Some question-begging objections to rule consequentialism

Perl, Caleb

SOME QUESTION-BEGGING OBJECTIONS TO RULE CONSEQUENTIALISM

Caleb Perl

Abstract: This paper defends views like rule consequentialism by distinguishing two sorts of ideal world objections. It aims to show that one of those sorts of objections is question-begging. Its success would open up a path forward for such views.

Keywords: Rule consequentialism, Ideal world, Contractualism, Cluelessness

Derek Parfit (2011) forcefully presents 'ideal world' objections. One such objection is that universal acceptance rule consequentialists predict that pacifism is morally required. For those consequentialists, we're morally required to ϕ iff the rules whose universal acceptance would make things go best require ϕ-ing. But universal acceptance of the pacifist rule *never use violence* would plausibly make things go best. If pacifism is implausible, so too is this kind of rule consequentialism. Even worse, even pacifists should find this explanation implausible. The explanation treats actual violence as *irrelevant*: the view idealizes and erases it.

A burgeoning consensus takes the ideal world objection to doom several core views, contractualist and Kantian as well as rule consequentialist. I'll call the target views 'pattern-dependent views', because they ground moral facts in facts about patterns of group behavior.[1] Some modify their views specifically to avoid the objection, including both Parfit himself (2011) and Michael Ridge (2006). But Gideon Rosen (2009) and Abelard Podgorski (2018) generalized the objections to those views as well. Rosen and Podgorski take their generalized objections to doom *all* pattern-dependent views.

I aim to disrupt the burgeoning consensus by distinguishing two sorts of ideal world objections. I grant that the two sorts of objections would *together* doom all pattern-dependent views. But the two sorts of objections spring from different sources. I'll argue that one sort of objection is much more compelling than the other. My

[1] Christopher Woodard (2008, 2019) similarly emphasizes 'pattern-based' reasons; it's such a good label for the class of views that I'm echoing his use.

success would open up a path forward for pattern-dependent views like rule consequentialists. The path forward remains fraught. Some versions of the ideal world objection remain forceful. But *some* kinds of pattern-dependent views avoid the genuine objections.

1 Reflective equilibrium and ideal world

Pattern-dependent views often emphasize that agents who coordinate their behavior can achieve better outcomes than they could achieve without coordinating; I'll say that they emphasize the *benefits distinctive to cooperative patterns.* Emphasizing those benefits helps explain how morality is something that we can intelligibly care about. We can intelligibly care about the benefits distinctive to cooperative patterns, because we can intelligibly care about what happens to us. So a core advantage of pattern-dependent views like rule consequentialism over rivals like Rossian pluralism is that they give a unified account of what ties all our moral judgments together – benefits of cooperative patterns – and thereby grounding moral facts ultimately in something we can intelligibly care about.

This paper focuses on rule consequentialism as the simplest pattern-dependent view. Other pattern-dependent views, like Scanlon's contractualism (1998), add other features of cooperative patterns as mattering too. This paper's lessons about rule consequentialism will generalize to these more complicated cousins.

Legal systems illustrate benefits distinctive to cooperative patterns. It's best if we all defer to shared legal norms rather than trying to follow whichever legal norms we individually judge to be best, as a case from Judith Jarvis Thomson illustrates:

> You are a sheriff in a small southern town. A murder has been committed, and you do not have the least idea who committed it, but a lynch mob will hang five others if you do not fasten the crime to one individual. (Thomson 1996, 50n2)

We judge that it's (objectively!) wrong for me to scapegoat in this case, even if I can get away with it. Act consequentialists struggle to agree. If I can get away with scapegoating, scapegoating could easily be the action with the best consequences: four more people survive.

In contrast, pattern-dependent views like rule consequentialists can vindicate our judgment that scapegoating is wrong. They'd emphasize the value of a stable, trustworthy justice system. It's valuable because people then trust the justice system to adjudicate disputes, rather than avoiding it because of its capriciousness. And a society where everyone trusts the justice system is much better off than one where they don't. Crucially, though, a trusted justice system isn't the result of any one person's actions. A single sheriff by himself can't collapse trust in the justice system, nor can he repair trust once lost. A trusted justice system is instead the result of a *pattern* of actions. So if the great good of a trusted justice system bears on what the sheriff objectively ought to do, benefits distinctive to cooperative patterns must help determine what's objectively required. (In contrast, act consequentialists can take benefits distinctive to cooperative patterns to affect what's *subjectively* required.)

Contemporary rule consequentialists defend their view as putting our considered moral judgments in reflective equilibrium. Hooker says: "the best argument for rule-consequentialism is that it does a better job than its rivals of matching and tying together our moral convictions, as well as offering us help with our moral disagreements and uncertainties" (Hooker 2000, 104). Past rule consequentialists have defended their view differently. Harsanyi, for instance, defended rule consequentialism by appealing to "its ability to take proper account of the implications that alternative systems of possible moral rules would have for people's expectations and incentives (expectation and incentive effects)" (Harsanyi 1982, 58). This justification for rule consequentialism threatens to collapse it into act consequentialism. It justifies actions by appeal to what's good but sometimes tells us to *not* bring about what's good (Arneson 2005). Even worse, act consequentialists can pick *decision procedures* to respond to expectation and incentive effects without abandoning the basic justification for the view (Railton 1984). This paper focuses just on contemporary pattern-dependent views, which center the method of reflective equilibrium. Their use of the method helps them avoid Harsanyi-style collapse (Hooker 2000, 93ff).

