Justifying thresholds

(*Early draft*)

You shouldn't scapegoat someone innocent to prevent five killings. But maybe you should scapegoat someone innocent when the numbers get large enough – if, say, scapegoating would save a million. That is, there might be *thresholds* where the numbers start to matter.

This paper pioneers a new justification for thresholds. The new justification rests on my earlier formulation of rule consequentialism (2021), tidied up significantly. I tidy it up by incorporating some important lessons from Ross, characterizing our reasons first.

1 Introducing thresholds

Judith Jarvis Thomson imagines:

> 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 think that it's wrong for me to scapegoat in this case, even if I can get away with it. We might say that the individual's innocence is an *exclusionary reason* that prevents us from considering the benefits of scapegoating.

Our confidence might waver as the numbers get larger – if scapegoating one would for some reason save a million lives. Maybe a villain will detonate a bomb that kills a million unless I scapegoat. But we might refuse to take the large-number cases to illuminate Thomson's original case. We might say that whatever is correct about each case, they're still different cases. Saying so is positing *thresholds*. We might posit thresholds for promise-keeping, too.

Rule consequentialism can look attractive as a way of putting our considered judgments in reflective equilibrium – including considered judgments about thresholds. For one thing, it promises to do better than act consequentialism in vindicating our considered judgments about ordinary cases of killing, promise-keeping, truth-telling, justice, and so on. This paper will suggest that the right formulation of rule consequentialism provides a deeper justification of our ordinary judgments – both our conviction about scapegoating in Thomson's case, together with the thought that thresholds might start to matter at some point. It should look like an especially promising way of vindicating these judgments, because the thresholds seem to be about consequences.

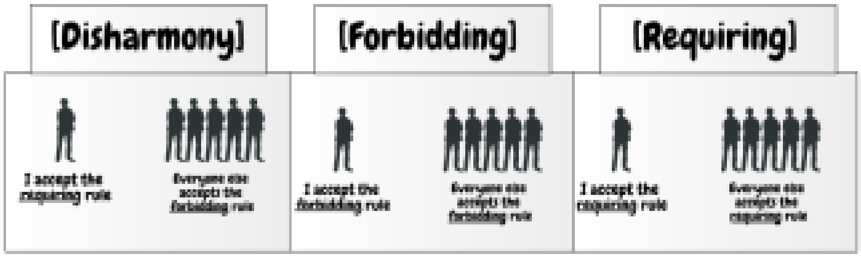
Rule consequentialists start from a much better position than act consequentialists in vindicating our judgment that scapegoating is wrong. The rule consequentialist vindication begins by emphasizing that it's very good for a society to have a stable justice system that people trust. People are then willing to have the justice system adjudicate disputes, instead of avoid ing it because of its capriciousness. And a society where everyone trusts the justice system is much better off than one where people 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.[1] So if the great good of a trusted justice system is relevant for what the sheriff ought to do, benefits distinctive to cooperative patterns must be relevant even when the agent can't herself bring them about.

Rule consequentialist try to capture the moral significance of benefits distinctive to cooperative patterns by contrasting rules about scapegoating:

- The requiring rule: I'm *required* to scapegoat when the odds of discovery are low enough.

  The forbidding rule: I'm *forbidden* to scapegoat no matter the odds of discovery.

Rule consequentialists standardly implement their view by contrasting states of affairs:


[Disharmony]
I accept the requiring rule
Everyone else accepts the forbidding rule
[Forbidding]
I accept the forbidding rule
Everyone else accepts the forbidding rule
[Requiring]
I accept the requiring rule
Everyone else accepts the requiring rule

Let's suppose that:

> [Disharmony] is better than [Forbidding], which is better than [Requiring]

This supposition should look plausible. It's a schematic representation of the facts emphasized earlier: that a trusted justice system is a great good, but not one that any one person can cause. Supposing that [Forbidding] is better than [Requiring] represents the fact that a trusted justice system *is* a great good – that good makes [Forbidding] better than [Requiring]. (I assume that enough people would be scapegoating in [Requiring] for trust in the justice system to collapse.) Supposing that [Disharmony] would be even better than [Forbidding] represents two important facts: first, that no

[1] The pattern may cause beliefs about the trustworthiness of the justice system, and those beliefs may be what secures the good. But beliefs don't always mediate the good effects of patterns. So I won't emphasize the mediating role of

beliefs in what follows, since I'm interested in the general significance of good patterns.

one person can destroy by himself destroy trust in the justice system, and second, that my scapegoating would create the good of five people living rather than just one person living.

