Criminal Law and Philosophy
https://doi.org/10.1007/s11572-023-09665-9

ORIGINAL PAPER

Too Objective for Culpability?

Alex Sarch[1]

Accepted: 22 February 2023

Abstract

To help explain in a principled way why criminal law doctrine tends to abstract away from motives and other individualized circumstances, I have defended an insufficient regard theory of criminal culpability that is more objective in certain respects than other views in the same camp. This has led Alec Walen to object that my view is too objective to be an account of culpability and is better understood as a theory of criminal wrongs. This challenge is important not least because it requires getting clearer about what wrongness and culpability are and what roles they play on a legal moralist picture of the criminal law. Here, I argue that Walen's objection is mistaken. Once we get clearer on what distinguishes wrongness and culpability, it becomes clear that my account is best seen as a theory of culpability. This is so even though it calculates degrees of culpability in a more objective way than other insufficient regard views. Besides just defending my account from Walen's objection, this paper aims to make a positive contribution by developing a more sophisticated version of the Manifestation Account I proposed earlier. To do this, I focus on the main area of criminal law doctrine where culpability comes apart from wrongness – namely, excuse cases. My earlier Manifestation Account is implausible – or at least limited – because it has nothing distinctive to say about how to determine the culpability of excused misconduct. Accordingly, the theory would be, if not outright false, at best incompatible with the orthodox view in criminal law theory, which takes excuses to accept wrongdoing but deny culpability. To solve this problem and enhance the Manifestation Account's explanatory value, I show how to extend the theory to account for the culpability of wrongs where a putatively excusing condition is in play. This not only fills a gap in the Manifestation Account but has the further benefit of providing a unified reasons-based account of the main categories of misconduct in the criminal law, which shows what separates criminal wrongness from culpability both in justification cases and excuse cases. The hope is that this effort will shed light on the distinct roles of wrongness and culpability in legal moralist theories of the criminal law and provide a better understanding of degrees of culpability for criminal wrongs – not only when justifications are at issue, but also when excuses are involved.

Extended author information available on the last page of the article

Published online: 21 March 2023

Keywords Culpability · Wrongness · Legal moralism · Justifications · Excuses · Manifestation · Objectivity · Duress

How objective is the notion of culpability? Suppose you think, as is common, that culpability is a matter of how poorly the actor engaged with the applicable reasons bearing on what to do.[1] Does this mean that the criminal law's liability rules, if they are to reflect culpability in the way legal moralists think would be ideal,[2] should (or at least could) be concerned with all the specific motivations and individualized circumstances that led offenders to break the law? After all, such factors may have been a part of their actual engagement with reasons. However, this would sit uncomfortably with the common observation that criminal law doctrine (unlike sentencing) should abstract away from many sorts of individual circumstances and remain largely indifferent to one's motivations for performing a criminal action (at least as the default position).[3]

To capture the latter sort of view in a principled way (rather than merely seeing it as a result of practical limitations or compromises), in earlier work I defended an insufficient regard theory of criminal culpability that is more objective in certain respects than other views in the same tradition.[4] Insufficient regard views generally understand culpability in terms of one's lack of due regard for, *i.e*. improper

[1] James Edwards & Andrew Simester, *Crime, Blameworthiness, and Outcomes*, 39 Oxford J. Legal. Stud. 50, 55 (2019) ("What makes D's φ'ing blameworthy…is D's engagement with those guiding reasons," such that "D is blameworthy for φ-ing in virtue of the fact that D's engagement with those reasons exhibits some shortfall or deficiency, relative to what we can reasonably expect from D"); *see also* Gideon Yaffe, Attempts 38 (2011) (endorsing the view that an action is culpable to the degree that "it is a product of a faulty mode of recognition or response to reasons for action"); Andrew Simester, Fundamentals of Criminal Law 16-17 (2021).

[2] *See*, *e.g*., Doug Husak, Ignorance of Law 34 (2016) (defending a version of legal moralism on which "we should recognize a presumption that the criminal law should…be based on, conform to, or mirror critical morality"); RA Duff, *Towards a Modest Legal Moralism*, 8 Crim. L. & Philosophy 217, 229–30 (2014) (distinguishing negative and positive legal moralism as a constraint on vs a reason in favor of criminalization; sketching a sophisticated legal moralist view in which public moral wrongs are a key ingredient; arguing the condemnation and censure of the criminal law "must, as a matter of justice, be directed at wrongful conduct and its culpable agent").

[3] *See*, *e.g*., Wayne LaFave, 1 Substantive Criminal Law § 5.3 (2d ed.) (2003) ("A defendant's motive, if narrowly defined to exclude recognized defenses and the 'specific intent' requirements of some crimes, is not relevant on the substantive side of the criminal law."); Kenneth Simons, *Does Punishment for "Culpable Indifference" Simply Punish for "Bad Character"?*, 6 Buff. Crim. L. Rev. 219, 234 (2002) ("the harsh sanctions of the criminal law should not be brought to bear on individuals who have not yet done anything wrong, but who merely have disreputable—or even dangerous—character traits"). Of course, some crimes take particular motives (as opposed to the traditional mens rea categories) to be elements of the offense. *See e.g*. Model Penal Code § 212.1 (specifying the requirement for a specified bad purpose in order to be guilty of kidnapping). However, I take the inclusion of motives as elements to be comparatively rare.

[4] *Compare* Alex Sarch, Criminally Ignorant 46–58 (2019) *with* Larry Alexander & Kim Ferzan, Crime and Culpability, 18–19, 27, 67–68 (2009). For discussion of the differences between my view and the view of Alexander & Ferzan, *see* Sarch, Criminally Ignorant 39.

engagement with, the applicable reasons for action.[5] Rather than pegging culpability to the *full level* of insufficient regard that impelled the wrongdoer to act, I argued that it is better to take one's level of culpability for an action to equal the degree of insufficient regard that the action *manifests*.[6]

This, however, led Alec Walen to object that my view is too objective to be an account of criminal culpability and is better understood as a theory of criminal wrongs,[7] which tend to be understood in a more objective way.[8] This is an important challenge because it requires getting clearer about what wrongness and culpability are and what roles they are meant to play on a legal moralist picture of the criminal law. Still, I argue here that Walen's objection ultimately is mistaken. Once we get clear on how wrongness and culpability are to be understood – though as criminal theorists we should ideally do this while remaining neutral on which substantive moral theory is correct – it becomes clear that my account is best seen as a theory of culpability. This is so even though the theory calculates degrees of culpability in a more objective way than other insufficient regard views.

The plan for this paper is as follows. Section 1 outlines what I take to be a fairly orthodox view among legal moralists about how wrongness and culpability are to be distinguished. This shows what theories of criminal wrongs or theories of culpability are supposed to be theories *of*. Section 2 motivates my Manifestation Account of criminal culpability and then presents Walen's objection to it. Section 3 provides the core of my reply to Walen.

Finally, Sect. 4 moves beyond playing defense and aims to develop a more sophisticated version of the Manifestation Account. To help strengthen the case for thinking that the Manifestation Account is properly seen as a theory of culpability, I focus on the main area of criminal law doctrine where culpability comes apart from wrongness – namely, excuse cases. My earlier defense of the Manifestation Account is limited because it has nothing distinctive to say about the culpability of excused misconduct. Accordingly, the theory as presented earlier would be, if not outright false, at best incompatible with the orthodox view presented in Sect. 1, which takes excuses to accept wrongdoing but deny culpability. To solve this problem and enhance the theory's explanatory value, Sect. 4 shows how to extend the Manifestation Account to account for the culpability of wrongs where a putatively excusing condition is in play. This not only fills a gap in the Manifestation Account but has the further benefit (explained in Sect. 5) of providing a unified reasons-based account of the main categories of misconduct in the criminal law, which shows what

[5] Alexander & Ferzan, *supra* note 4 at 67–68 ("insufficient concern [is] the essence of culpability"); Yaffe, *supra* note 1 at 38; Victor Tadros, Criminal Responsibility 250 (2005) (arguing that conviction for serious offenses communicates that one's "behaviour manifested an inappropriate regard for other citizens and their interests"); Peter Westen, *An Attitudinal Theory of Excuse*, 25 Law & Philosophy 289, 373–74 (2006); Simester, *supra* note 1 at 237–38.

[6] Sarch, *supra* note 4 at 51–54; Alex Sarch, *Reply to Commentators*, 12 Jurisprudence 291, 302–07 (2021) (responding to worries about the theory of culpability defended previously compared to competitor views).

[7] Alec Walen, *On Blame and Punishment: Self-Blame, Other-Blame, and Normative Negligence*, 41 Law & Philosophy 283–304, 289 (2022).

[8] *See infra* note 20.

separates criminal wrongness from culpability both in justification cases and excuse cases. My aim through this discussion is to shed light on the distinct roles of wrongness and culpability in legal moralist theories of the criminal law and provide a better understanding of degrees of culpability of criminal wrongs – not only when justifications are at issue, but also when excuses are involved.