1.1 Ideal world objections from supervenience failures

This paper aims to distinguish two sorts of ideal world objections. Considering one often makes you consider the other. But the two objections have very different sources.

Parfit's original ideal world objection illustrates the first kind of objection. It applies to views that depend on facts about what *would* happen in certain ideal possibilities, which includes Kantians as well as rule consequentialists. He notes that "Kant's formula mistakenly requires us to act in certain ways even when, because some other people are not acting in these ways, our acts would make things go very badly, and for no good reason" (Parfit 2011, 314). And he generalizes the problem to universal acceptance rule consequentialism (UARC), noting that it enjoins pacifism for the wrong reasons.[2]

Parfit himself develops a modified version of rule consequentialism that surmounts his own objection. Parfit's variant enjoins the rules that do the best at *any* level of adherence (Parfit 2011, 317). Since the pacifist rule does *not* do best at lower levels of adherence, Parfit's variant avoids the objectionable prediction that pacifism is morally required. His suggestion works by centering a wide range of *embeddings* – a wider range of 'tests' on the moral significance of the rules. Rules themselves don't have consequences. Consequences are consequences of particular *embeddings* – for example, embeddings where [everyone/an overwhelming majority/more than half...] [adhere to/internalize/accept/...] the rules (Kagan 1998, 227). Parfit takes the ideal world objection to show that no *single* embedding matters uniquely.

Gideon Rosen generalizes the original ideal-world objection to show that no embedding or class of embeddings could be morally significant. Rosen formulates his objection by imagining *gremlins*:

> The gremlin will do his worst if unanimity is achieved, unless the consensus embraces the moral system of (say) the Roman aristocracy under Caligula. It is then a nontrivial fact about this world that this repugnant moral system is the system whose universal acceptance would make things go best. (UARC) entails that if the world contains this sort of gremlin, it is morally permissible to beat one's slaves, even if the gremlin will never make his presence felt. But that's absurd (Rosen 2009, 86).

[2] Can UARC avoid this problem by incorporating a 'disaster avoidance' rule? Maybe it'd vindicate a rule that permits using violence if others are violent, in combination with a rule enjoining pacifism if others are pacifists. Unfortunately not: UARC must classify those two rules as *tied* with a simpler rule enjoining pacifism. At the embedding where everyone accepts the same rule, the simpler rule does just as well as the more complicated set of rules that includes the disaster avoidance rule. The basic problem is that UARC then ends up *indeterminate* – it simply doesn't tell us whether to be a pacifist. Holly Smith spends five pages arguing that this problem can't be solved (2010, 421–426); she tackles the responses that you may be imagining, and shows that they fail. In addition, Abelard Podgorski (2018, 286) describes an ideal world objection that dooms this strategy: a gremlin could cause disaster if enough people accept conditionalized rules but not otherwise. And that objection is one of the legitimate ones, even if this paper is right.

Rosen plausibly suggests that this gremlin can't affect what we're permitted to do. He objects that Parfit's view mistakenly allows that it could: the gremlin affects what rules do best at any level of adherence by affecting what does best at universal adherence. This gremlin can't affect what we're permitted to do because it doesn't change anything that *actually* happens. Rosen takes this sort of gremlin to eliminate any view that allows merely *possible* patterns of action to determine what's obligatory. Gremlins could change how the merely possible patterns go without changing our actual moral obligations. And that seems intuitively implausible.

This objection can show that its targets must deny plausible theses about *supervenience*. Supervenience theses look something like conceptual truths; Rosen (2020) elsewhere introduces a supervenience thesis as the least controversial thesis in metaethics.[3] One thing Rosen's initial objection shows is that its targets must deny that moral facts *supervene* on the consequences of agents' actions. After all, the views Rosen targets predict that the gremlins can change the moral facts without changing any actual consequences.

Now some theories can intelligibly deny that moral facts supervene on the consequences. For instance, the doctrine of double effect denies moral facts supervene on the consequences, since differences in *intention* needn't always affect the consequences. (A soldier could bomb the innocent with the intention of killing them or just with the intention of following orders – the consequences would still be the same.) But supervenience on the consequences is plausibly part of a 'foundational consequentialist thought.' Although non-consequentialists can intelligibly deny that morality supervenes on the consequences, we might doubt that *consequentialists* in particular can.

Most importantly, though, Rosen's targets must reject an especially uncontroversial supervenience thesis:

> Weak Supervenience: nonnormative duplicates cannot differ in normative respects (Rosen 2020, cf 228).

Weak Supervenience should be uncontroversial. For one thing, it's compatible with the doctrine of double effect. Nonnormative duplicates have the same intentions. So Weak Supervenience remains true even if intentions affect what's permissible. Rosen's gremlins show

[3] R. M. Hare gives a classic discussion (Hare 1952, 145). Debbie Roberts (2018) and Rosen give a critical discussion. But Rosen at least is open to the weak supervenience thesis two paragraphs from now.

that his targets reject Weak Supervenience. My nonnormative duplicate could coexist with one of Rosen's gremlins, since the gremlin doesn't affect my mental states or the consequences of my actions. And Rosen's targets predict that the gremlin could change what's obligatory for my nonnormative duplicate. Even though Rosen's gremlins are silly, they illustrate failures of supervenience that look very troubling.