The good of a trusted justice system illustrates a benefit distinctive to cooperative patterns: a good that's available only if agents coordinate. Since act consequentialists take this kind of benefit to be morally irrelevant, they'll predict that I'm sometimes required to scapegoat. There are only two states of affairs that I can bring about: [Disharmony] and [Forbidding].[2] Of those two, [Disharmony] is best – and I scapegoat in [Disharmony]. Since act consequentialism tells me to bring about the best state of affairs I can, it'll predict that scapegoating is sometimes required. But predicting that scapegoating is sometimes required is out of sync with our considered moral judgments.[3]

Similar points apply even to the most sophisticated kinds of act conse quentialism. Some act consequentialists deny that scapegoating is morally required even if it has the highest expected utility, because they do not focus on the expected utility of acts considered on their own. They instead focus on decision procedures that maximize agents' chances of bringing about what's best. They may insist that the decision procedure that maximizes my chances of doing so forbids acts of scapegoating. After all, scapegoating often has bad consequences, and we're not in a position to know when it does. So you can maximize your chances of doing what's best by refraining from scapegoating. Such act consequentialists are then in a better position to vindicate our considered moral convictions that scapegoating is wrong.[4]

However, even sophisticated act consequentialists still face a problem. Many of us think that scapegoating is *never* morally required. And even though sophisticated act consequentialists can predict that scapegoating *usually* isn't required, it's very hard for them to predict that it's *never* required. Contrast a decision procedure that always forbids scapegoating with a decision procedure that forbids scapegoating unless the stakes are high enough and the odds of discovery low enough. The latter decision pro cedure could be best. Scapegoating *sometimes* has the best consequences; we should expect the decision procedure that maximizes the chances of do ing what's best to require scapegoating when it's likely enough that you can get away with it.

In contrast, pattern-dependent views like rule consequentialism take the great good of a stable justice system into account. For example, a universal acceptance rule consequentialist can predict that scapegoating is always forbidden, because [Disharmony] is simply *irrelevant* for her. Her view is sensitive only to states of affairs where everyone accepts the same rules. Since [Disharmony] involves agents accepting different rules, we put

[2] For an act consequentialist, we would focus on slightly different states of affairs: ones where I scapegoat or don't. I'm suppressing this complication in the

it aside. Between [Forbidding] and [Requiring], [Forbidding] is the only one that contains the great good of a stable justice system, so it is best and determines our obligations: scapegoating is always forbidden.

I've been using the term 'pattern-dependence'; to regiment the use, I'll say that a view that appeals to features of cooperative patterns is (group)- pattern-dependent:

> (Group)-Pattern-Dependence: What's forbidden, required, or permitted is explained by features of a group's pattern of behavior, and not explainable by features of any individual's pattern of action considered on its own.

Act consequentialism isn't (group)-pattern-dependent, because it focuses on individual acts considered on their own. I'll use 'pattern-dependent' to abbreviate '*group*-pattern-dependent'.[5]

2 Introducing the Reflexive View

2.1 Morality As Reaction

I now introduce two abstract pictures of morality. The first picture emphasizes what our actions *produce*. This picture makes act consequentialism compelling. Christine Korsgaard helpfully emphasizes how consequentialists see "action [as] essentially production, and therefore its function is to bring something about, to achieve some end" (Korsgaard 2008, 216).

> Morality As Production: moral action produces some good.

Thus Mill, for example: "all action is for the sake of some end, and rules of action, it seems natural to suppose, must take their whole character and color from the end to which they are subservient" (Mill 1863, 2). I'll use Morality As Production as a foil. I want to describe a rival.

My rival picture links moral action with our reactive attitudes – resentment, indignation, guilt, and the like. Acting morally can make those attitudes *in*appropriate. For instance, it'd be inappropriate to resent a judge for holding fair trials as long as we thought that the judge did what she ought to. My rival picture takes the appropriateness conditions for reactive attitudes to reveal *objective* moral norms.

> Morality As Reaction: objective moral norms govern the reactive attitudes.

Morality As Reaction picks up a different strand in Mill: his insistence that "we do not call anything wrong, unless we mean to imply that a person

main text, to facilitate the comparison.

[3] The same problem arises for Regan's cooperative consequentialism, because his final theory has the property that he calls 'PropAU'.

[4] Moore (1903), Smart (1973), Norcross (1990), Mason (2004), and Railton (1984) have all sketched or defended suggestions with this kind of structure.

ought to be punished in some way or other for doing it– if not by law, by the opinion of his fellow-creatures; if not by opinion, by the reproaches of his own conscience" (Mill 1863, 48-9). Now Mill himself took Morality As Production to reveal the nature of objective moral reality. He relegates these sanctions to just revealing subjective moral norms.

