DOI: 10.1111/phpr.12705

ORIGINAL ARTICLE

Philosophy and Phenomenological Research

The Moral Grounds of Reasonably Mistaken Self-Defense

Renée Jorgensen Bolinger* Assistant Professor of Politics, Human Values

Princeton University, Princeton, United States

Correspondence
Renée Jorgensen Bolinger, Princeton University.
Email: bolinger@princeton.edu

Abstract

Some, but not all, of the mistakes a person makes when acting in apparently necessary self-defense are *reasonable*: we take them not to violate the rights of the apparent aggressor. I argue that this is explained by duties grounded in agents' entitlements to a fair distribution of the risk of suffering unjust harm. I suggest that the content of these duties is filled in by a social signaling norm, and offer some moral constraints on the form such a norm can take.

1 | INTRODUCTION

On a commonly held view, it is permissible to harm someone in self-defense when and because doing so is necessary to avoid unjustly suffering a similar or worse harm that they will otherwise responsibly impose on us.[1] Suppose this is true. Ordinary decisions about whether to defend oneself must be made in ignorance of many of these facts: one can rarely be certain whether an apparently lethal threat is genuine (rather than a bluff or misleading impression), whether the threatener is morally responsible, whether the contemplated defensive action is actually proportionate to the threat, etc. Given the

[1]Abstracting away from nuances, this is the core thought behind most prominent moral justifications of self-defense, including the causal accounts defended by Thomson (1991) and Uniacke (1994), the Responsibility account as defended by Otsuka (1994), McMahan (2009), and Gordon-Solmon (2017), the Moral Status account defended by Quong (2012), and is central to the more culpability-driven accounts offered by Alexander (1993), Rodin (2003), and Ferzan (2005), as well as many others.

*My thanks to Christian Barry, John Hawthorne, Robin Jeshion, Richard John, Seth Lazar, Mark Schroeder, Seana Shiffrin, and especially Jonathan Quong for comments that greatly improved this paper. Thanks also to Bob Goodin, Robert Gooding-Williams, Kerah Gordon-Solmon, Alan Hájek, Philip Pettit, Katie Steele, Laura Valentini, Mark van Roojen, Brian Weatherson, and audiences at the 2015 Harvard Political Theory Conference, 2016 Manchester Center for Ethics and Political Theory panel on Risk and Uncertainty for valuable discussion. This paper was also presented in the regular colloquium series in philosophy at Australian National University (2018), and the University of Queensland (2019).

pressures and ignorance defenders must act under, they are bound to make some mistakes, failing to defend themselves when necessary or imposing unnecessary defensive harm. Focusing on one type of mistake—imposition of harm on agents who actually pose no threat—reveals an asymmetry in the moral evaluation of defensive errors. Some, but not all, of the mistakes a defender makes while acting in accord with her best evidence are *reasonable*. To illustrate, contrast two cases:

- *Stalker* – Dahlia is walking home alone one night when she notices an imposing man (Alex) following behind her in the shadows. Dahlia stops, and so does Alex. Unnerved, Dahlia grips the tazer she carries for self-defense, and starts walking again, faster. Alex keeps following, and when Dahlia turns down an alley Alex rushes at her. Dahlia, believing she will be severely assaulted otherwise, tazes Alex once he is only an arm's length away. In fact, Alex only intended to scare Dahlia without harming her.
- *Testimony* – Dignitary has been told by an extraordinarily reliable (but not infallible) source, whom she trusts, that Alishan is a highly skilled assassin intending to kill her. Dignitary knows that if this is true, her only chance of surviving is to disable Alishan before he begins to attack. So when Alishan approaches to shake hands, Dignitary tazes him. In fact, the reliable source was wrong, and Alishan posed no threat to Dignitary.

The widely shared intuitive reaction to cases like these is that while Alex is not wronged by the defensive mistake in *Stalker*, Alishan *is* seriously wronged by Dignitary's blameless defensive mistake in *Testimony*. The explanation usually given for this verdict in cases like *Testimony* runs as follows: the intentional imposition of harm wrongs someone if they have not made themselves liable to it. Alishan has done nothing to incur liability, so the Dignitary's defensive mistake wrongs him. Indeed many authors go further, noting that even though Dignitary was appropriately attentive to evidence, and justified in believing that her action was necessary and proportionate defense against an unjust threat, she knew (or should have) that there is always *some* risk of being mistaken. So in defending herself, she responsibly engaged in a foreseeably risky action, and unluckily that risk eventuated. Though blameless for the wrong she does, Dignitary thus makes herself liable to be harmed in Alishan's defense, and owes him an apology and compensation.[2]