Rule consequentialists might hope to avoid Rosen's problems by centering *actual* patterns of action, as Conrad Johnson (1991), Richard Miller (2009), and Caleb Perl (2021) all suggest. After all, we already act in patterned ways, as our legal systems illustrate. Centering actual patterns of action preserves the crucial supervenience theses. But doing so faces new objections.

1.2 "Active" ideal world objections

Views that center actual patterns face objections from Rosen-style gremlins that are *actually* active.

> Active Gremlin The gremlin creates a new disaster for each person who fails to embrace the moral system of (say) the Roman aristocracy under Caligula.

Rosen could insist that this gremlin doesn't affect what's permissible, and reject any view that takes actual adherence to matter. He might grant that everyone subject to the gremlin should *believe* in the moral system of the Roman aristocracy, to the extent that they can, and even that they should believe in it for moral reasons. But he'd insist that they have moral reasons to have false moral beliefs, perhaps comparing it to the ways that rationality might sometimes require us to act irrationally (Parfit 1984). For pattern-dependent views that center actual patterns, though, the active gremlin does affect the truth about morality: the gremlin *makes* the Roman moral system the true moral system. Active gremlins then seem to provide good evidence against those pattern-dependent views, too.[4]

Active gremlins vividly illustrate a broader class of related objections. One related objection starts by emphasizing that we should ordinarily keep our promises even in a Hobbesian world of interminable war of all against all – a world where cooperative patterns never exist. A rule consequentialist who grounds moral facts only in actual

[4] Podgorski points out that views like Johnson's and Miller's face problems with active and inactive gremlins both, in effect (Podgorski 2018, 291).

patterns predicts otherwise, since the pattern that makes promise-keeping obligatory doesn't exist in that Hobbesian world. Rossians could press this objection by insisting that promise-keeping remains obligatory. They might then use judgments about the Hobbesian world as evidence against the pattern-dependent view. This Rossian challenge illustrates the same objection as active gremlins without the gremlins gimmick.

This paper aims to diagnose the source of this second sort of objection. Objections from active gremlins cannot rest on the supervenience theses that ground the objections noted in §1.1. After all, they envision changes to what actually happens.

Diagnosing the sources of these objections matters because the two sorts of objections jointly eliminate *all* pattern-dependent views. Pattern-dependent views can ground moral facts only in *actual* patterns, or they can ground them in merely possible patterns. If they ground them in merely possible patterns, they'll mistakenly allow that inactive gremlins can change what's right. And if they ground moral facts only in actual patterns, they'll mistakenly allow that *active* gremlins can change what's right. Abelard Podgorski (2018) gives this challenge with particular force; I'm giving a schematic presentation of his powerful argument. Most rule consequentialists have struggled with versions of this challenge, though Rosen and Podgorski give particularly vivid presentations; Holly Smith (2010) gives a sharp version of a related challenge.

My diagnosis of the objections described in this section will rest on a fairly new notion: the notion of an *transparent method* for forming moral beliefs. This notion plays a starring role in what follows; understanding the argument requires understanding it. To preview: I'll argue that objections from active gremlins must assume that we have transparent methods. Then I'll show that rule consequentialists must already deny that we have transparent methods for altogether independent reasons.

Roughly, transparent methods are methods that reveal how pure moral facts depend on empirical facts. Transparency comes in degrees. If universal acceptance rule consequentialism is true, a maximally transparent method would reveal how the moral facts would change given *any* changes in what'd make universally accepted rules go best. *Maximal* transparency would reveal exactly which theory is true. It'd reveal how the moral facts would vary systematically across all the different ways that things could go best. It'd thus reveal what kind of rule consequentialism captures the systematic variation. A maximally transparent method would be very striking and very powerful.

Other kinds of transparency are more moderate. One kind would involve understanding how the moral facts would change given a *narrower* range of changes – for example, given changes in our material conditions from moderate scarcity of resources to severe scarcity, say, but not given changes to human nature. A distinct kind of moderate transparency would involve a less *articulated* understanding of the dependency. For instance, a transparent method might reveal that either universal acceptance rule consequentialism or Parfit's alternative is true – it might reveal how the moral facts would vary in the cases where the two kinds of rule consequentialism would agree. But it might also fail to reveal how the moral facts would vary in the cases where the two kinds of views disagree.

I intend to use "transparency" so that Rossians would deny that we have *any* transparent method. Rossians deny that *pure* moral facts depend on further empirical facts. They do allow that *im*pure moral facts depend on further empirical facts. For instance, the permissibility of a particular lie depends on empirical facts about the lie itself. But the pure moral facts about lies don't depend on empirical facts. In contrast, rule consequentialists do take pure moral facts to depend on empirical facts: consequences of embedded rules determine the pure moral facts. Rossians would thus insist that my kind of transparency is *im*possible.

2 A Puzzle Paired with a Millian Solution

I just characterized transparency abstractly. This section fleshes out the abstract notion by describing a puzzle for contemporary rule consequentialists, and introducing a Millian solution to that puzzle. Then I show that the Millian solution disarms the second sort of ideal world objection.

Rule consequentialists ground moral facts in substantive empirical facts. Development economists might investigate those substantive facts. For instance, a development economist might find herself interested in the following three rules.

> R0: you may sacrifice someone's toe to save another's life *only with* the toe-owner's consent.
>
> R1: you may sacrifice someone's toe to save another's life *even without* the toe-owner's consent.
>
> R2: you may sacrifice someone's toe to save another's life *only if you expect* that the toe-owner would consent.