In contrast, Morality As Reaction presents itself as revealing the nature of objective moral norms. Objective moral norms directly govern how to hold people accountable, governing the reactive attitudes. Now we can often produce more good by *not* holding people accountable just for what they produce. For instance, we can often produce more good by holding people accountable for following fair rules, even when following fair rules sometimes prevents them from producing the best option. So if Morality as Production is right, objective moral norms *wouldn't* directly govern how to hold people accountable. So even though Morality As Reaction picks up some threads in Mill, it weaves those threads into a tapestry that Mill himself rejects.

For instance, it'd be inappropriate to resent judges who hold fair trials. They're together sustaining valuable trust in the justice system. Morality As Reaction then predicts that the judges act as they *objectively should*. Morality As Production disagrees. Fair trials aren't always the way of producing the most good. Judges could sometimes scapegoat without affecting trust in the justice system. For Morality As Production, the objective requirements needn't directly line up with the standards governing resentment. That's just what the previous section observed, in noting that we might want a reason for an alternative to act consequentialism.

Morality As Reaction specifies a more general idea of moral action as *publicly justifiable*. Someone in the grip of Morality As Production could agree that it'd be inappropriate to resent the judges who refuse to scapegoat. But they wouldn't take that agreement to constrain their understanding of *objective* moral requirements. They'd instead take the agreement to constrain what's *subjectively* required. I'll just assume that objective moral requirements are public. This paper explores whether this sort assumption can justify rule consequentialism. I'm not innovating in grounding rule consequentialism on this sort of publicity requirement. As Darwall notes, Parfit's shift from the global consequentialism of *Reasons and Persons* to the Triple Theory of *On What Matters* rests on Parfit's coming to accept "a view of the deontic concept of moral rightness that ties it closely to blame worthiness and accountability in a way that effectively concedes a Rawlsian publicity condition" (Darwall 2014, 79).

I'm introducing Morality As Reaction as a principled foundation for rule consequentialism. But it's embarrassingly difficult for rule consequentialists to make good on this kind of picture. For instance, one

[5] The abbreviation may be misleading; some like Fred Feldman (1975) defend alternatives that center *individual* patterns of action rather than group patterns. But I hope that the context will make the intended meaning of 'pattern dependent' clear enough.

natural kind of rule consequentialism supposes that we should obey the rules that'd make things go best if universally accepted. Pacifist rules plausibly would make things go best if everyone accepted them. But that plausible point *doesn't* mean we should resent those who defend themselves! We don't live in a world where everyone is a pacifist.

More generally, the reactive attitudes ordinarily react to what *actually*

happens. They don't react to what could possibly happen. Morality As Reaction then doesn't look like an apt foundation for any kind of rule con sequentialism that centers center *counterfactuals* about what could happen. It could only be an apt foundation for reacting to what *actually* happens.

Worse, extant forms of rule consequentialism tend to center counter factuals, rather than what actually happens. They often compare rules by imaging what would happen if different *embeddings* took hold – if, say, the rules were followed, or accepted, or internalized, by some or most or all of the community.[6] Comparing those embeddings means comparing at least some counterfactuals, about what merely could happen. (For instance, uni versal pacifism may never take hold.) I interpret formulations of the 'ideal world' problem as showing that extant formulations of rule consequential ism do rest on counterfactuals, and go wrong for that reason (Rosen 2009; Podgorski 2018).

2.2 The Reflexive View

My (2021) introduced one specification of Morality As Reaction. That spec ification focuses exclusively on what *actually* happens. It thus promises to fit Morality As Reaction, since it eliminates the counterfactuals that previ ous accounts have used. I'll call my proposal the *Reflexive View*, because it rests on reflexive evaluation of what actually happens. This section intro duces my initial version of the proposal; the next section identifies a central problem and gives a different solution.

My new view begins with the platitude that the demands of morality are the demands of the rules that are morally best. The view will hold, roughly, that the set of morally best rules is the set of sufficiently general rules that has optimal moral value.[7] My development of the platitude incorporates two requirements. The first is the *Generality* Requirement:

> Generality Requirement: the morally best rules
> are sufficiently general.

Contrast two formulations of rules about scapegoating:

- *Irreducibly Singular* : If my scapegoating is beneficial enough and unlikely to be discovered, then my scapegoating is morally right.
- *General*: If anyone's scapegoating is beneficial enough and unlikely

> to be discovered, then that person’s scapegoating is morally right.

The Generality Requirement eliminates the Irreducibly Singular rule. This Requirement starts from our conviction that “because it’s me!” can’t be a brute justification for a moral claim. If it ever looks like a justification, it’s functioning as an elliptical way of referring to general facts about me.

[6] Richard Brandt (1963), John Harsanyi (1977), Brad Hooker (2000), Derek Parfit (2011), Holly Smith (2010), and Michael Ridge (2006) describe a representative range of the options.