1 An Orthodox View of Wrongness and Culpability in Criminal Law

To set the stage, let me sketch what I take to be a fairly orthodox view of the role of wrongness and culpability in the criminal law. I won't defend this view,[9] but merely introduce it to clarify the landscape against which the debate with Walen takes place. On this view, wrongness and culpability are distinct concepts with different roles to play in legislative decisions about which actions should be criminalized. Moral wrongness is a property of actions. To say an action is morally wrong is to say that it is *decisively* disfavored (or in some other binding sense not to be done[10]) by the full set of reasons that bear on how to behave under the circumstances according to the best moral theory.[11] Wrongness thus is about how strong the case in fact is against doing the action in question. Culpability (or equivalently moral blameworthiness) is a property of agents *in respect of their actions*. Culpability concerns not the actor's character in general, but rather what one's action says about how much lack of respect or concern for others (or their interests, rights or other important values) one acted with on the occasion in question. As Andrew Simester nicely puts

[9] I defend its basic features – particularly the difference between the reasons that determine wrongness vs culpability as explained below – in other work. Alex Sarch, *The Role of Wrongness and Culpability in Criminalization* (draft).

[10] This qualification is important because of cases where the balance of reasons count against an action and yet it is not morally prohibited. Suppose one must choose between action A and B both of which are supererogatory and thus both permissible but neither obligatory. Suppose A is much better than B. The reasons for B are not as weighty as they are for A. Thus, B is disfavored by the totality of applicable reasons. But that does not make it wrongful. By hypothesis, both are permissible and indeed supererogatory. S*ee, e.g.,* Stephen Darwall, "What Are Moral Reasons?" *The Amherst Lecture in Philosophy* 12 (2017) 1, 6 (http://www.amherstlecture.org/darwall2017) ("when we say that someone morally ought to act… [w]e mean that they are under a moral duty or obligation and that it would be wrong in the fully deontic sense for them not to do it. We do not mean just that they would be acting against the balance of moral reasons."). Thus, an action's merely being disfavored by the balance of reasons is not sufficient for it to be wrong. The qualification "decisively disfavored" is included in the formulation of moral wrongfulness in the text in order to capture whatever this additional ingredient is that is required for moral wrongness. *See* Edwards & Simester, *supra* note 1 at 55 (discussing different views of what this extra ingredient beyond merely being disfavored by the reasons is that's required for full moral wrongness).

[11] *Id*. at 55 (noting that on all the main views of moral wrongness understood in terms of reasons, "what makes an action wrong is some property of the guiding reasons that bear on that action"); Victor Tadros, The Ends of Harm, 217–18 (2011) ("If it is wrong for D to v, there is a morally decisive reason for D not to v. If it is permissible for D to v, there is no morally decisive reason for D not to v," where decisive here means not defeated e.g. by being excluded or outweighed by other reasons). *Cf*. Heidi Hurd, *Justification and Excuse, Wrongdoing and Culpability*, 74 Notre Dame L. Rev. 1551, 1559 (1999) ("Moral wrongdoing consists of doing an action that violates the maxims of our best moral theory – whatever that theory may be, be it consequentialist or deontological."). I think it's somewhat more illuminating to explain wrongness in terms of reasons, as stated in the main text, than to say it's what the best moral theory prohibits.

it, culpability "arises not from doing the wrong thing, but from how D engages with the reasons why she ought not to have done it."[12] In particular, "D becomes culpable for φ'ing when, in her engagement with the reasons she has not to φ, she was insufficiently motivated by – cared insufficiently about – the interests of others."[13]

This distinction between wrongness and culpability is typically thought to bear on the proper content of the criminal law in a number of ways. Most importantly, the orthodox view assumes that both the wrongness of some bit of conduct and its culpability are both, at the very least, *requirements* for properly criminalizing that conduct and making one who performs it liable to conviction (assuming no defense applies).[14] One might also think that wrongness or culpability, or perhaps both, are reasons that affirmatively support criminalization (perhaps only within the public sphere), although there is more debate about this claim.[15] I won't assume this positive claim is necessarily a part of the orthodox view, though it can naturally supplement the negative constraints that will be the primary focus here. A further aspect of the orthodox view worth emphasizing is that justifications and excuses are also typically understood in terms of wrongness and culpability.[16] Justifications deny the all-things-considered wrongfulness of an action that is pro tanto wrong.[17] Excuses accept the all-things-considered wrongness of the conduct but instead deny the defendant's culpability for it in order to exculpate her.[18] These distinctions in the orthodox view can be mapped onto the conceptual structure of the criminal law, such that 1) offense definitions would ideally correspond to pro tanto wrongs, 2) offenses that lack a justification (like necessity or self-defense) – *i.e.* the unlawful actions – would ideally correspond to wrongs simpliciter (or all things considered), and 3) unjustified offenses that also lack any excusatory defense (like duress) – for which one can be fairly convicted – would ideally track the wrongs simpliciter that also are culpable.[19]

[12] Simester, *supra* note 1 at 17.

[13] *Id*. at 238. *See generally id*. at 237–38.

[14] *See id*. at 32 ("What the law criminalizes, and what the criminal law condemns, is not wrongdoing per se; but *culpable* wrongdoing."); *see also* Berman, *infra* note 23 (discussing the widespread support for desert-constrained pluralism).

[15] RA Duff, *Towards a Modest Legal Moralism*, 8 Crim. L. & Philosophy 217, 229–30 (2014) (distinguishing negative and positive legal moralism as a constraint on vs a reason in favor of criminalization; arguing that the condemnation and censure of the criminal law "must, as a matter of justice, be directed at wrongful conduct and its culpable agent").

[16] This understanding of justifications and excuses is widespread but not universal. *See* Mark Dsouza, Rationale-Based Defences in Criminal Law 3–9 (2017) (discussing the widespread acceptance of this view, which he dubs the "wrongness hypothesis," before raising doubts about the view).

[17] *See* Simester, *supra* note 1 at 17–33.

[18] Westen, *supra* note 5 at 290 ("unlike defenses of lack of actus reus and justification, excuses obtain even when a defendant has done something that society regards as undesirable or regrettable under the circumstances").

[19] *See* Simester, *supra* note 1 at 19 (explaining justification excuse distinction), 29–30 (distinguishing pro tanto wrongs from wrongs simpliciter), 32 (introducing culpable wrongs as proscribed acts that are both unjustified and unexcused) and 39 (noting that a "pro tanto offence done without justification is an unlawful act and a wrong simpliciter").

Given the many substantive views one might take in moral theory, can we say something more in general terms (*i.e.* without appealing to any substantive moral theory) about how wrongness differs from culpability? Perhaps the most important difference is that wrongness is often thought to be more objective in certain respects than the more agent-focused (or subjective) notion of culpability[20] – although strictly speaking this will depend on the specific substantive view of moral wrongness one favors. Still, I think we can fairly say that, at a minimum, the orthodox view takes it that the notion of wrongfulness relevant to criminalization is a function of not only the reasons that were epistemically available to the actor to reason from (or be motivated by, take into account, etc.) at the time in question, but sometimes also reasons that were not epistemically available to the actor at the time. These might include a range of considerations that the actor reasonably was unaware of or would not be expected to act in light of, but which the legislature properly may consider in deciding which conduct to prohibit – considerations like general deterrence and the expressive benefits of criminalization and punishment, and perhaps even considerations like the facts about how much harm one's conduct actually goes on to cause (even if only due to luck).[21]

What does this mean for how to understand the wrongness constraint on criminalization? There are two options. It turns on whether one thinks moral wrongness similarly can depend a least in part on such objective, non-agentially available reasons like those just sketched. If one *does* think non-agentially available reasons can bear directly on moral permissibility, then one can simply take moral wrongness on this more objective view to be a requirement for proper criminalization. Alternatively, if one has a more agent-centered or perhaps deontological view of moral wrongness (which does not look to considerations the actor could not take into account or reason from), then moral wrongness would merely be one consideration among many (sitting alongside deterrence, expressive benefits, etc.) that determine whether there is a sufficiently weighty positive case in favor of criminalizing the conduct. Thus, moral wrongness on this more agent-centered view would not be required for criminalization; rather, the constraint would be that the conduct to be criminalized must be what we might call a "bare wrong"[22] – *i.e.* an action that is decisively disfavored by the balance of the relevant reasons (whether available to the agent or not) that bear on how to behave on the occasion in question. Such a constraint requiring the bare wrongfulness (or objective unjustifiability) of the conduct is, I take it, the *minimum* that would be built into the orthodox view, which its adherents could

[20] Peter Westen, *Unwitting Justification*, 55 San Diego L. Rev. 419, 420 (2018) ("it is commonplace to conceptualize wrongdoing as consisting of the objective harms and risky acts that the state deploys criminal statutes to prevent, all things considered, while culpability consists of an actor's mental capacities and states regarding those harms and risky acts"). For a view that in general normative reasons have moral force even when one is not aware of them (unlike explanatory reasons), *see* Joseph Raz, Practical Reason and Norms 17–19 (1978).