The economist might be interested in how we'd behave if we expected others to have internalized R0 rather than R1, and R1 rather than R2. We might take more risks, secure in our own bodily integrity. Or we might take fewer risks, because we know we couldn't rely on others to save us. Maybe, for instance, our economist can find societies with different expectations, and work to identify the difference that the expectations make. The economist could develop an empirical method for comparing these rules. (I focus on development economist just for concreteness. I could use other kinds of economists, or even sociologists, to illustrate the point.)

I'm not a development economist. Even still, I might have strong opinions about R0–R2, and articulate reasons why one rule is better. But I don't have an empirical method for comparing them. If you're tempted to disagree, ask yourself whether a serious economist has any reason to ignore her empirical work in favor of whatever reasons I'm able to articulate. I don't think she does. I infer that she can have an empirical method for comparing rules that I don't have. I'll say that I don't have the ability to go "toe-to-toe" with development economists – meaning that the reasons that I myself am able to articulate in comparing R0–R2 aren't reasons that the economist needs to consider in addition to their empirical work.

Even though I can't go toe-to-toe with development economists, I still make moral judgments that rule consequentialists would take to reflect comparisons between the rules. Parfit gives one example:

> Third Earthquake: You and your child are trapped in slowly collapsing wreckage, which threatens both your lives. (Parfit 2011, 222). ... In Third Earthquake, you cannot save your child's life except by crushing Black's toe, without Black's consent. This act, I believe, would be justified (Parfit 2011, 231)

He uses our belief here as evidence against Kantian strictures on treating agents as mere means. Parfit assumes that we can have *justified* beliefs about Third Earthquake that provide evidence.

Our inability to go toe-to-toe with development economists creates a puzzle for contemporary rule consequentialists like Parfit. Since those consequentialists use our considered moral judgments as evidence, they need to explain how they provide evidence even though we can't go toe-to-toe with development economists. As a first-pass illustration of the need, imagine that I learned that my considered moral judgments are systematically out of sync with the rule consequentialists' complex empirical facts. Then I shouldn't use those considered moral judgments as evidence for rule consequentialism. The problem wouldn't be that we *lack* evidence that we have a reliable method that tracks the rule consequentialists' facts. Instead, our inability to go toe-to-toe with development economists threatens to provide *positive* evidence that we lack a reliable method.

Mill gives one classic strategy for handing our ignorance:

> Defenders of utility often find themselves called upon to reply to such objections as this – that there is not time, previous to action, for calculating and weighing the effects of any line of conduct on the general happiness. ... The answer to the objection is, that there has been ample time, namely, the whole past duration of the human species. During all that time mankind have been learning by experience the tendencies of actions; on which experience all the prudence, as well as all the morality of life, is dependent. ... that mankind have still much to learn as to the effects of actions on the general happiness, I admit, or rather, earnestly maintain. (Mill 1863, 420)

Mill likely wasn't a rule consequentialist, but rule consequentialists could emphasize similar points. Such rule consequentialists would insist that our *community* can track the consequences of embedded rules, even if we individually cannot. Perhaps our community has converged on the best rules through generations of experience, so that intuitions/ considered judgments reflect the best rules, even though we can't go toe-to-toe with development economists in showing that they do. We'd have a derivative ability to track the consequences of these rules, derivative from our community.

The Millian strategy explains how we can rely on our considered judgments even though we can't go toe-to-toe with development economists. We can appropriately rely on our considered judgments because they reflect our *community's* convergence on what's best. But I'm not individually in a position to articulate the reasons why my community has converged on those judgments. So I can't go toe-to-toe with development economists because I can't myself articulate those reasons.

The Millian strategy disarms objections from active gremlins. That objection gives a Simple Argument against a view P that grounds moral facts in actual patterns:

Simple Argument:

1. Rosen's active gremlins couldn't make beating slaves permissible.

2. If P is true, Rosen's active gremlins could make beating slaves permissible.

Conclusion: P isn't true.

The Millian strategy classifies Premise 1 as a persistent illusion. The history of our community wouldn't track the activity of any of Rosen's active gremlins, since our history doesn't include any active gremlins. Our history instead includes lots of mistaken rationalizations of slavery that make us especially leery of rationalizations of slavery. That's why the Millian strategy would make Premise 1 a persistent illusion. The Millian strategy only justifies moral judgments about cases *where the empirical facts remain the same*, and Rosen's active gremlins change the empirical facts.

The Millian strategy classifies Premise 1 as *unjustified* for creatures like us. Then the Simple Argument would misfire because we lack justification for the premises that could transmit to the conclusion. We might still remain very confident that Premise 1 is true. The Millian strategy would explain that confidence as a projection from our past. We've converged on an absolute prohibition on slavery: we don't think it's *ever* permissible. Since we accept that absolute prohibition, the Millian strategy predicts that we wouldn't find slavery permissible even given active gremlins. We'd lack a transparent method capable of responding to active gremlins.

The Millian strategy treats our overwhelming confidence in Premise 1 in the way we should treat persistent optical illusions. Pencils appear bent in glasses of water – but we know that appearances deceive. But that knowledge doesn't undermine the appearance. The pencil still looks bent. The Millian strategy treats intuitions about active gremlins similarly. If it's correct, we'd remain confident that active gremlins don't change our moral obligations, but our confidence wouldn't provide evidence against rule consequentialism.

The Millian strategy illustrates how we might *fail* to have a transparent method. It takes our community to track facts about embedded rules, even though individuals can't. Our ordinary moral beliefs remain justified. An *external* feature – the reliability of our community – justifies them. Since an external feature justifies, individuals are not in a position to see how pure moral facts depend on empirical facts.