[7] I’m grateful to an associate editor for this helpful gloss.

A wide range of philosophers accept some version of the Generality Requirement; Rawls, for example, requires that “principles should be general. That is, it must be possible to formulate them without the use of what would be intuitively recognized as proper names, or rigged definite descrip tions” (Rawls 1971, 113). I take Rawls to identify paradigmatic *violations* of the Generality Requirement. But I think he should leave open what else violates this Requirement; for instance, R. M. Hare thinks this sort of requirement should also exclude indexicals (Hare 1952, 129). Whatever Rawls himself thought, I intend to leave the exact contours of the General ity Requirement open for now. Rawls’ endorsement at least illustrates the range of philosophers who can endorse a version of this Requirement.

The second requirement gives my new account its consequentialist flair. In formulating it, I’ll assume that some states of affairs are good in an agent neutral way. For example, a state of affairs where I scapegoat an innocent person might be agent-neutrally good, where the benefits of scapegoating outweigh the costs. Act consequentialists can identify moral goodness with agent-neutral goodness: the morally best act produces the greatest amount of agent-neutral good.

The innovation in my new second requirement is to give a different account of the moral significance of what’s agent-neutrally good. Roughly, a rule is credited with agent-neutral good/ bad only when the rule “approves” of its production. More carefully:

> Consilience Requirement: the moral value of a rule *R* is everything actual that’s agent-neutrally good or bad to the extent it’s caused by actions that *R* classifies as morally right.

Go back to the earlier example about scapegoating. In the example, my *individual* scapegoating produces something good, but a *pattern* of scapegoating destroys something good – it destroys trust in the justice system.

- The scapegoating rule: scapegoating is right when the odds of discovery are low enough, and fair trials are wrong.
- The fair trials rule: scapegoating is wrong even when the odds of discovery are low enough, and fair trials are always right.

Given the Consilience Requirement, the great good of a trusted justice system is credited to the fair trials rule and is not credited to the scapegoating rule. Fair trials are what sustains trust in the justice system. But according to the scapegoating rule, fair trials are wrong when the person could be scapegoated. So the actions that cause trust in the justice system are wrong according to the scapegoating rule – which means that the great good of a trusted justice system isn't credited to the scapegoating rule. In contrast, the great good of a trusted justice system *is* credited to the fair trials rule. The actions that cause that great good are *right* according to the fair trials rule. Now there is also a genuine good that's credited to the *scapegoating* rule but not to the fair trials rule: the agent-neutral good of successful scapegoating. In the Thomson-style case mentioned earlier, scapegoating

would save five people's lives at the cost of only one life. So the agent neutral good of four more people surviving is credited to the scapegoating rule and not to the fair trials rule. To compare the scapegoating and fair trial rules, we add up the agent-neutral good net bad credited to each.

Our driving observation is that the good of a trusted justice system is greater than all the goods that come from successful scapegoating.[8] We're looking for an account of the moral significance of our driving observation free of counterfactuals. That's *exactly* what the Consilience Requirement provides: it predicts that the fair trials rule is best given that driving observation. The Consilience Requirement thus predicts that fair trials are right and scapegoating is wrong, because that's what the best rules say.

I call this view the *Reflexive* View because it centers reflexive evaluation – evaluation of what actually happens according to the very rules being evaluated. The next section makes some dramatic changes to this formulation. But it retains the central move: reflexively evaluate rules.

This paper focuses directly on the basic idea of reflexive evaluation. It aims to see if any version of it at all can be made to work. The easiest way to do that is to screen off separate problems by stipulating that they have answers. I make two assumptions to do that. I first assume that we're working from the stock of the act-descriptions that the Generality Requirement would vindicate. I'll talk about holding fair trials, scapegoating, preventing disaster, and maximizing welfare. I'll thus be treating those act-descriptions as *eligible* candidates in moral reasoning. Appeal to act-descriptions like "preventing disaster", "scapegoating", and the like is an idealization; the Generality Requirement may only vindicate cousins or distant relatives. I ignore those complications from here on.

I'll also ignore objections that rest on changes to the actual empirical facts. I've argued elsewhere in detail that those sorts of objections are question-begging (Perl forthcoming). Relatedly, I'm going to assume that the actual empirical facts pattern in the way that's the most helpful for the rule consequentialists. I'm exploring a more abstract question: whether there's a way that makes good on the guiding idea of Morality As Reaction.

3 Redoing the Reflexive View

My early formulation of the Reflexive View simply doesn't work. It doesn't vindicate the simple rules that I said it does: for instance, a rule scapegoat ing is always right. It rather ends up vindicating rules where scapegoating is one of several right actions. It vindicates what I'll call a *Scheffler* rule:

- *Scheffler* rule: a rule where fair trials are right but *maximizing wel fare* is also right.