[21] I argue for this view elsewhere. *See* Sarch, *supra* note 9.

[22] I adopt this term from Findlay Stark, *Tort Law, Expression, and Duplicative Wrongs*, *in* Civil Wrongs and Justice in Private Law 441, 443 (Paul Miller and John Oberdiek, eds. 2020).

accept regardless of their view of what makes act tokens morally wrong.[23] But those who would defend a more objective view of moral wrongness might also go further and say that their preferred form of moral wrongness is required for criminalization. I cannot take a substantive position on the difficult and contested concept of moral wrongness, so I prefer to say that the relevant constraint on criminalization according to the orthodox view is just that the conduct in question must be a *bare wrong* in order for it to be properly prohibited.

Unlike the notion of bare wrongness that the orthodox view (at a minimum) requires for criminalization, culpability is assumed to be a function only of the reasons bearing on how to behave that were *epistemically available* to the actor to reason from (or be motivated by, etc.) on the occasion in question – particularly how defective the agent's engagement with these reasons was shown to be by the misconduct performed.[24] Generally it is assumed that conduct must also be culpable (not just wrong in the relevant sense) in order for it to be properly criminalized and punished (ideally that each act token of the proscribed type of conduct be culpable). Further, the amount of criminal blame and punishment that the defendant becomes liable to upon conviction must be proportionate to their culpability for the action they were convicted of.[25] At a minimum, the culpability of an action (understood in terms of the quality of the actor's engagement with the reasons available to her to take account of when acting) sets the upper limit on the amount of criminal blame and punishment that may fairly be imposed for this action pursuant to a conviction,[26] although some think that wrongness might also be an independent ingredient in the amount of criminal blame that is deserved (e.g. if some less harmful forms of culpable wrongdoing might deserve less punishment than equally culpable but unluckily more harmful conduct).

[23] Indeed, it would also be compatible with a form of desert-constrained consequentialism about the justification of punishment, which Berman describes as something approaching the dominant view among criminal law scholars. Mitch Berman, *The Justification of Punishment*, *in* The Routledge Companion to Philosophy of Law 141, 144–45 (Andrei Marmor ed., 2012).

[24] Thus culpability, on the orthodox view, is typically assumed to require a wrongful action – or at least an action properly deemed unlawful by the legislature – in order for one to count as culpable. As Simester puts it, "[w]e cannot blame a person for doing the right thing." Simester, *supra* note 1 at 17. Of course, one may *possess* significant ill will when acting in ways that are legally permissible but this is not manifested in one's conduct in a way that is "visible" to the criminal law unless one violates a legal prohibition. *See infra* Section III.

[25] *See* Mitchell Berman, *Proportionality, Constraint, and Culpability*, *forthcoming* in Criminal Law & Philosophy (defending the view that the proportionality principle applicable to criminal law—particularly in the view of "responsibility-constrained pluralists"—requires punishments to be proportionate to *culpability*).

[26] *See e.g.* Michael Moore, Placing Blame 247 (1992) ("Culpability sets the outer limits of desert and thus of proportionate punishment").

2 Walen's Objection to My Theory of Culpability

The orthodox view as just described serves as the backdrop to Walen's objection to my theory, which was intended to be a theory of culpability in the above sense. Let me briefly explain the view before explaining Walen's objection.

2.1 The Manifestation Account of Culpability

Following the insufficient regard line of theories of culpability,[27] the *Manifestation Account* I defended previously says that a prohibited action is criminally culpable to the degree that it manifests insufficient regard for the interests, rights and values that are properly protected by the law.[28] Insufficient regard is understood in terms of the defects one's action reveals in how one attached weight to the reasons that exist in the circumstances as one believes them to be.[29] That is, an action is culpable to the extent it displays a valuation of reasons that diverges from the correct weights the law ideally says should be ascribed to the reasons available for one to reason from or be motivated by on this occasion.

The distinctive aspects of my view are motivated by the observation, as Walen puts it, that "there is often a mismatch between how bad a person's will was and the amount of blame the criminal law assigns them. Two people with equally bad wills may commit different crimes, or two people may commit equally grave crimes with different levels of bad will."[30] This makes trouble for simple versions of the insufficient regard view (or quality of will theory). The most important competitor to my view is the Causal Account, which takes an offense's culpability to equal the amount of insufficient regard the agent *possessed* and that was causally active in producing the action. The Causal Account says we find an act's culpability by comparing the difference between the agent's actual valuation of reasons and the valuation of reasons the law takes as correct.[31] Consider an unjustified act of theft, for example[32]:

Theft–Causal Account

[27] *See supra* note 5.

[28] Sarch, *supra* note 4 at 50–55.

[29] This formulation, "the circumstances as one believes them to be," is intentionally ambiguous between the facts constituting one's evidence and the facts one infers from one's evidence. Thus, I leave open whether the theory looks to i) the reasons that exist on the facts as one *should* believe them to be (i.e. on one's evidence) and ii) the reasons that exist on the facts as one *honestly* believes them to be (whether reasonably or not). This matters to how culpable one is in cases involving unreasonable mistakes of fact and whether one can be culpable for actions stemming from the unreasonable failure to be aware of risks that one should have seen on one's evidence (*i.e*. negligent conduct). I leave these tricky questions unresolved for present purposes. The theory can be fleshed out in either way.

[30] Walen, *supra* note 7 at 289.

[31] Some complications and amendments to the Causal Account are laid out in my *Reply to Commentators*, *supra* note 6 at 302–07.

[32] This example is based on the one given in *id*. at 303.

Agent	Pro	Contra	Culpability (degree of error) on the Causal Account
Perfectly law-abiding citizen's valuation	1	10	0
Offender's valuation	20	2	(20–1) + (10–2) = 27

The perfectly law-abiding citizen's valuation is the weight the law recognizes as attaching to the reasons in favor of the prohibited act (Pro) and the reasons against (Contra). The degree of difference between this and the agent's actual valuation is then determined as indicated in the table to arrive at an overall assessment of the action's culpability.

Trouble arises for the Causal Account when the defendant harbors tremendous ill will towards her victims, which causes her to perform the crime, but the crime itself is not especially serious. Perhaps the actor commits only a minor offense because this is all she could do on this occasion to harm her victim. If more harm had been possible, suppose she would have inflicted it. In the case reflected in the table, suppose the actor hates the victim and would impose suffering and death if only she could. Instead, the only harm she can impose in this instance (without getting caught and punished, perhaps) is to steal the victim's car. That explains the numbers in the table: the actor sees 20 units of benefit to the action given that it harms the person she hates and she sees little (just 2 units of cost due to the risk of getting caught) that counts against it.

It is implausible, as the Causal Account implies, that the large amount of ill will the agent possesses and acts from means her action is *tremendously* culpable regardless of what that action is or involves. The Causal Account overestimates this action's culpability. Because auto theft is less serious compared to the tremendous ill will the offender acted from, the theft does not fully manifest all the insufficient regard she possessed and acted on. Sometimes the full depth of one's ill will is not fully manifested in what one does, and our culpability judgments should reflect this.

My Manifestation Account was designed to avoid such problems. To find a criminal action's culpability, it says, we look not at the agent's *actual* degree of error in weighing reasons, but the degree of such error the action *manifests*. This is assessed in a more objective way based on the circumstances as she believed them to be, *i.e.* using a principle of lenity to interpret the offender's behavior.[33] More precisely, the culpability of a criminal action equals the degree of error in weighing reasons that exists between i) the perfectly law-abiding citizen's (PC's) valuation of the reasons the law properly recognizes as bearing on the act, and ii) the otherwise law-abiding citizen's (OC's) valuation of these reasons. OC represents *the least departure* in weighing reasons (under the circumstances as the actual agent believes them to be) that would be needed to get someone to do the crime that the agent actually committed. This amount of error, as explained elsewhere,[34] is equal to the difference between PC's valuation of the reasons Contra and the reasons Pro. Why? Because

[33] Sarch, *supra* note 4 at 51.

[34] *Id.* at 53.

to turn PC into OC, all we need to do is assume an error in overvaluing Pro and/or undervaluing Contra that leads to a valuation of reasons where Pro *just slightly outweighs* Contra. Recall the theft example. The next table[35] shows ways one might adjust the weightings of PC to get OC, who is the smallest departure from PC needed to get one to do the offense. Each version of OC is just as culpable.