3 Generalizing

The rest of this paper asks in general if we have transparent methods for forming moral beliefs. In asking, I'm going to be granting that development economists can have empirical methods for comparing rules, and asking what we can learn about our *moral* belief-forming methods from our inability to go toe-to-toe with them.

This question matters because the Simple Argument fails for *anyone* who lacks a transparent method. Transparent methods reveal how the permissibility of beating slaves *depends* on empirical facts, like the existence of active gremlins. Absent a transparent method, we couldn't tell how active gremlins change what's morally permissible. We might still have *confident* beliefs: that slavery remains wrong. But those confident beliefs won't be evidence one way or another. The Millian strategy illustrates one way that the Simple Argument could fail.

This section argues that all contemporary rule consequentialists must already deny that our moral judgments rest on transparent methods. This argument would show that the Simple Argument is question-begging, resting on assumptions that rule consequentialists must already reject.

3.1 With unjustified initial judgments

I argue by case. I begin with *constructivist* views where our initial considered moral judgments needn't be justified at all. Those views don't require transparent methods, thereby disarming arguments from active gremlins. Then I turn to views that require starting with *justified* judgments. I show that those views must also deny that our judgments rest on transparent methods.

Some Rawlsians allow that the method of reflective equilibrium can confer justification on previously unjustified judgments. Rawls distinguishes 'narrow' and 'wide' reflective equilibrium; the former achieves coherence within our moral judgments, while the latter achieves coherence with our moral and non-moral judgments both (Rawls 1974). Those who think reflective equilibrium can confer justification should think that it's *wide* reflective equilibrium that confers justification. If *narrow* reflective equilibrium did, narrow reflective equilibrium would let us go toe-to-toe with a development economist. After all, rule consequentialists want their view to put my purely moral considered judgments in reflective equilibrium. If I learned that it did and thought that my judgments were justified, I seem able to go toe-to-toe with development economists. But that conclusion should look implausible. Since narrow reflective equilibrium doesn't advance my understanding of the empirical facts, it still wouldn't allow us to go toe-to-toe with development economists.

In contrast, wide reflective equilibrium *could* let me go toe-to-toe with a development economist. I could go toe-to-toe because wide reflective equilibrium itself *involves* the same sort of empirical method that development economists use. It involves fitting our considered moral judgments with non-moral beliefs, including the sort of beliefs that development economists seek to justify.

Philosophers who allow that the method of reflective equilibrium can confer justification are metaethical *constructivists*. They take our starting judgments to constrain the final equilibrium by revealing the sort of *problem* that we need to address, as Christine Korsgaard explains: "the most general normative concepts, the right and the good, are names for problems – for the normative problems that spring from our reflective nature" (Korsgaard 1996, 114). Since the starting judgments reveal the problems we need to address, they constrain the final equilibrium even if they're initially unjustified. Crucially, though, our problems arise from our *actual* circumstances, which do not include the sort of active gremlins that Rosen imagines. That's why Rawls insists that "there is no objection to resting the choice of first principles upon the general facts of economics and psychology" (Rawls 1971, 137), and why he vigorously resists fanciful counterexamples (Rawls 2001, §19). And Rosen concedes that constructivism disarms his objections (Rosen 2009, 80).

The constructivist strategy illustrates the same point as the Millian strategy. The point is that judgments about active gremlins are question-begging absent transparent methods. The constructivist strategy grounds the justificatory role of considered moral moral judgments in their revealing the problems that need solving. The constructivist strategy thus resembles the Millian strategy in rejecting transparent methods. So we shouldn't be surprised that it also disarms objections from active gremlins. Rather than further developing the constructivist or the Millian strategies, though, the rest of the paper turns to establishing a general point. The general point is that *all* contemporary rule consequentialists must disavow transparent methods. That point matters because objections from active gremlins *require* transparent methods.

3.2 With justified initial judgments

We now focus on theorists who assume that they're *starting* with justified considered moral judgments. This assumption looks natural after setting constructivism aside.

Contemporary rule consequentialists defend their theory as putting our considered judgments in reflective equilibrium – for example, putting judgments about Third Earthquake in equilibrium with other judgments. Non-constructivists should deny that unjustified moral judgments provide evidence for the view. If they admit that they started with unjustified judgments, they should see their resulting castle as built on sand. I'll keep calling philosophers who agree with this whole paragraph the 'contemporary' rule consequentialists. (Extant rule consequentialists disavow constructivism, or at least don't want their view to *require* constructivism.)

Contemporary rule consequentialists need an account of our justified judgments that explains why I can't go toe-to-toe with development economists. I'll now argue that any such account predicts that we lack transparent methods. I'll do that by arguing for a pivotal thesis that I label Link:

> Link: *if* our initial considered judgments incorporate a transparent method *and* are themselves justified, development economists *shouldn't* feel free to ignore what I can articulate.

This paper succeeds if it establishes Link. Link forces contemporary rule consequentialists to insist that we lack transparent methods, by *modus tollens.* After all, §2.2 already noted that economists *should* feel free to ignore what I can articulate. And I'm now focusing on contemporary rule consequentialists like Parfit and Hooker, who aren't constructivists and who agree the starting judgments do need to be justified.