The Consilience Requirement credits the Scheffler rule with the great good of a stable justice system. But it also credits it with the good of successful

[8] That observation is necessary for rule consequentialism to be a genuine rival to act consequentialism. And as §1 emphasized, we're here assuming that it is a genuine rival in the course of trying to find the best formulation.

scapegoating. So it credits the Scheffler rule with strictly more goods than the goods credited to the simpler rule where fair trials are always right. My initial formulation thus fails to vindicate its central ambition: to explain our considered judgment that scapegoating is wrong.

I diagnose my Consilience Requirement as going wrong because it tries to characterize what's *right*. It should instead characterize the *reasons* that determine what's right and wrong. Step back for a moment and note that the Consilience Requirement is picking out features that seem to have some kind of moral significance: fair trials and maximizing welfare. But my formulation of the Consilience Requirement adds a further interpretation: that both are themselves *right*, all things considered. And it's that further interpretation that creates the problems.

I'll propose reworking the idea behind the Consilience Requirement as characterizing our *reasons*, rather than characterizing what's right. This reinterpretation thus needs a further characterization of how our reasons determine what's right and wrong. So I'll reformulate in two steps: rework the Consilience Requirement as an account of our reasons, then *add* a new characterization of how our reasons weigh together. This reformulation is my tenuous connection with Ross and the conference – he taught us the importance of characterizing reasons first.

Reworking the Consilience Requirement to be an account of reasons requires a more complicated picture of the structure of the *rules*. First, those rules must specify reasons:

- R_{fair} = the fact that φ-ing is a fair trial is a reason for x to φ.
- $R_{maximizing}$ = the fact that φ-ing maximizes welfare is a reason for x to φ.

- ...

Other sets of rules could model act consequentialism – for instance, the set containing only $R_{maximizing}$ could model a kind of act consequentialism: I propose replacing the Consilience Requirement with an account of our *base reasons*.

> Base Reasons: There is a reason to do whatever type of intentional action does the most to jointly explain each actual net good.

Consider the good of a trusted justice system. Holding fair trials does the most to explain that trust. Base Reasons thus predicts that there is reason to hold fair trials. It thus vindicates the R_{fair} reason: the fact that φ-ing is a fair trial is a reason for x to φ. Base Reasons also vindicates a reason to maximize welfare. Intending to maximize welfare does the most to explain the good of giving some particular chunk of money to GiveDirectly.

Crucially, though, Base Reasons does *not* generate a separate reason to scapegoat. Acting to scapegoat *doesn't* help explain trust in the justice system. It instead exploits the trust that already exists.

Base Reason fits my guiding idea, of Morality As Reaction: we have to see the people who sustain trust in the justice system as acting rightly. Seeing them as acting rightly means seeing them as acting for sufficient reason. Base Reasons generates reasons for their acting like they do. It doesn't yet vindicate them as acting rightly, because it doesn't yet offer an account of the priorities between reasons. But Base Reasons does *preserve* the features that made the Consilience Requirement work: reflexive eval uation of actual goods. It uses those features to characterize our reasons, rather than to characterize what's right directly.

I then need an additional account of *priorities* between our reasons. Fair trials plausibly *exclude* the benefits of scapegoating – roughly, just judges ignore those benefits when ruling. Capturing these observations means capturing when one reason excludes another. I modify the Reflexive View to include:

> Exclusion: R_x excludes R_y iff the actual value due to R_x-actions is greater than the actual value due to R_y-actions *in the common domain when someone is choosing between an x-action and a y-action*.

Judges can choose between holding a fair trial and maximizing welfare. They needn't think explicitly about the choice. But their 'option-set' includes both. (I'm belaboring this point now so I can *later* distinguish ordinary judges from judges whose option set also includes include a terrorist who'll kill a million if a fair trial is held. That distinction matters for my official justification of thresholds – but not yet.)

Exclusion then predicts that R_{fair} excludes $R_{maximizing}$. • R_{fair} = the

fact that φ-ing is a fair trial is a reason for x to φ.

- $R_{maximizing}$ = the fact that φ-ing maximizes welfare is a reason for x to φ.

The value of trusted justice systems swamp the value of successful scape goating. As a result, the actual value of holding fair trials is greater than the actual value of maximizing welfare *in the common domain of holding trials*. Exclusion then applies: R_{fair} excludes $R_{maximizing}$. This paper assumes that ideal moral reasoning attends to all and only the genuine, un excluded reasons. Given that assumption, Exclusion predicts that judges should simply ignore the benefits of scapegoating in their deliberation – which seems right.