Theft–Manifestation Account

Agent	Pro	Contra	Degree of error [Manifestation Account]
Perfect Citizen (PC)	1	10	0
OC1 (Raise Pro)	10.1 (raised from 1)	10	9.1
OC2 (Lower Contra)	1	0.9 (lowered from 10)	9.1
OC3 (Mix)	6 (raised from 1)	5.9 (lowered from 10)	5 + 4.1 = 9.1

One benefit of the Manifestation Account, I've argued,[36] is that it reflects a normatively attractive ideal of the criminal law as making charitable assumptions about us as citizens when interpreting our behavior for criminal law purposes. On this view, the criminal law aims to take us to be the least bad version of ourselves that is compatible with what we did.

2.2 Walen's Objection

Walen objects that my theory of culpability is actually a theory of degrees of wrongness. The reason, as Walen puts it, is that on my theory "[t]he actual quality of a wrongdoer's will is, to a large extent, disregarded."[37] Instead, the theory is "concerned with only what [amount of ill will] must be implied to explain that particular wrong"[38] – the amount *manifested*. This avoids the "mismatch" problem of minor offenses that don't manifest the full depth of the actor's ill will; however, the theory manages to correctly assign amounts of criminal blame only by collapsing into a theory of wrongness. Why? Rather than taking culpability to be about the defects in one's actual practical reasoning (as Simester suggests[39]), my theory is concerned with the degree of disrespect that actions manifest, which is objectively assessed in terms of the balance of reasons that exist given the circumstances as the actor believes to them to be. This objective aspect of the assessment seems to push the theory in the direction of an account of wrongness. As seen above, the orthodox picture takes wrongness (whether moral wrongness or just bare wrongness) to be

[35] This table is based on the one in my *Reply to Commentators*, *supra* note 6 at 304.

[36] Sarch, *supra* note 4 at 75–80.

[37] Walen, *supra* note 7 at 289.

[38] *Id*.

[39] Simester, *supra* note 1 at 16–17; *see also id*. at 388–98 (discussing how actual engagement with reasons matters in justification cases).

more objective in nature.[40] Thus, the objective assessment of what amount of ill will a given action manifests would seem to more naturally fit the more objective assessment that wrongness involves. Thus, it may seem questionable whether the theory is an account of culpability at all. Wouldn't it be simpler to just take it to be an account of degrees of wrongness?[41]

This objection matters not just because it reveals confusion about the proper target of the theory I sought to develop, but because it would mean the theory of culpability is not compatible with the orthodox picture sketched above. Culpability, we saw, has a distinct focus on the agent's practical reasoning – the quality of one's engagement with the available reasons bearing on what to do.[42] But if my theory of culpability is not really a function of one's actual reasoning, then it faces difficulties in capturing what the orthodox view takes culpability to be about. My theory might seem most directly to be a form of act evaluation (as is wrongness), not an assessment of agents in respect of their actions (the lack of respect that went into the act), as the orthodox view assumes is true of culpability.

Moreover, if my account is supposed to be a theory of wrongfulness instead, then it would be a poor fit with how the orthodox view conceives of criminal wrongs. My theory makes culpability a function solely of the reasons available to the actor to reason from, be motivated by or otherwise take into account. But the orthodox view took it that wrongness is not necessarily a function just of the reasons that are available to the agent to take into account, but also other more objective reasons not available to the agent. Thus, if my theory really is to be seen as an account of what actions the law may deem to be criminal wrongs, it does not amount to a very plausible account. At best, the theory would capture a subjective picture of moral wrongness which is merely one ingredient in the overall legislative determination of which actions to prohibit. This would threaten to render my theory at best quite limited in scope – and at worst outright false.

3 Answering Walen's Objectivity Objection

Walen objected that my theory is in fact concerned with wrongness because it does not look to the quality of one's actual reasoning (one's actual level of ill will) to assess the culpability an agent incurs from doing a given action. While Walen is right that my theory is somewhat more objective in its operation than other views of culpability,[43] I'll argue it's incorrect to say that the theory "disregards"[44] the actual

[40] *See supra* note 20 and accompanying text.

[41] On my theory, Walen suggests, "[c]riminal blame primarily tracks the wrong, not the quality of the agent's will; and the blame that tracks the quality of an agent's will is not criminal blame." Walen, *supra* note 7 at 289.

[42] *See* Simester, *supra* note 1; *see also* Edwards & Simester, *supra* note 1 at 55. ("[w]hat makes D's φ'ing blameworthy…is D's engagement with those guiding reasons").

[43] *See* Sarch, *supra* note 4 at 34–39 (discussing differences from causal accounts as well as Ferzan & Alexander's view), 40–42 (discussing differences with Yaffe's view).

[44] Walen, *supra* note 7 at 289.

quality of the agent's will – *i.e*. the defects in her actual engagement with the applicable reasons bearing on how to act.

In fact, the culpability attributions licensed by my account do speak in important ways to the actual ill will (or defectiveness in one's engagement with the reasons one had access to) at the time of one's misconduct. Even on my Manifestation Account, ill will must still be *present* for culpability to be attributed for an action. In particular, as I previously explained my view, attributing criminal culpability requires a "repulsion failure" to be present which was causally active in producing one's misconduct.[45] Here is what this means. A convenient way to think of the demands of the criminal law is that we have a "repulsion mechanism" whose job is to be motivationally responsive to the factors in virtue of which actions are criminally prohibited and which is supposed to kick in to motivate us not to perform actions that are properly prohibited.[46] The law is "indifferent to the content of one's repulsion mechanism,"[47] as self-interested motivations can be as effective in getting one to avoid criminally wrongful conduct as a concern for others or even respect for law. However, it's a requirement for a prohibited action to manifest insufficient regard (*i.e*. be criminally culpable) that the action was caused, at least partially, by a failure of one's repulsion mechanism.[48] This, in turn, requires that one's action was actually called upon to do some motivational work under the circumstances to get one to refrain from a criminally prohibited action, but the repulsion mechanism failed to do its job and one actually went ahead and performed the action nonetheless. Only when one's action is in this way at least partially caused by a failure of one's repulsion mechanism would the ill will one possesses be *manifested* in the action one performed.[49]

Thus, my theory takes it that a prohibited action is culpable only if caused at least in part by ill will. As a result, the theory always imputes at least a portion of the ill will that was actually present and helped produce the conduct. Once we know that some ill will helped cause the prohibited action, the next question is *how much* ill will a given action manifested. This is where the more objective aspect of the theory comes in. As explained in Sect. 2, the theory applies a principle of lenity under which an action is taken to manifest only the least amount of ill will needed to get an otherwise law-abiding citizen to do the act the defendant did under the circumstances. Because of this lenity-driven approach to determining how culpable an action is (an approach adopted in part to resolve the mismatch problem), my view will sometimes impute *less* than the full amount of ill will that was present at the

[45] Sarch, *supra* note 4 at 48.

[46] This talk of a "repulsion mechanism" is meant just as a convenient way to speak about the motivational expectations of the criminal law. As used here, it's not meant to suggest that there is a unified psychological phenomenon or natural kind that corresponds to this sort of repulsion mechanism. It is used merely as an expository device.

[47] Sarch, *supra* note 4 at 47.

[48] *Id*. at 48.

[49] *Id*. at 47-48 ("Only when the failure of your repulsion mechanism helps *cause* your conduct…is it *manifested*…. [E]ven if the safeguard mechanism whose job it is to keep you within the bounds of legally justifiable conduct is faulty, this failure won't be *manifested* until the mechanism is actually called on to do its job, but doesn't. Before that, the failure of the mechanism is just a latent defect that has not yet been manifested.").

time of acting and that was causally active in producing the wrongful action. But the theory never imputes *more* ill will than one actually possessed.[50] As a result, there is a meaningful sense in which the theory is still about one's actual engagement with reasons – *i.e.* how much lack of respect for others we can discern in how one acted. As such, it is still an instance of what Edwards and Simester call the engagement view, which I take it reflects the orthodox position among criminal law theorists.

One might question this and ask if there really are cases where my view imputes more ill will, and thus culpability, for an action than the amount the agent possessed when acting. Most importantly, consider a very hesitant and anguished offender who does a criminal action – suppose it's a theft – with high degree of reluctance and regret. Call her Jane. She strongly feels the pull of the reasons against the action – suppose she gives exactly the correct weight to the reasons Contra – and just barely manages to do the criminal action. Thus, one might think she possesses only a small amount of ill will. To be more precise, suppose Jane's actual valuation of reasons takes it that Pro = 10.1 and Contra = 10 (assuming units of reason come in increments of 0.1). However, suppose that the correct valuation (as in the Theft example above) is Pro = 1 and Contra = 10. Isn't this a case in which my theory might impute more culpability than the real amount of ill will that the agent actually possessed at the time? The actual amount of ill will Jane possessed was 9.1. After all, this is the degree to which Jane overvalued the reasons Pro, and she gave the right weight to Contra.