To establish Link, suppose its antecedent. That is, suppose that my initial considered judgments incorporate a transparent method *and* are themselves justified. For instance, the supposition would mean that I'm justified in judging that I may sacrifice your toe in Third Earthquake. Then imagine that I become fully confident that rule consequentialism is true. (Rule consequentialists should allow that I could become fully confident – presumably they aim to get us to that point!) If I see myself as having a transparent method and see my judgment about Third Earthquake as justified, I should *deny* that development economists are free to ignore what I can articulate. My transparent method would reveal which embedded rules ground the permissibility of the toe-sacrifice, transmitting moral justification into empirical justification.

We can illustrate the abstract argument for Link by returning to the Millian strategy. That strategy explains why moral justification doesn't transmit to empirical justification. The Millian strategy takes the history of our species to *enable* moral justification even absent justification that that enabling condition is met. James Pryor gives a helpful parallel.

> Suppose you're reading some proof of the Pythagorean Theorem. H1 is the claim that you understand and correctly follow the proof. Presumably, for you to be justified in believing the theorem, H1 does have to be true. But you don't need to have evidence that H1 is true. It's the proof itself that justifies you in believing the theorem. H1 is just some condition that enables this to happen. It's not itself one of the premises that your justification for believing the theorem rests on-not even a suppressed, background premise. (Pryor 2004, 354)

Understanding the proof is a condition that enables justification, though the justification doesn't itself require justification that the enabling condition is met. The Millian strategy takes the history of our species to enable justification, though the justification doesn't itself require justification that the enabling condition is met. As best as I can tell, vindicating moral justification without transparent methods requires distinguishing *justification* from the *enabling conditions for* justification. That's why the Millian strategy illustrates what *all* contemporary rule consequentialists need: an account of justified considered judgments, *without* transparent methods. Otherwise they end up allowing that I could go toe-to-toe with development economists.

The need to avoid transparent methods constrains how we should interpret extant rule consequentialist responses to our empirical ignorance. For instance, Brad Hooker concedes that actual consequences of embedded rules "are too difficult (indeed, effectively impossible) to find out" (Hooker 2000, 72), and instead appeals to *reasonable expectations* about their consequences. In conceding that the consequences are too difficult for ordinary agents to discover, he's admitting in part that he can't go toe-to-toe with development economists. He might also think development economists can't find them out. I use those economists only to establish our empirical ignorance, which he grants.

Hooker veers closer to positing a transparent method. One way of reading his strategy is as requiring that moral judgments depend on reasonable expectations about the consequences. But he shouldn't think that the reasonableness of our moral judgments *depends on* the reasonableness of our expectations about the consequences. If he did, he'd be committed to thinking that it's "too difficult, indeed effectively impossible", to learn that rule consequentialism is true – learning that rule consequentialism is true would then transmit into learning the empirical facts, too. That's just what my general Link thesis means.

Hooker should instead accept some more modest thesis, perhaps only that our considered moral judgments need to be capable of *cohering* with *actual* reasonable expectations. That thesis allows that we could learn that rule consequentialism is true, without learning the empirical facts that are too difficult to discover. The more modest thesis doesn't incorporate a transparent method. Instead, it treats the *availability* of actual reasonable expectations as an enabling condition: were actual reasonable expectations *un*available, our judgments would be unjustified. But since they're available, our judgments remain justified, independently of our empirical abilities. Crucially, though, the more modest thesis also disarms arguments from active gremlins. It doesn't take our moral judgments to rest on a transparent method – so active gremlins won't be evidence one way or another.

4 Prioritize the deeper questions

This paper distinguishes two different sorts of ideal world objections: those that appeal to *in*active gremlins, and those that appeal to active gremlins. It aims to show that the latter sort of argument requires transparent methods. I've tried to show that contemporary rule consequentialists must already insist that we lack transparent methods. Arguments from active gremlins are then question-begging because they require an assumption that contemporary rule consequentialists already must reject.

For all I've said so far, rule consequentialism may still be indefensible. I've identified an assumption that rule consequentialists must reject. But I haven't shown that they *can* defensibly reject it. This paper aims instead at helping us focus on the deepest questions about rule consequentialism. It aims to establish that questions about transparency are deeper than questions about active gremlins. They're deeper because answers to the former questions also answer the latter questions. This paper succeeds even if it's *impossible* for rule consequentialists to do what they need to do: explain how moral judgments don't rest on transparent methods. This paper succeeds by focusing attention on the deeper problem, even if the deeper problem proves unsolvable.

To see how the deeper problem might prove unsolvable, consider a spectrum of agents:

- Ignorant Igor, who lacks *any* (propositional) justification about the consequences of embedded rules.

  ...

- Ordinary agents like you and I,

  ...

- God, whose moral-belief forming methods rest on a transparent method for comparing rules

Caleb Perl (2019) has argued that the method of reflective equilibrium *couldn't allow* Ignorant Igor to learn that rule consequentialism is true – at least once he's aware of his empirical ignorance. He should instead recognize that his empirical ignorance defeats his considered moral judgments. And that recognition should prevent him from seeing those judgments as supporting rule consequentialism. Perl focuses on contemporary rule consequentialists, setting aside constructivists and views like Harsanyi's.

I concede that we could turn out to be like Ignorant Igor. We could lack transparent methods and *also* lack externalist features that support propositional justification. Extant discussions of our empirical ignorance focus on a much more extreme challenge that James Lenman (2000) has developed. He emphasizes that embedded rules are *identity-affecting*: different people would exist on some embeddings than others. If, for instance, everyone always accepted rules forbidding toe sacrifices without consent, maybe one of Hitler's great^n-grandparents didn't exist. So Hitler's actions might be the consequences of an embedded rule permitting those sacrifices. Or someone much worse than Hitler – Lenman's 'Malcolm the Truly Appalling' – might exist only on some embeddings. We're unable to tell what would happen; we're ignorant of the consequences.