The hard question is why we don't think the same is true of *Stalker*. Alex's actions don't fit the normal explanation of how a person makes themselves liable to harm. On standard accounts, an agent is only liable to harm that is both proportionate to and necessary for the prevention of the threat for which he is responsible; he retains rights against, and would be wronged by, harm that does not meet these conditions. Dahlia's mistake violates both, imposing harm that is in fact unnecessary, and out of proportion to what Alex has actually *done*. And yet, assuming that the mistaken defender did all they reasonably could to try to determine whether the threat was genuine, there is a strong intuition that in cases like this, the apparent aggressor does not have grounds to complain that his rights were violated. The defender has neither failed to respect his moral status, nor wronged him, nor does she owe

[2]The verdict that apparent aggressors are not wronged in cases structurally like *Stalker* is endorsed in many places, including discussions of the *Joker* case in McMahan (2005), Lazar (2009, 724), Frowe (2010), and the *Unloaded Gun/Bluff* case in Ferzan (2012, 690), Ferzan (2017), McMahan (2011) and Quong (2009), among others. For a brief sampling of theorists who endorse the verdict that apparent aggressors *are* wronged in cases structurally like *Testimony*, and explicitly appeal to the unlucky eventuation of responsibly-taken risk to justify it, see discussions of the *Dignitary* case in Otsuka (1994, 91), Otsuka (2016, 63–64); the neighbor in Draper (1993, 84); police officer in Draper (2009, 74) and the *Mistaken Resident* in McMahan (2005, 387), McMahan (2009, 162–7), and *Police Intervention* in Quong (2012, 60–61, 67).

compensation. In other words, the mistake is *reasonable*.[3] To secure this verdict, many append a second clause to their core account of liability; for instance, Quong (2015, 263 fn 10) extends his core account in a footnote to cover this case, adding "Or by being responsible for causing others to reasonably believe one is going to violate someone's rights." McMahan (2011, 555–56) similarly extends his account by simply stipulating that "when a person bears full responsibility for appearing to pose a threat of wrongful harm, the apparent victim does not make himself liable to counterattack by attacking in apparent self-defense." Quong and McMahan (and others) make these extensions in passing; the moral grounds for reasonably mistaken defense is not their primary focus, and so gets little explication. But it would be good to have theory to explain *why* addenda like these are appropriate; why does merely causing a false belief have so severe an effect as to forfeit stringent rights against suffering serious physical harms?

It is important to have an articulated account rather than relying on our intuitions to sort the reasonable mistakes from the others, because we might hope moral theory will have something to say about real cases of mistaken self-defense, which are rarely as clear-cut as *Stalker* and *Testimony*. For example:

- Parking Garage – Dodge is walking back to her car in a parking garage alone one evening when she notices a man (Aston) following behind her in the shadows. Dodge knows that assaults by strangers are statistically most likely to be perpetrated by males against women in places like this parking garage. Unnerved, she grips the tazer she carries for self-defense and walks faster. Aston keeps following all the way to the corner where her car is parked, and when she stops to find her keys, his pace quickens directly toward her. Dodge, believing she will be assaulted otherwise, tazes Aston once he is only an arm's length away. In fact, he had no intention of attacking her and was only taking the fastest route to his car, which was parked just past hers.
- *ATM* – It's raining hard the night that Dollars withdraws cash from an ATM in a motel lobby. She knows she's in a neighborhood where armed robberies and muggings are statistically most likely to be perpetrated at ATMs and by black males. As she makes her withdrawal, Dollars notices a black man (Avon) quickly crossing the street toward the ATM. Unnerved, she pulls a tazer from her purse. As Avon rushes out of the rain and into the cramped lobby, he reaches into a coat pocket and begins to pull out what looks like a handgun. Fearing that he means to shoot and rob her, and lacking time to say anything, Dollars tazes him. The object was actually his wallet.[4]

The project of this paper is to build an account to illuminate not just the first two cases, but these more realistic ones as well. While most theorists writing on self-defense pass over reasonable mistakes, some (particularly Ferzan (2017) and Bolinger (2017a)) have offered a more sustained treatment. I will note the differences between the account I ultimately offer and Ferzan's proposal in §5. While I engaged with the question of which mistakes we should treat as reasonable in Bolinger (2017a), I there set aside moral underpinnings in order to focus on policy recommendations. I only gestured vaguely at how the moral story might go, viz: "society can treat some behaviors as marked 'signals' of aggression, licensing agents to assume (absent countervailing evidence) that the performer

[3]Dahlia has of course acted suboptimally; it would be better if her mistake had not occurred. She should feel regret on discovering her mistake—perhaps she should even feel prompted to apologize or explain—but this is a poor test of whether an action violates rights; one should feel sorry to have done unnecessary harm, even permissibly. We can say that in the evaluative sense of 'ought', her mistake ought not have happened, but it does not follow that it was objectively impermissible, let alone that it wrongs Alex.