Nonetheless, my account does not impute more culpability than the amount of ill will Jane actually possesses. The amount of ill will the theory imputes in its culpability assessment of this action is also 9.1. After all, this is the least amount of ill will that must be present to get an actor to do the wrongful action under the circumstances. Thus, in this case, the Manifestation Account attributes the *same amount* of culpability for the action as the amount of ill will that actually was present. The theory sometimes attributes less culpability than the full amount of ill will possessed at the time of acting, and sometimes the same amount – but never more.[51]

For this reason, the culpability judgments generated by the Manifestation Account does always reflect on the agent's actual engagement with reasons at the time of acting – her actual valuation of the weight of the applicable reasons. It does not always reflect the full depth of depravity the offender may possess in their heart (which we might even *know*, through admissions or other communications, was causally active in producing the conduct in question). But the theory always attributes an amount of culpability that bears some relation to how much ill will was actually present. Sometimes it attributes less, but it doesn't attribute more. In this way, the theory – despite its greater objectivity than other views – still concerns the quality of one's actual engagement with reasons and does not entirely disregard it. It looks to the degree of ill will that is manifested by one's action – *i.e.* an amount one's action showed one

[50] *See* Sarch, *Reply to Commentators*, *supra* note 6 at 304 (arguing that "the Manifestation Account always produces culpability judgments that are no harsher (and often less harsh) than the Causal Account").

[51] Note that I previously proposed a different reply to the worries about this sort of case, but subsequently abandoned it. *See* Sarch, *Reply to Commentators*, *supra* note 6 at 302–03.

to possess *at least as much as*. For this reason, the theory does not disregard one's actual practical reasoning and so I resist Walen's suggestion that the theory really concerns wrongness rather than culpability as the orthodox view conceives of it.[52][53]

4 Extending the Manifestation Account to Excuses

To strengthen this response to Walen, I now want to answer a further concern. Another way the Manifestation Account might be thought to be overly objective is that it fails to account for the sort of case where the agent-centered notion of culpability, as distinct from the more agent-neutral notion of wrongness, is most clearly on display – namely, cases of excused wrongdoing. As I presented it in earlier work, my theory has little to say about culpability in excuse cases. Thus, it not only is incomplete as a theory of culpability, but it seems incompatible with the orthodox view presented in Sect. 1. Here I set aside cases of exemptions, where one's cognitive faculties are so impaired that one falls below the minimum threshold for being an agent who is aptly subject to judgments of blame.[54] Instead, the worry for my account concerns "rationale-based excuses" like duress and provocation, where the agent remains a full actor who performed a genuine action but has reduced

[52] Note that my Manifestation Account also answers the three other arguments against quality of will views that Walen presents. Walen, *supra* note 7 at 288–89 (drawing on Doug Husak, Ignorance of Law 166–67 (2016)). First, the Manifestation Account does not attribute culpability merely for character. Instead, in assessing the ill will displayed by a particular action, the Manifestation Account only attributes culpability for a particular action. Second, the Manifestation Account does not require that the agent *intended* to manifest a certain amount of ill will for the action to do so. The theory assesses the weight given to reasons that is needed under the circumstances to explain why the actor did what she did – regardless of whether the agent had any intent or desire to express anything (ill will or otherwise) via her conduct. Third, the Manifestation Account does not consider the whole history or available evidence about everything the actor did on other occasions to assess culpability. What matters to culpability, on my view, is the amount of ill will the particular bit of misconduct puts on display – not the full level of ill will we might otherwise know she possessed. *See* Sarch, *supra* note 4 at 41–42 (discussing similar problems for a view like Yaffe's evidential approach).

[53] An anonymous reviewer suggests that I could further strengthen the response to Walen by appealing to the role of mens rea. One might think causing a certain amount of criminal damage intentionally is more culpable than causing that same amount, say, recklessly. If one can in this way commit the same wrong with different levels of culpability, it might seem to further strengthen the case for thinking that culpability, on my picture, is distinct from wrongness. While tempting, I think this response to Walen ultimately won't suffice. The reason is that it will be difficult to rule out that these distinctions can all be captured within the concept of wrongness. Unless one has a substantive view of wrongness that suggests otherwise, it's not clear why we shouldn't say these are simply two different wrongs: one of intentionally causing X criminal damage and the other causing X damage recklessly. Thus, if we allow for fine-grained distinctions between wrongs (and it's not clear why we shouldn't), then the wrongness standard could capture mens rea gradations without having to appeal to a distinct notion of culpability. Given this threat from a fine-grained theory of wrongness, I think I can't rest my response to Walen solely on the contribution to culpability that mens rea might make. It could in principle all be captured within wrongness. For more on this sort of collapse worry, *see infra* notes 58, 60 and 67.

[54] *See e.g.* Antony Duff, Answering for Crime 285–86 (2007).

culpability for their wrong due to their excuse.[55] (What to say about incapacity defenses like insanity, and whether they sometimes function as genuine excuses not exemptions or denials of an offense element, is a complex issue, which I mostly set aside in what follows.[56])

The challenge for the Manifestation Account is this. The orthodox view understands culpability as improper engagement with the reasons available to the agent, while taking it that excuses accept the wrongness of one's action but deny one's culpability for it (at least partly). However, as presented, the Manifestation Account offers no distinctive way of calculating the culpability of excused actions. It would simply apply the culpability calculus presented in Sect. 2 to such actions. But this incorrectly entails that excused wrongs remain fully culpable. This is what follows from the perfectly law-abiding person's weighting of the relevant reasons, given that excused actions (supposing there are any) are assumed to remain all-things-considered wrongs that at least a *perfectly* law-abiding person would manage not to do.[57] Accordingly, the Manifestation Account as presented fails to capture what the orthodox view assumes about genuine excuses (as opposed to exemptions or justifications): namely, that they fully accept wrongness but still eliminate or reduce culpability. [58]

Thus, to make the Manifestation Account compatible with the orthodox view and allow it to explain why genuinely excused wrongs (if any) are properly viewed as having reduced culpability, I show how the account can be amended to apply also to excused wrongdoing. This would allow the theory to account for the cases in which the distinctive contribution of culpability is most easily discernable, namely those where a prohibited action is admittedly wrongful (at least in the sense of bare wrongs from Sect. 1) but which nonetheless have reduced culpability. This reinforces the conclusion that the theory properly concerns culpability rather than a

[55] Simester, *supra* note 1 at 15, 422–23 (distinguishing "irresponsibility defenses" like insanity and infancy, which speak mainly to eligibility for moral evaluation, from "rationale-based" excuses, which speak to culpability).

[56] *But see infra* note 80.

[57] As noted, on the orthodox picture, the line between lawful and unlawful conduct corresponds to the law's view of what actions are permissible vs. wrong all things considered (equivalently, simpliciter). *See* Simester, *supra* note 1 at 39 ("A pro tanto offence done without justification is an unlawful act and a wrong simpliciter.") So the perfectly law-abiding citizen is assumed to abstain from acts that are all things considered wrong in the law's view, including excused actions.

[58] This assumes that there really *are* cases of excused actions – *i.e*. those the law should see as all things considered wrongs that lack a justification, but for which the actor has reduced culpability. For any particular case of a rationale-based excuse like duress or provocation, one might argue that it should really be recast as a justification, in which case the perfectly law-abiding person would *not* be expected to abstain from it. Indeed, as David Owens notes, *rationalists about responsibility* hold that all excuses reduce to either justifications or exemptions. David Owens, *Excuse, Capacity and Convention*, ROUTLEDGE HANDBOOK OF RESPONSIBILITY (Max Kiener, ed, forthcoming) 4. As I argue elsewhere, it is implausible that we can reduce *all* rationale-based excuses to justifications – at least without losing something important. *See* Alex Sarch, *Excuses, Excuses: A Fair and Humane Account* (draft). Nonetheless, for present purposes, I set aside this problem and simply assume that there are genuinely excused actions of the rationale-based type, and my aim is to provide an account of how these would function. The question confronted here is how to make sense of this category of excused actions, if any, on the Manifestation Account.

theory of wrongness as Walen suggested. It also offers additional theoretical benefits, summarized in the conclusion.

4.1 What We Want to Capture: A Closer Look at Duress

To show what we want a theory of culpability to capture, consider a case of the excuse of duress:

> *Duress*: The Spider Crew wants to sell some enriched uranium on the black market. They go to Dr Alina Johnson who works in the lab in the city and tell her to steal it for them after closing tonight or else they will kidnap her daughter and "hurt her in ways you can't imagine." She knows they have committed acts of violence in the past and has no doubt they wouldn't hesitate to carry out the threat if she didn't comply (or went to the police). So she gives in to the threat, steals the uranium and hands it over to the Spider Crew.