Lenman's argument might establish that we lack an empirical method for comparing rules. He can more easily establish ignorance about the consequence of token actions than about the consequences of rules, since statistical reasoning may help with the latter but not the former. But his argument may also establish ignorance about the latter, too. If you think it does, you agree with me about what needs prioritizing: rule consequentialists should prioritize explaining how our moral judgments can be evidence absent transparent methods. Then they must see intuitions about active gremlins as projections from our actual convictions.

But I haven't foregrounded Lenman's challenge, because some dismiss it as a kind of skeptical challenge. After all, it's tempting to formulate Lenman's point with closure principles: if we knew that embeddings of toe-sacrifice-permitting rules were best, we'd know that Malcolm the Truly Appalling wouldn't exist on those embeddings; we don't know that Malcolm wouldn't exist, so we don't know that those embeddings are best. I've foregrounded development economists rather than Lenman's challenge to head off this reaction. The comparison with development economists doesn't invite a skeptical construal, since it *assumes* that they can have genuine knowledge. I myself don't see Lenman's challenge as a skeptical one. I think it also succeeds without skeptical assumptions.

Lenman's argument may then doom rule consequentialism. Contemporary rule consequentialists use our considered judgments as evidence. And Lenman's argument might show that our considered judgments are *defeated* if rule consequentialism is true. Then we're like Ignorant Igor. The paper still advances our understanding of contemporary rule consequentialism, by centering the most fundamental problems.

Lenman's argument helps foreground a distinction between two different kinds of belief-forming mechanisms: *transparent methods* for forming moral beliefs and methods for forming *empirical* beliefs about the consequences. I've suggested that ordinary finite agents like you and I don't have the latter, and inferred that we don't have the former. I used development economists as evidence that we don't have those empirical methods. If we did, we could go toe-to-toe with development economists. Lenman's argument provides another kind of evidence that we don't have those empirical methods.

If we set Lenman's argument aside, though, we might worry that development economists still could have transparent methods, since they have the relevant empirical methods. This worry is misguided. Those economists made the pivotal considered moral judgments even before they develop their empirical methods – judging, for instance, that promise-keeping is usually required. Those moral judgments cannot depend on the empirical methods they develop later, since they came first. Development economists might also insist that active gremlins don't affect our moral obligations. Their insistence reflects the ways that their moral-belief-forming mechanisms remain *unintegrated* with their empirical methods for comparing rules. My fingers can feel that a pencil submerged in a glass of water is straight even while it still looks bent; my vision and touch remain unintegrated. The same remains true for the economist's moral-belief forming mechanisms. Those mechanisms remain unintegrated with her empirical methods.

My pivotal claim is that we must lack transparent methods. The claim is only about *finite* creatures like you and I, since we have a limited ability to track the consequences. God is different. Rule consequentialists should insist that God would take active gremlins to affect our moral obligations, and not just what we should believe about them. You probably find this response incredible – you probably think that God *wouldn't* take active gremlins to affect our obligations. My Feuerbachian diagnosis is that your confidence just reflects projection from our finite epistemic situation. We'd gain new evidence if God told us that active gremlins couldn't affect our moral obligations. Then this paper would become irrelevant. This paper addresses only our actual epistemic situation, without divine deliverances on active gremlins.

5 Generalizing to other pattern-dependent views

My overt agenda is to show that some instances of the ideal world objection are question-begging. They're question-begging because they appeal to judgments that'd be persistent illusions if rule consequentialists are right. Rule consequentialists would then have a path forward. We should ground moral facts in *actual* patterns, rather than merely possible patterns. Conrad Johnson (1991), Richard Miller (2009), and Caleb Perl (2021) develop different strategies for doing so – others should too. Though those strategies conflict with intuitive judgments about active gremlins and Hobbesian worlds, those intuitive judgments aren't genuine evidence.

But I also have a covert agenda. I want to convince rule consequentialists to articulate our implicit methodological assumptions more explicitly. I've shown that the absence of transparent methods brightens our prospects. I haven't explained why we lack a transparent method, though I've indicated some options. For what it's worth, I think that all the options discussed here *fail* – they don't do what rule consequentialism needs. I have my own explanation of why our considered judgments can justify rule consequentialism even absent transparent methods, sketched in my (2017). But in developing my own explanation, I realized that it illustrated a highly general point.

This paper factors out the highly general point from the baroque details of my own explanation.

I've focused on rule consequentialism as the simplest pattern-dependent view. The lessons of this paper generalize to other pattern-dependent views like Scanlon's contractualism. He holds that

> if, for example, I lived in a desert area and were obligated to provide food for strangers in need who came by my house, then I would have to take account of this possibility in my shopping and consumption, whether or not anyone ever asked me for this kind of help (Scanlon 1998, 203).

Parfit, Rosen, and Podgorski all emphasize that ideal world objections apply to contractualists as much as rule consequentialists, since they also center group patterns of behavior as Scanlon does here.