In fact, we'd *resent* judges who weighed the benefits of scapegoating in their deliberation. We'd take their weighing to reflect a bad quality of will to those who are innocent. Morality As Reaction takes our appropriate resentment to reveal the judges' objective reasons. The pivotal Exclusion claim vindicates that resentment as appropriate: we'd resent such judges for reasoning from an excluded reason. The method of reflective equilib rium thus justifies Exclusion, as an abstraction that unifies more particular considered judgments.

The Reflexive View then incorporates a standard characterization of how our reasons determine what we may and must do.

- I may φ iff φ-ing is one action I have reason to do
- I must φ iff φ-ing is the *only* action I have (most) reason to do

Exclusion predicts that the reason to hold fair trials exclude the reason to maximize welfare. So it predicts that holding a fair trial is the *only* action I have reason to do. Given that prediction, it holds that I must hold fair trials. It thus delivers precisely the verdicts about scapegoating that we'd want a pattern-dependent view to deliver.

The Reflexive View differs from act consequentialism by incorporating Base Reason and Exclusion. The standard characterization of what we must do in terms of our reasons is formally compatible with act consequentialism. Consider again the set of rules R_{act} that acknowledges exactly one reason:

- $R_{maximizing}$ = the fact that φ-ing maximizes welfare is a reason for x to φ.

Maximizing welfare is then the only action I have reason to do. It's obligatory as well as permissible. That is, my *formal* characterization of obligation and permission doesn't eliminate act consequentialism; it's the dis tinctive elements of the Reflexive View that do.

Base Reason plus Exclusion together vindicate the following set over any act consequentialist set of rules:

- R_{fair} = the fact that φ-ing is a fair trial is a reason for x to φ.
- $R_{maximizing}$ = the fact that φ-ing maximizes welfare is a reason for x to φ.
- R_{fair} excludes $R_{maximizing}$
- ...

Base Reason generates R_{fair} – so it generates a more expansive set of (objective!) reasons than act consequentialists acknowledge. And Exclusion predicts that R_{fair} excludes $R_{maximizing}$. Exclusion still predicts that $R_{maximizing}$ is a genuine reason when asking whether to GiveDirectly. Fair trials aren't in play, so that reason isn't excluded.

4 Thresholds

A terrorist threatens to detonate a nuke that'd kill a million unless a judge scapegoats someone innocent. That'd be a *disaster* ! Happily, the Reflexive View also predicts that that judge must scapegoat when disaster looms.

The crucial point is that the judge faced with disaster doesn't face the same choices as ordinary judges. Ordinary judges don't confront terrorists with nukes. So our reactions to ordinary judges needn't determine our reactions to judges confronted with nukes.

4.1 Exclusion vindicates thresholds

Vindicating this crucial point requires vindicating a reason for judges to avoid disaster. That is, our base set of reasons needs to be:

- R_{fair} = the fact that φ-ing is a fair trial is a reason for x to φ.
- $R_{maximizing}$ = the fact that φ-ing maximizes welfare is a reason for x to φ.
- $R_{disaster}$ = the fact that φ-ing prevents disaster is a reason for x to φ.

If the Reflexive View generates this base set of reasons, Exclusion would then predict that $R_{disaster}$ excludes R_{fair}.

> Exclusion: R_x excludes R_y iff the actual value due to R_x actions is greater than the actual value due to R_y-actions *in the common domain when someone is choosing between an x action and a y-action*.

Exclusion restricts attention to cases with nukes. Given this restriction, the Reflexive View roughly mimics act consequentialism. It compares the (epistemic) possibility where the judge holds fair trials and the nuke goes off with the (epistemic) possibility where the judge scapegoats. If the judge holds fair trials, the Reflexive View *discounts* the good of a trusted justice system by the extent to which that good is due to the judge's holding fair trials. That discount is massive: this *particular* judge's fair trials are just a drop in the pond. Of course, the judge's fair trials also causally contribute to the resultant explosion. The Reflexive View also discounts the bad of that explosion by the judge's causal contribution. But that discount is much less drastic. For one thing, the judge's holding fair trials is necessary for the explosion. Maybe the View credits the judge's fair trials with about half the bad of the explosion. In contrast, the View doesn't credit scapegoating with any good at all: it's not part of holding fair trials.

R_{fair}-actions are massively bad: maybe the bad of half a million deaths plus the good of being a drop in the bucket for fair trials. In contrast, $R_{disaster}$-actions are mainly neutral. Given the comparisons that Exclusion makes, the actual value of $R_{disaster}$-actions is higher than the actual value of R_{fair}-actions in the common domain. So Exclusion predicts that $R_{disaster}$ excludes R_{fair}. That's how the Reflexive View justifies a threshold for holding fair trials.