Alina plausibly has a duress excuse to the crime of theft. On the English law test, the duress defense is available only if a "sober person of reasonable firmness" with the legally cognizable characteristics of the defendant (like age, gender, illness, pregnancy) would have behaved the same way.[59] For this to be a case of excuse not justification, it would have to be the case that the action remains all things considered, or in an impersonal sense, wrongful (which a *perfectly* law-abiding person would manage to abstain from even if many of us would not) while Alina's culpability for it is reduced, resulting in acquittal. How might this be the case?[60] Assume the probability-adjusted harm of allowing enriched uranium to get onto the black market is much greater than of one individual experiencing grievous bodily harm and possible death. This would preclude a necessity (or lesser evils) justification, which typically requires the harm one's conduct can reasonably be expected to prevent is greater than the harm it can reasonably be expected to cause (such that it's really the lesser evil).[61] Alina therefore has done something all-things-considered wrong and unlawful[62] – something, let's suppose, that the weight of reasons on the facts as she believes them to be decisively disfavors. Despite being wrongful (at least a

[59] *R v Graham* [1982] 1 All ER 801, 806 (Lord Lane CJ).

[60] For present purposes, I simply assume this case involves an excuse not a justification. I cannot here refute the view that any excuse can be recast as either a justification or an exemption. If you disagree that Alina's case involves an excuse, you can substitute in your preferred case. My aim is just to explain how such cases, assuming there are some, are to be understood.

[61] *See e.g. Re A (Children) (Conjoined Twins: Surgical Separation)* [2001] Fam. 147, HL (approvingly discussing Sir James Stephen's account of necessity on which the defense requires, among other things, that "the evil inflicted by [the misconduct] was not disproportionate to the evil avoided"); *see also* Westen, *supra* note 5 at 301 (using the notion of justification in general to refer to "a claim by a defendant that in so far as he effectuated the harm or risk that an actus reus prohibits, he was allowed to do so because the harm or risk he effectuated was no greater than the alternative evil that he would have had to choose under the dilemmatic circumstances in which he found himself").

[62] *See supra* note 57.

bare wrong in the sense outlined above[63]), however, the duress she faces reduces her culpability sufficiently to prevent conviction. She engaged with the available reasons in such a way that, although wrongful (not supported by the balance of reasons), we nonetheless have sufficient sympathy for her that we do not see her as properly meriting legal condemnation. Her act didn't manifest *enough* insufficient regard to merit the condemnation built into a conviction.

There are two ways to understand why Alina, despite having acted in a way the law takes to be wrongful, still merits exculpation. The dominant model sees excuses as cases where the actor's rationality (or will) was overcome by emotion in ways we find understandable, even if the circumstances do not sufficiently justify the conduct to render it non-wrongful.[64] While the balance of applicable reasons disfavors Alina's conduct, we would find an emotion-driven reaction along the lines of what Alina did to be understandable and sympathetic enough to merit exculpation even though her conduct was wrongful.

One concern, however, is that some cases that seem to merit a duress defense don't involve an emotional reaction that overwhelmed the actor's rational assessment of the reasons.[65] Sometimes actors who seemingly merit an excuse are cool and collected and yet still do the crime – not from an emotional disturbance but from calmly assessing the circumstances. If some such actors should be exculpated too, as plausibly would be the case for Alina even if she calmly assessed the situation and chose to behave as she did in the above case, this would require an alternative theoretical basis than emotion overwhelming reason. One plausible account distinguishes the objectively applicable (third-personal) reasons bearing on the action and the agent-relative (first-personal) reasons that reflect the individual prerogatives or privileges citizens have to pursue personal priorities. It's not implausible to think there is an agent-relative prerogative to inflict some degree of impersonally unjustified harm to defend oneself and loved ones from grave threats. This might be recognized on an excusatory basis out of sympathy for the defendant's plight and as a "concession to human frailty," rather than the objective balance of reasons which justifications focus on.[66]

[63] *See supra* note 22. Roughly, to say that A is a bare wrong is to say that the relevant reasons bearing on what to do that the legislature can legitimately look to for criminalization purposes decisively disfavor A.

[64] Christopher Bennett, *Excuses, Justifications and the Normativity of Expressive Behaviour*, 32 Oxford J. Legal Studies, 563–81 (2012). He observes that on the typical model of excuses, "[t]he defendant's will is said to have been 'overborne' by the nature of the circumstances [and] although she acted wrongly, [an excuse like] duress represents a necessary 'concession to human frailty' to mark the moral difference that her situation and its attendant emotions makes." *Id*. at 564. *See also* Paul Robinson, *Criminal Law Defenses: A Systematic Analysis*, 82 Columbia L. Rev. 199, 221–22 (1982) (offering a general account of excuses).

[65] A similar point is recognized by Mitch Berman and Ian Farrell in *Provocation as Partial Justification and Partial Excuse*, 52 William & Mary L. Rev. 1027 (2011), although they make the point in the provocation context and take it to support reduced wrongfulness rather than reduced culpability. *See id*. at 1056. On the Manifestation Account of culpability, Berman and Farrell's example of an excuse that does not involve being overborne by emotion could just as well be read as a case of reduced culpability.

[66] *See* Westen, *supra* note 5 at 301 (explaining justifications generally as instantiating a "lesser evils" structure).

Regardless of one's account of excuses, duress exculpates those who commit offenses in response to a threat that a "sober person of reasonable firmness" would not be able to withstand. This helps show how culpability comes apart from wrongness in criminal law theory. Suppose that, following the orthodox view, we understand the notion of wrongness relevant to the criminal law as the conduct that is decisively disfavored by the applicable reasons, where these reasons do not all have to be available to the agent at the time of acting and can include a range of agent-neutral or impersonal considerations that the legislature may legitimately look to in deciding which conduct to prohibit (such as harm, deterrence, expressive benefits, and providing reassurance to victims). If we make this assumption, then culpability stands as a distinct notion focusing on what an action says about whether one has engaged with reasons in at least a minimally tolerable way. Considering excuses shows us that what is "minimally tolerable" here is not marked out just by the line between unlawfulness (wrongfulness) and lawfulness, but rather suggests that we also can tolerate some unlawful (wrongful) conduct in certain especially sympathetic circumstances.

This shows a distinct contribution made by the notion of culpability. Suppose one wanted to understand all of criminal law simply in terms of wrongness and wrongness were the only relevant normative concept in one's legal moralist view – perhaps for reasons of conceptual parsimony.[67] This would produce an impoverished view of excuse cases. To secure the result that Alina should be exculpated, a theory of the criminal law that could only appeal to wrongness would have to say that Alina's conduct simply is not wrong – *i.e*. not properly prohibited. But this is an impoverished account of Alina's case. It provides no acknowledgment of the reasons in virtue of which the legislature may have legitimately chosen to mark out the conduct at issue as an offense, including the reasonable concern to deter theft and facilitation of terrorist activities and to express our condemnation thereof through criminalization. Surely these reasons stand even in cases like Alina's where we can also have sympathy for giving in to such extreme threats. Therefore, it would be more accurate and illuminating to accept that the conduct is wrongful, thus fully acknowledging the legislature's reasons for marking it out as a wrong even in cases like Alina's. Instead, the exculpation is better seen as stemming from viewing the conduct as non-culpable. Why? Because abstaining from the offense under Alina's circumstances is not something it seems fair and humane to expect citizens to do, as even a person of "reasonable firmness" would behave similarly in her place. Alina's conduct, even if wrongful and legitimately deemed unlawful, does not cross beneath the threshold of action that displays the minimum level of regard for others that we expect everyone in society not to dip below.

[67] There are good examples of high-quality work in the legal moralist tradition that proceed solely in terms of wrongness without much loss. *See* Law Commission, 2010 *Consultation Paper on Criminal Liability in Regulatory Contexts* (https://www.lawcom.gov.uk/app/uploads/2015/06/cp195_Criminal_Liability_consultation.pdf).

4.2 Extending the Manifestation Account to Excuse Cases

As presented above, the Manifestation Account of culpability still faces difficulty in such cases. I have offered no worked-out account of how to understand wrongful but excused actions. The Manifestation Account as presented earlier said an action manifests an amount of insufficient regard equal to the degree of error in weighing reasons that exists as between: i) the perfectly law-abiding citizen's (PC's) valuation of the applicable reasons and ii) the otherwise law-abiding citizen's (OC's) valuation of these reasons. OC represents *the least departure* in weighing reasons (given the circumstances as the agent believes them to be) that is needed to get someone to do the offense that the agent actually committed. Schematically, Alina's conduct can be represented as follows:

Unjustified Nuclear Theft/Facilitation of Terrorism

Agent	Pro	Contra	Degree of error [Unamended Manifestation Account]
PC	20	50	0
OC	50.1	50	30.1

Thus, the theory as stated entails that even crimes done with excuses are culpable. The theory needs to be amended to explain why offenses done with excuses are not culpable.

