to be gained in saying that Bob is morally required to perform an action he could not have performed as a result of his own deliberation. To say so, it seems to me, is to confuse two very different questions: whether a particular action at a particular time possesses certain properties that we regard, in the abstract, as morally valuable (given a theory of moral valence), and whether those properties generate moral reasons or requirements for particular individuals given the proper theory of moral requirability (which must necessarily depend on a range of factors about the individual in question, including their capacity to perform such actions on the basis of reasons).

I humbly submit that these are my intuitive reactions, but I am certainly not convinced that they will be widely shared. It may very well be that I have uncovered a seriously implausible verdict of AGENT. But the mere fact that some results of AGENT may be counterintuitive does not entail that we should reject AGENT forthwith. Rather, this consequence should be weighed against the implausible verdicts of competitor theories of moral requirability. In particular, one might wonder whether it is worse or better to accept a theory of moral requirability that entails that on virtually any theory of moral valence, one can be doomed to moral failure simply given the hum-drum fact of cognitive limitations. For my money, even if AGENT delivers implausible results in the case of Bob or anyone else, it is far less implausible than simply accepting a view on which we are required to perform actions we will not, given cognitive limitations, perform. And for this reason, it seems to me, we should accept AGENT rather than TRAD, warts and all.

6. Conclusion

The fact of cognitive limitations has very serious consequences for virtually every moral theory if we accept a traditional theory (and some non-traditional theories) of moral requirability. I have argued that a fix for this problem is to adopt an alternative: to insist that x can be morally required to ϕ if and only if x can ϕ as an agent. This proposal solves the problem of cognitive limitations: in determining what someone can do as an agent, his or her cognitive limitations are *factored in.* In addition, AGENT permits of a compelling rationale: AGENT is a natural extension of the plausible thought that moral requirements apply to individuals who have reasons-based control of their first-order motivations.

Of course, the points in favor of AGENT must be weighed against the points against. Such points there are. However, we must be careful not to jettison AGENT too quickly. First, it is likely that many of our considered judgments, especially considered judgments about just what sorts of reasons we have, are heavily influenced by the tacit assumption of TRAD rather than a more restrictive theory of moral requirability. Second, and more importantly, on virtually any theory of moral valence, TRAD, and a wide range of potential alternatives, seems to succumb to the problem of cognitive limitations. Though I am perfectly willing to admit that there may be as yet unexplored options, in weighing the plusses and minuses, AGENT must be taken very seriously indeed.