Exclusion justifies thresholds by modulating the domains of comparison. In ordinary cases – without disaster – we compare all the ordinary cases of fair trials with the ordinary cases of scapegoating. There are lots of those cases. So even though we discount the effects of each trial by its causal contribution, there are *so many* fair trials that they get credited with most of the good of the trusted justice system. We compare a different domain when disaster looms: we compare just the domain of actions where disaster looms.

4.2 No collapse

Modulating the domain of comparison threatens a familiar kind of collapse. After all, Thomson's case is itself exceptional: ordinary trials don't involve a lynching outside. Exclusion then threatens to make our obligations in Thomson-style cases depend on just the domain of actions where a lynching looms. When we restrict attention just to *that* domain, judges should scapegoat for precisely the same reason they should scapegoat in disaster type cases. The Reflexive View would then credit fair trials with the bad of five deaths plus the tiny good of trusted justice systems, discounted by the causal contribution of fair trials in the face of lynchings. And it'd credit scapegoating with no good. So it'd predict that *undetected scapegoating* is also a reason that excludes fair trials. And it'd make that bad prediction for *precisely* the reason that it makes the good prediction that disaster is a reason that excludes fair trials. No surprise here – it's really easy for rule consequentialist views to accidentally collapse.

In introducing this problem about collapse, I implicitly appealed to an important assumption. I just implicitly assumed that undetected scape goating provides a *separate, self-standing* reason – that is, that the set of basic reasons includes both $R_{scapegoating}$ and R_{fair}.

- R_{fair} = the fact that φ-ing is a fair trial is a reason for x to φ.
- $R_{scapegoating}$ = the fact that φ-ing would be undetected scapegoating is a reason for x to φ.

The base set of reasons itself *modulates* the domain of comparison. We'd modulate the domain to only include cases like Thomson's when we're com paring these two reasons. If $R_{scapegoating}$ isn't part of the base set of rea sons, no collapse would loom. Exclusion only applies when $R_{scapegoating}$ is part of the base set of reasons.

So here's the general strategy for avoiding collapse: don't generate reasons to scapegoat in the first place! In particular, I suggest that trusted systems of justice depend causally on judges intending to not scapegoat in cases like Thomson's. Those intentions causally explain the trust in systems of justice. Disasters are different. Trusted systems of justice don't depend causally on what judges intend to do in case of disaster. Maybe the judges actually intend to give up on fair trials when disaster looms. The presence of that intention doesn't affect trust in systems of justice. That is, the presence or absence of intentions about disasters don't causally explain trust in the justice system. The Reflexive View then uses this causal difference to justify thresholds.

Given this causal difference, the Base Reasons claim predicts that $R_{scapegoating}$ is *not* one of the base reasons.

> Base Reasons: There is a reason to do whatever type of intentional action does the most to jointly explain each actual net good.

Instead, Base Reasons vindicates the following set of base

reasons.

- R_{fair} = the fact that φ-ing is a fair trial is a reason for x to φ.
- $R_{maximizing}$ = the fact that φ-ing maximizes welfare is a reason for x to φ.
- $R_{disaster}$ = the fact that φ-ing prevents disaster is a reason for x to φ.

Base Reasons doesn't vindicate a separate reason to scapegoat because it does not recognize any actual net good due to scapegoating. Successful scapegoating *does* have good effects. But those good effects are part of the good effects of *trusted systems of justice*. Remember that the overarching ambition of this paper is to vindicate *pattern-dependent* views, where moral obligations depend on facts about patterns of action. And scapegoating causally depends on trust in systems of justice; it wouldn't work without it. There is *no* actual net good *distinctive* to successful scapegoating; it's rather distinctively part of trusted systems of justice more generally.

In contrast, I'll suggest that there is a distinctive good associated with disaster avoidance, distinct from the good of a trusted system of justice. Given that suggestion, Base Reasons would vindicate a reason to prevent disaster. Intending to prevent disaster would do the most to explain that distinctive good. Establishing this suggestion means establishing that the good of disaster avoidance does *not* causally depend on the good of a trusted system of justice. Establishing that point would distinguish dis aster avoidance from successful scapegoating, since successful scapegoating *does* causally depend on a trusted system of justice.

I'll start with an implausible simplifying assumption. Then I'll relax the simplifying assumption. The implausible simplifying assumption is that we all know that all judges intend to hold fair trials absent disaster. (This assumption is implausible in part because judges ordinarily *don't* think about disasters when holding fair trials.) People will still trust the justice system. For one thing, nuclear-armed terrorists are rare enough that even iterated disasters won't destabilize and collapse trust. For another, it's more transparent to judges when disaster looms than when they can get away with scapegoating. I think that any rule consequentialist is going to have to emphasize these sorts of empirical points; I don't have anything new to say about them. What I'm providing is an adequate systematic framework for integrating these points if they're correct.