Nonetheless, I contend, the theory can be extended to account for such cases.[68] Here is my proposal for how to do it. When an offense is done without sufficient justification but excusing conditions are present (whether sufficient or not), the Amended Manifestation Account would say the culpability of the action is calculated differently than when no excuse is present. Rather than finding the difference between the weights attached by PC and OC, as can normally be done, we should instead look to the weights attached to the reasons Pro and Contra that would be attached[69] (on the facts the actor is aware of) by a *person of reasonable firmness* (RF) who has the same legally cognizable characteristics and faces the same pressure (threats, coercion, provocation, etc.) as the actual actor. RF thus is supposed to reflect our shared judgments about which actions done in response to pressurized circumstances meet the minimum standard of acceptability that we insist on as a

[68] Note I mainly seek to extend the Manifestation Account to cover rationale-based excuses like duress or provocation, not irresponsibility defenses like incapacity or infancy, which likely function mainly as exemptions. The account offered here applies only assuming the defendant meets whatever capacity thresholds are properly recognized by the criminal law. *But see infra* note 80 (discussing what to make of incapacity excuses, if any).

[69] In assessing the weight "attached" to reasons here, we are not asking what the actor would say if asked. Particularly someone facing an overwhelming emotional reaction could perhaps not be expected to reliably answer such questions in the moment. Rather, what matters to the "weight" at issue here is the degree of motivational pressure to do or abstain from the action one feels owing to the particular consideration in question – or *should* feel if we're looking at PC or RF.

society – *i.e*. which we would have sufficient understanding and sympathy for to tolerate in our community even if wrongful.[70] Note that in saying that even a person of reasonable firmness would perform a given wrongful action, I don't mean to suggest that this makes the conduct *itself* reasonable or non-wrongful. It's only the *firmness* in question that is reasonable, although it admittedly might be clearer to use "minimally acceptable" to modify "firmness." If the action is genuinely excused (not justified), it must remain all-things-considered wrongful such that it would not be performed by a perfectly law-abiding citizen (who may have especially high levels of fortitude). And for excuse cases, this will be so even if the wrongful action also has reduced culpability because a person of reasonable (or minimally acceptable) firmness would not manage to abstain from it in light of their vulnerabilities or limitations that we can have sympathy for. I continue to present the view in terms of what a person of reasonable firmness (RF) would do to fit better with the legal test (though would be equally happy with other terminology, such as speaking of the person of "minimally tolerable" firmness).

Given all this, the Amended Manifestation Account would hold that when there is credible evidence (enough to meet the burden of production) that D was in excusing conditions when she committed offense A, and the issue is whether these conditions are enough to exculpate this otherwise unjustified wrongful conduct, then *the culpability of A equals the degree of error as between i) OC's valuation of the applicable reasons and ii) RF's valuation thereof (not PC's).* However, when it's clear that *no excusing* conditions apply, then the unamended account would still give the right results. Thus, on the amended account, when an action is fully excused – *i.e*. when RF would also do the action – no insufficient regard is manifested.[71] Even if PC would (perhaps heroically) abstain from this conduct, such that doing this excused action involves a departure from the standard of legal perfection, this is not *manifested* by the action – *i.e*. we do not find the regard given to the interests of others to be *insufficient* or below the minimum standard of due regard expected in our community. Normally, RF's valuation of reasons will coincide with PC's; they come apart only in excuse cases. While the Amended Manifestation Account takes RF to be the operative measuring stick for culpability in general, this yields different culpability results from the unamended theory, which assesses culpability by reference to PC, only where excuses are in play.

More precisely, the Amended Manifestation Account gives distinctive results (compared to the original theory) in two cases where D finds herself in excusing conditions: 1) those where RF *would* still commit the offense even though it's unjustified and 2) those where RF would be expected to resist the emotional pressure that is present and refrain from the action. Consider each in turn. First, recall Alina who

[70] Note that I briefly suggested a way to understand excuses under the Manifestation Account. *See* Sarch, *supra* note 4 at 106, ftn. 65. The proposal in the main text can be seen as one way to flesh out that brief suggestion, although it departs in some ways from what I gestured at previously.

[71] One might wonder if this also holds in cases involving those who act out of cold indifference. I address this worry at the end of this section and show how even such cases can be made to fit the claims in the main text here.

is exculpated because a person of reasonable firmness would not be expected to refrain from behaving as she did. The culpability of her action is given as follows:

Excused and Unjustified Nuclear Theft

Agent	Pro	Contra	Degree of error [Amended Manifestation Account]
PC	20	50	0
RF	55	50	0
OC	50.1	50	x

Alina's action manifests no insufficient regard because although PC (perhaps heroically) would not do the wrongful action, culpability on the Amended Manifestation Account is measured against the yardstick of RF (as compared to OC, not PC). And RF *would* do the action. So OC's valuation of reasons would contain no *error* relative to RF's.

Now take a case where D is in excusing conditions but RF *would* be able to abstain from the offense in the agent's shoes. Thus, consider Mark. Suppose the criminal gang threatens to smash the windows of Mark's convertible unless he steals the nuclear material for them. Perhaps the car has sentimental value for him, so he is overwhelmed by an emotional response that induces him to commit the theft. Even if RF's weighing of reasons were somewhat distorted by an understandable and sympathetic emotional response to the pressure applied to Mark, RF nonetheless would not give in to the threat. Here, to find the culpability of the action, we compare OC's valuation of reasons to RF's. We ask what the smallest departure from RF's valuation of reasons would be that is needed to get someone with D's legally cognizable characteristics to do the crime under the circumstances, and that is the insufficient regard it manifests.

Partially Excused Unjustified Nuclear Theft

Agent	Pro	Contra	Degree of error
PC	10	50	0
RF	15	50	0
OC	50.1	50	35.1 [= 50.1 – 15 + 50 – 50]

The action here manifests insufficient regard because both PC and RF will refrain from the theft despite the pressure that is present. The amount manifested is determined by comparing how OC weighs reasons to how RF does. (This amount, 35.1, which is obtained by comparing OC to RF, is less than if we'd compared OC to PC, which would have yielded culpability 40.1.)

Thus, the Amended Manifestation Account can account also for excusing conditions whose influence on the agent's practical reasoning we find sympathetic and are willing to tolerate as a society. When such excuses are present, it becomes clear that the operative yardstick for measuring culpability comes

not from comparing OC with PC but rather with RF. Outside these two sorts of cases where excusing conditions are present, RF's weighting of reasons will be the same as PC's and so the Amended Manifestation Account lines up with the results of the original version of the account in cases involving no excuses.[72]

The Amended Manifestation Account thus can capture the distinct contribution made by the notion of culpability as opposed to wrongness, which is most visible in excuse cases. Usually, unjustified (bare) wrongs will be criminally culpable as well (provided no excuse applies). This is what the original Manifestation Account entailed. However, by extending the account to explain why acts like Alina's remain wrong but aren't culpable, the Amended Manifestation Account is on stronger footing as a theory of culpability. The Amended Manifestation Account plausibly accounts for cases like Alina's: it fully recognizes the legislative considerations that make it legitimate to deem Alina's conduct to be a (bare) wrong while still giving full recognition to the aspects of the case that lead us to sympathize with and ultimately condone her practical reasoning, thus negating her culpability.

Furthermore, by focusing on culpability as distinct from wrongness in excuse cases, we can see that the threshold of minimally tolerable engagement with reasons does not perfectly track the line between (bare) wrongful and lawful conduct. Rather, certain tokens of act types that are properly deemed legal wrongs might nonetheless *not* involve engagement with reasons that dips below the level of due regard that we expect from everyone in society. That is, actions like Alina's, even if properly deemed legal wrongs, do not manifest *insufficient* regard. The line between actions that make a conviction deserved and those that don't should track this line between acts that manifest *insufficient* regard and those that don't, not the line between unjustified wrongs and lawful action.

4.3 An Objection: Indifference and the motivations of excused actors

An anonymous reviewer helpfully points out that one might still worry that the Amended Account has trouble with deviant cases involving especially indifferent actors. Suppose Alina* is actually motivated to do the theft not because of the threat or a concern for her children's well-being, but because she is insufficiently repulsed by the wrong-making features of the theft. Here the theft might still seem quite culpable. Can the Amended Manifestation Account capture this?

[72] This raises a worry: Can we really distinguish cases where excusing conditions exist but are *insufficient*, *i.e.* where RF would refrain from doing the crime, from cases where *no* excusing conditions exist at all, where RF likewise would not do the crime? Aside from looking to the elements of the full excuse itself, is there any way to separate these cases? This is a good practical concern for implementing the amended account, but it is not a problem for the substance of the theory. I assume excusing conditions come in degrees such that there can be partial excuses – *i.e.* excusing conditions that are insufficient to fully exculpate. Mark was such an example. However, I accept that if this approach does not prove workable, then for practical reasons we might have to adopt a simpler alternative. For example, we could say simply that there is no culpability when the elements of the full excuse *are* met, and otherwise (when those elements are not satisfied), culpability would be calculated as the unamended version of the theory suggests. However, such practical issues are not paramount in the present context, where we just are aiming to excavate the deep structure of criminal law concepts.