[4]Based on a hypothetical case given by Armour (1997, 1–2).

is an aggressor.[…] agents who perform these behaviors have no complaint against mistaken defenders, since they induced the error (whether intentionally or not) through their own easily avoidable behavior." The project in this paper is to articulate an account that could work as a moral foundation for this sort of view.

I'll here develop a *how-possibly* justification: I will start by assuming that rights are fact-relative, protecting agents' interests against suffering certain outcomes, and show how we could reach the conclusion that reasonable mistakes do not violate rights, and are thus permissible rather than merely excused. Here's the gist: the distribution of harms and the risk of harms that arises naturally from agents living in close proximity, with reasonable fear of aggression, is unjust. Members of a society have a fairness-based complaint against leaving the distribution unaddressed. So, justice demands the institution of a (formal or informal) social norm to adjust the distribution of these harms and risks. Agents who violate the norm breach their duty to cooperate with the collective endeavor to more justly distribute the costs, and cannot reasonably demand that others accept disproportionate risk to accommodate them. Consequently, they lack a complaint against other agents acting as prescribed by the norm by imposing apparently necessary and proportionate defensive harm. Thus, reasonable mistakes are permissible as the byproduct of a justified social practice for fairly managing the risk of suffering unjust harm.

2 | THE DISTRIBUTIVE PROBLEM: HARM AND RISK

Let's start by just sketching the intuitive profile of reasonable mistakes. They occur when the mistaken defender had *good reason*—in some to-be-explicated sense—to believe that she faced a genuine threat. The relevant sense of 'good reason' isn't quite equivalent to justified credence above some relevant threshold, since otherwise *testimony* should be a reasonable mistake if anything is. It instead depends on, as Ripstein (1996, 691) puts it, "whether it is reasonable to ask the [defender] to bear the risk of a perceived attack, or to make the apparent aggressor bear the risk that the [defender] was mistaken."

Under normal circumstances, intentionally imposing unnecessary non-consensual harm on another person violates their rights against such treatment, and an agent who does so without moral justification incurs a liability to be preventatively harmed. If an agent is *liable* to suffer some harm H, he lacks a right against suffering H. However, since moral philosophers sometimes use 'liable' to mean not only this, but that he has forfeited his right and lacks permission to counterdefend against suffering H, I will use 'vulnerable' to indicate the simple absence of a right. Someone who is *vulnerable* might have forfeited their right against the harm, or have waived it, or might never have had such a right. Similarly, they might still be permitted to counter-defend, or they might not. The mere fact of their vulnerability does not entail anything except that H does not violate their rights.

Every member in a community has a strong interest in bodily integrity and autonomy, which grounds a permission to 'enforce' this right against violations by using proportionate force when necessary to avert unjust threats. The standard explanations of the permissibility of defensively harming genuine threats leverage these two facts: a defender (D) does not wrong an aggressor (A) in enforcing her right, because A has forfeited his right against (becoming liable to) proportionate defensive harm by threatening D. But if defense is *only* permissible against culpable aggressors or genuine threats, agents face an epistemic problem: merely apparent threats are systematically indistinguishable from genuine ones, and to the extent that culpability depends on the mental states or intentions of the individual, one cannot confirm it by observable behavior alone. This opacity leaves agents uncertain,

when facing a probable threat, whether they can permissibly defend themselves. They face a high-stakes gamble with non-trivial risks of error.

There are two kinds of mistake a defender in this position might make. If she fails to defend herself against a genuine threat, she makes a *false-negative* error. If she defensively harms someone who wasn't a genuine threat, she makes a *false-positive* error. All agents have incentives to minimize their chances of making either error, maximizing their odds of using defensive force when and only when actually necessary. But agents vary in their risk-aversion thresholds, and in the disvalue they assign to each error type.