The empirical facts that matter for Scanlon are more complex than the facts that matter for rule consequentialists. Scanlon allows that patterns can be morally significant because of their effects on each of us. Unlike rule consequentialists, he gives a *non-aggregative* account of the moral significance of patterns. I can reasonably reject a principle because of the effect the correlative pattern would have on me, and you can too. Doing so requires sensitivity to the same empirical facts that matter for rule consequentialists. But Scanlon must disavow certain kinds of simplifying dominance reasoning. Merely showing that the summed benefits of one policy dominates another isn't enough for him. He also requires consideration of each person's perspective.

Scanlon also should dismiss intuitions about active gremlins or about Hobbesian worlds as question-begging. He should also insist that we lack a transparent method that'd support those intuitions, since the facts that matter for him are strictly more complicated than the facts that matter for rule consequentialists. Some Kantians should also dismiss intuitions about active gremlins or Hobbesian worlds as unjustified. Such Kantians allow that the empirical facts that matter for rule consequentialists also matter for discharging our obligations of beneficence. They'd agree with Kant that "to be beneficent, that is, to promote according to one's means the happiness of others in need, without hoping for something in return, is every man's duty" (Kant 1797, 6:453), and ground obligations of beneficence in empirical facts about patterns. For those Kantians, too, we'd lack a transparent method that'd provide genuine evidence from active gremlins or from Hobbesian worlds.

Acknowledgments: For extremely helpful comments and conversations about this material, I'm grateful to Daniel Immerman, Julia Staffel, Brian Talbot, and the referees and editors for the *Australasian Journal of Philosophy.* I'm especially grateful to Joe Horton for a pointed formulation of the 'active gremlins' problems for my (2021).

References

Arneson, Richard. 2005. "Sophisticated Rule Consequentialism: Some Simple Objections." *Philosophical Issues* 15:235–251.

Hare, R. M. 1952. *The Language of Morals.* Oxford: Oxford University Press.

Harsanyi, John. 1982. "Morality and the Theory of Rational Behaviour." In Amartya Sen and Bernard Williams (eds.), *Utilitarianism and Beyond,*, 39–62. Cambridge: Cambridge University Press.

Hooker, Brad. 2000. *Ideal Code, Real World: A Rule-consequentialist Theory of Morality.* Oxford: Oxford University Press.

Johnson, Conrad. 1991. *Moral Legislation.* Cambridge: Cambridge University Press.

Kagan, Shelly. 1998. *Normative Ethics.* Boulder: Westview.

Kant, Immanuel. 1797. *The Metaphysics of Morals.* Cambridge University Press.

Korsgaard, Christine. 1996. *The Sources of Normativity.* Cambridge: Cambridge University Press.

Lenman, James. 2000. "Consequentialism and Cluelessness." *Philosophy and Public Affairs* 29 (4):342–370.

Mill, John Stuart. 1863. "Utilitarianism." In Russ Shafer-Landau (ed.), *Ethical Theory: An Anthology.* Malden, MA: Wiley-Blackwell;.

Miller, Richard. 2009. "Actual Rule Utilitarianism." *Journal of Philosophy* 106 (1):5–28.

Parfit, Derek. 1984. *Reasons and Persons.* Oxford: Clarendon.

—. 2011. *On What Matters*, volume 1. Oxford: Oxford University Press.

Perl, Caleb. 2017. *Positivist Realism.* Ph.D. thesis, University of Southern California.

—. 2019. “Empirical ignorance as defeating moral intuitions? A puzzle for rule consequentialists (and others).” *Analysis* 79 (1):62–72.

—. 2021. “Solving the ideal world problem.” *Ethics* 132 (1):89–126.

Podgorski, Abelard. 2018. “Wouldn’t it be Nice? Moral Rules and Distant Worlds.” *Nous* 52 (2):279–294.

Pryor, James. 2004. “What’s Wrong with Moore’s Argument.” *Philosophical Issues* 14:349–77.

Railton, Peter. 1984. “Alienation, Consequentialism, and the Demands of Morality.” *Philosophy and Public Affairs* 13 (2):134–171.

Rawls, John. 1971. *A Theory of Justice.* Cambridge, MA: Harvard University Press.

—. 1974. “The Independence of Moral Theory.” *Proceedings and Addresses of the American Philosophical Association* 48:5–22.

—. 2001. *Justice as Fairness.* Cambridge, MA: Belknap Press.

Ridge, Michael. 2006. “Introducing Variable-Rate Rule-Utilitariansim.” *Philosophical Quarterly* 56:242–53.

Roberts, Debbie. 2018. “Why believe in normative supervenience?” In Russ Shafer-Landau (ed.), *Oxford Studies in Metaethics 13*, 1–24. Oxford: Oxford University Press.

Rosen, Gideon. 2009. “Might Kantian Contractualism be the Supreme Principle of Morality?” *Ratio* 22:78–97.

—. 2020. “What is Normative Necessity?” In *Metaphysics, Meaning and Modality: Themes from Kit Fine.*, 205–233. Oxford: Oxford University Press.

Scanlon, T M. 1998. *What We Owe to Each Other.* Cambridge, MA: Harvard University Press.

Smith, Holly. 2010. “Measuring the Consequences of Rules.” *Utilitas* 22:413–33.

Thomson, Judith Jarvis. 1996. *Rights, Restitution, and Risk: Essays in Moral Theory.* Cambridge, MA: Harvard University Press.

Woodard, Christopher. 2008. *Reasons, Patterns, and Cooperation.* New York: Routledge.

—. 2019. *Taking Utilitarianism Seriously.* Oxford: Oxford University Press.

Affiliation: Australian Catholic University (Melbourne)