One response suggested by some of my earlier work on manifestation in the context of justifications might be to bite the bullet and say that even if Alina* *possesses* significant ill will in this case, it is not *manifested* in this action because of the threat that she is aware of.[73] However, this seems not to be the most plausible route to take for excuse cases. After all, the basis for excuses offered above is a sense of sympathy or compassion toward the defendant in light of the especially challenging circumstances they faced.[74] But given the indifferent way Alina* was motivated, it is not very plausible that our sympathy would be engaged. It is for similar reasons that excuses are typically conditioned on one's conduct being caused in the particular way the excuse specifies, rather than that one caused the conditions of one's own excuse or was motivated by the prospect of avoiding liability due to having an excuse.[75] If one is motivated in either of these dubious ways, our sympathy likewise would not be engaged. So there is little normative basis for an excuse.

How to capture this on the Amended Manifestation Account? The simplest way is to make use of another bit of machinery from my original account of culpability. I take it that in assessing culpability, what matters is how large a defect in practical reasoning it would take to get someone in the defendant's circumstances to do what defendant did *under whatever the privileged description of this act is*. In legal contexts, the relevant description is naturally supplied by the applicable legal doctrine (or for normative culpability judgments, what this legal description should be). Thus, when an excuse is in play, the relevant description of the actor's conduct, for purposes of assessing culpability, should specify the manner in which the putatively excused conduct was motivated.[76] This is needed to ensure that the action is one where our sympathy is genuinely engaged. The relevant description of Alina*'s conduct, thus, should be something like: *committing the nuclear theft from indifference to the harms and risks involved,* rather than due to the threat or concern for her children. It will require a large amount of insufficient regard for legally protected interests to get one to do the action described in this fine-grained way.

Accordingly, for duress to genuinely reduce one's culpability, it plausibly must be the case that one performed the wrongful action *from duress – i.e*. in response to and out of concern for the interests of those threatened. Only then would our sympathy be likely to be sufficiently engaged to see the actor as having reduced culpability. Thus, in assessing the culpability of a putatively excused action, to rule out cases of indifferent motivation like Alina*, it's plausible we should apply the Amended Manifestation Account to the actor's conduct described so as to include the particular motivations that were causally operative. This accords well with legal doctrine

[73] Sarch, *supra* note 4 at 54–58.

[74] *See* Erin Kelly, *What is an Excuse* (Coates & Tognazzini, eds.) 244–62 (2012) (exploring the idea of sympathy as the basis for excuses); s*ee also* Sarch, *supra* note 57.

[75] *See* David Owens, *Excuse, Capacity and Convention*, Routledge Handbook of Responsibility (Max Kiener, ed, forthcoming) 4; John Gardner, Offences and Defences at 138 ("To attempt to benefit from a legal excuse by being guided by it is to forfeit that excuse.").

[76] Tadros, for example, thinks this approach should be used for justifications as well, though I officially remained neutral on this issue in my earlier work. *See* Sarch, *supra* note 4 at 37-39.

as well.[77] (Depending on one's views about the relevance of motives to justifications, this may mark a difference from the way the culpability of putatively justified actions are to be assessed, where motives may not matter since justifications do not similarly involve a sympathy-based relaxation to our expectations.[78])

5 Concluding Remarks: Toward A Unified Framework for Criminal Wrongs and Criminal Culpability

In this paper, I have defended the Manifestation Account from Walen's objection and shown that it is best seen as a theory of culpability as distinct from wrongness, as these notions are understood on an orthodox view. Moreover, I showed how the Manifestation Account can be extended to cover the cases that most clearly show the separate role of culpability as distinct from wrongness, namely cases of excused wrongdoing.

One payoff for extending the Manifestation Account to cover culpability not only for legally wrongful (and unjustified) actions but also for excused actions is that it provides a convenient way to give a unified reasons-based account of all the main categories of criminal misconduct as conceived on the orthodox view. That is, it suggests a simple picture, which proceeds in terms of reasons, of a) when acts are wrongful and how wrongful they are, as well as b) how culpable wrongful unexcused actions are, plus c) how culpable wrongful excused actions are.

First, to account for the degree of bare (or legal) wrongness of an action, we can simply apply the conception of wrongness baked into the orthodox view. To assess the bare wrongness of an action, we would consider the degree to which all the reasons that bear on what to do in a given choice situation decisively disfavor the action under consideration, where these reasons do not need to be all epistemically available to the actor to reason from, but which may properly be taken into account by the legislature for purposes of deciding what conduct to deem unlawful.

Second, the culpability of unjustified wrongful actions in a non-excuse case is to be understood along the lines of the Manifestation Account (whether the original or the amended version, as they are indistinguishable in such cases). This involves narrowing our assessment from considering all the reasons bearing on how to act – regardless of their availability to the agent to reason from at the time – to instead considering just those reasons the agent could reasonably be expected to reason from on the occasion in question. Thus, on the Manifestation Account, to assess the culpability of a wrongful action which may or may not be justified, but in which no

[77] *See*, *e.g.*, *R v Martin* (1989) 88 CR App R 343 (requiring for duress of circumstances that the defendant acted reasonably "*in response* to a threat of death or injury" and was "*impelled* to act as he did because…he had good cause to fear death or…injury") (emphasis added); *see also R v Graham* (1982) 1 WLR 294 at 300 (Lord Lane) (adopting a similar test).

[78] *Compare* Larry Alexander, *The Means Principle*, in Legal, Moral, and Metaphysical Truths: The Philosophy of Michael Moore 28 (K. K. Ferzan and S.J. Morse, eds. 2014) (suggesting that motives do not matter to justification cases); Larry Alexander and Kimberly Ferzan, Crime and Culpability 60–61 (2009) (defending a similar view); Victor Tadros, The Ends of Harm 156 (2011) (suggesting motives do matter to justification cases).

excuse is implicated, we can compare: i) the weights attached by PC (the perfectly law-abiding citizen) to the reasons bearing on how to act *that were available to the actual agent to reason from at the time* with ii) the weights attached to these reasons by OC (the otherwise law-abiding citizen). The culpability of the action equals the degree of error in OC's weighting of the epistemically available reasons relative to PC's weighting thereof.[79]

Finally, to account for culpability in cases of wrongful action where a putative rationale-based excuse (as opposed to a justification or exemption) is also in play, we can look to the Amended Manifestation Account. This would tell us that, to assess culpability, we should compare (i) the weights we can in sympathy expect to be attached by RF (the person of reasonable, or we might say minimally tolerable, firmness) to the reasons that are epistemically available for the actual agent to reason from at the time with (ii) the weights attached to these same reasons by OC. The culpability of the action (if any) equals the degree of error in OC's valuation of these epistemically available reasons relative to RF's valuation thereof, in the way the Amended Manifestation Account advises.[80]

In this way, a reasons-based framework can be used to account for (a) the wrongness of actions, (b) their culpability when justifications might be in play but not excuses and (c) their culpability when an excuse is in play. This reasons-based account helps crystallize the distinct roles of wrongness and culpability, as well as show how the culpability of an action differs in cases of unjustified wrongs where no excuse is relevant and cases of wrongs that are unjustified but have an excuse.

[79] Note that focusing on the divergence between RF's and OC's valuations here, as does the Amended Manifestation Account, yields the same results in such cases. After all, outside excuse cases, RF and PC attach the same weights to the relevant reasons.

[80] Note that if one thinks there are incapacity defenses that function as genuine excuses, not exemptions or the denial of an offense element, the present account can be extended to capture them as well. Suppose, for example, one thinks a given defendant in the grips of a powerful conspiracy theory meets the legal definition of insanity but has sufficient agential capacities to remain eligible for judgments of blame and counts as having committed an action that is attributable to them. If one thinks the actor merits a culpability-reduction in virtue of their incapacity, one could apply the same approach as to rationale-based excuses. This would ask: To what extent can we fairly and humanely expect the actor, given her impairment (which let's suppose we can have sympathy for or at least pity), to take account of and be motivated by the applicable reasons against the wrong in question? The impaired actor's culpability would be pegged to the weighting of reasons that would be applied by an agent with the actor's impairments assuming we can have sympathy with these. The action's culpability would correspond to the degree to which it is disfavored by the applicable reasons as we can at a minimum expect them to be weighed by such an impaired actor. More work is needed to flesh out this picture, however – in particular, to cash out how we can expect reasons to be weighed by partially impaired actors. For more, *see* Sarch, *supra* note 58.

Authors and Affiliations

Alex Sarch[1]

✉ Alex Sarch
a.sarch@surrey.ac.uk

[1] University of Surrey School of Law, Guildford, England