Given my implausible simplifying assumption, then, the good of disas ter avoidance doesn't causally depend on trusted systems of justice. We can all know that judges intend to avoid disaster while still trusting the justice system. That's why I distinguish two different goods. Given those two goods, Base Reasons generates two reasons to act. Since I hold that the goods of successful scapegoating are *not* distinct from the good of trusted systems of justice, Base Reasons will not generate a third reason to scape goat. That's what blocks collapse.

Base Reasons thus generates reasons to avoid disaster given the further claim that the good of disaster avoidance doesn't causally depend on trusted systems of justice. That further claim should look independently plausible,

even apart from my implausible simplifying assumption. Maybe you're skeptical. What's the evidence for that further claim? Well, one important piece of evidence comes from *counterfactual* comparisons. Imagine that every judge intended to hold fair trials unless they could get away with scapegoating. Then we wouldn't trust the justice system – it wouldn't work. For one thing, finite creatures like us are bad at telling when we can get away with scapegoating. So scapegoating will be discovered more than we expect. In addition, iterating something unlikely enough times and something will be discovered. In contrast, we'd still trust the justice system even if judges intended to hold fair trials unless disaster looms. That's just what we saw with my initial discussion of our reasons to hold fair trials.

I've emphasized that I want to ground moral obligations fully in what actually happens. But I've just started talking about counterfactuals in the way that other rule consequentialists traditionally have. Crucially, though, I'm using those counterfactuals as *heuristic evidence*, rather than the genuine grounds themselves. The pivotal grounds are about causal dependence: I'm claiming that the good of successful scapegoating causally depends on trusted systems of justice, but that the good of disaster avoidance doesn't. I'm appealing to counterfactuals as heuristic evidence for these claims about causal dependence. But I'm not resting the justification on the counterfac tuals themselves.[9]I'm rather resting it on causal dependencies – which we can often see by considering counterfactuals, but which do not *depend* on counterfactuals.

My Base Reasons claim then captures my guiding idea, Morality As Reaction. The guiding idea is that we have to see those who jointly produce the trusted justice system as acting as they're required to. The trusted justice system causally depends on their refusing to scapegoat. Admitting $R_{scapegoating}$ as a genuine reason prevents us from seeing them as acting as they're required to. In contrast, the trusted justice system does *not* depend causally on their holding fair trials in the face of disaster. So our reaction to ordinary trials or cases like Thomson's needn't dictate how we react in the face of disaster. We can then recognize a reason to avoid disaster without imperiling our ordinary obligations to hold fair trials.

References

Berker, Selim. 2019. "The Explanatory Ambitions of Moral Principles." *Nous* 53 (4):904–936.

Brandt, Richard. 1963. "Toward a Credible Form of Utilitarianism." In H.-N. Casta˜neda and G. Nakhnikian (eds.), *Morality and the Language of Conduct*, 107–43. Detroit: Wayne State University.

Darwall, Stephen. 2014. "Agreement Matters: Critical Notice of Derek Parfit, *On What Matters*." *The Philosophical Review* 123 (1):79–105.

Feldman, Fred. 1975. "World Utilitarianism." In Keith Lehrer (ed.), *Analysis and Metaphysics*, 255–273. D. Reidel Publishing Company,.

[9] In any case, contemporary counterfactual accounts like Judea Pearl (2000) and James Woodward (2003) don't take counterfactuals to figure in fundamental explanatory grounds. But as Selim Berker (2019) emphasizes, normative theorizing aims at identifying the fundamental explanatory grounds. So incorporating a Pearl/Woodward-style view in the Patterned View still wouldn't mean that counterfactuals play a fundamental role.

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

Harsanyi, John. 1977. “Rule Utilitarianism and Decision Theory.” *Erkenntnis* 11:25–53.

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

Korsgaard, Christine. 2008. *The Constitution of Agency*. Oxford University Press: Oxford.

Mason, Elinor. 2004. “Consequentialism and the Principle of Indifference.” *Utilitas* 16 (3):316–321.

Mill, John Stuart. 1863. *Utilitarianism*. Indianapolis, IN: Hackett Publishing Company.

Moore, G E. 1903. *Principia Ethica*. Cambridge: Cambridge University Press.

Norcross, Alastair. 1990. “Consequentialism and the Unforeseeable Future.” *Analysis* 50 (4):253–256.

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

Pearl, Judea. 2000. *Causality*. Cambridge: Cambridge University Press.

Perl, Caleb. forthcoming “Some question-begging objections to rule conse quentialism.” *Australasian Journal of Philosophy* .

—. 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.

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.

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

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

Smart, J J C. 1973. *Utilitarianism: For and Against*. New York: Cambridge

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.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.