If a community of agents living in close proximity with reasonable fear of suffering aggressive harm are left to make their own judgments about how to balance the risks of each type of error, the resultant distribution of the risk of suffering either aggressive or mistaken harm will almost certainly be unjust. Under these conditions, the less careful an agent is to avoid violating others' rights—the less weight she assigns to false-positive errors—the more 'reactive' she will be, hesitating less in defending herself, and hence the lower her risk of suffering aggressive harm will be. Conversely, the more weight she gives them, the more 'cautious' she will be, and the greater her risk will be. So an agent's risk exposure would vary inversely with her respect for others' rights, except insofar as it depends on her brute luck in how often she encounters reactive agents. Worse, these prospects create perverse incentives for agents to be reactive, increasing community members' absolute risk of suffering rights-infringing (aggressive *or* mistaken) harm. This is far from a just distribution, and is in fact the best-case scenario. In actual societies, reactive defenders are more suspicious of some social groups than others, so rather than being randomly distributed, risk of suffering mistaken defensive harm pools disproportionately on these groups.[5] Both members of such groups and cautious defenders have a legitimate moral complaint against this distribution: their rights against harm are not adequately secured so long as they are at disproportionate, unjustified, and unchosen risk of suffering violations.

There are two distinct arguments we could offer to explain how the moral imperatives to minimize harm and redress the unjust distribution ultimately motivate taking reasonable mistakes to not be wrongings. The first invokes individuals' obligations to act jointly to remedy wrongs arising from their aggregate actions. Since the agents who face disproportionate risk do so as a direct result of the other community members' behavior, they have a moral claim against those others to take steps to counteract or compensate for the risks imposed. If the other members can act to redistribute the risk more fairly, they owe it to the disproportionately exposed to do so, even if this means accepting some risk of suffering mistaken harm themselves. If they cannot fulfill this duty acting alone, but could by acting jointly, then they are obligated to cooperate to address the injustice. Because they have a moral duty to take these steps, they do not have a right against the costs of fulfilling the duty; they are vulnerable to bear those costs.

The second argument appeals to the obligation of the state to secure the rights of its citizens. If a group of individuals are at disproportionate, unjustified, and unchosen risk of suffering violations of their rights against harm, their rights have not been adequately secured. To fulfill its obligations to members of the high-suspicion group, the state must act to counteract the perverse incentives to be reactive, and correct the worst distributive injustices arising from agents' acting on their own judgment. So, just as it must for agents' property rights, to fulfill its obligations to secure its citizens' rights, the state must adjudicate agents' rights against harm, issuing authoritative judgments to set their contours and determine what constitutes a wronging.

[5]For discussion, see especially Richardson and Goff (2012).

Either of these justifications is a solid foundation for taking agents to generally be vulnerable to reasonable mistakes, *if* we can offer a clear articulation of how fulfilling the duty entails treating a particular class of mistakes as not wrongings. If there is no way to fulfill the duty, then these arguments instead imply that compensation is owed to the disproportionately exposed. What would it take achieve a more just distribution? To minimize the secondary source of risk, and to distribute the remaining risk more equitably, reactive defenders must be made to act more cautiously, particularly with respect to members of the high-suspicion group. Cautious defenders meanwhile must act more decisively, which they will only be willing to do if we can reduce the odds that they will thereby make a false-positive defensive mistake. Acting alone, no individual can bring about these changes. However, the shape of error propensities *can* be influenced by a change in the social norms that govern defensive decisions, and these norms are responsive to the ways that costs are distributed by decisions in the criminal law. So, the rest of this paper will focus on articulating what sort of norm could fulfill the duty, and what it entails about the profile of reasonable mistakes.

A social norm can affect agents' defensive error propensities by altering either the costs of mistake, or the strength of evidence available to defenders, or both. A norm that permits some mistakes—'the reasonable' ones—and holds agents responsible for others, potentially alters both, and can go a long way toward eliciting a more just distribution. To succeed, it must reduce the space of unguided defensive gambles by putting agents in a position to usually know whether the norm permits defense in their actual circumstances. It cannot require sensitivity to facts that are in principle inaccessible to agents at the time of action. Call this *transparency*. To get cautious defenders to act more decisively, it must also help reduce the odds of a false-positive mistake, which it can do if it helps well-intentioned agents to anticipate, and thus avoid, actions which would make them misleadingly appear to be aggressors. Call this *predictability*. Finally, it must of course distribute the risk of suffering aggressive and mistaken harms among members of the community *fairly*, ideally in a choice-sensitive way.

3 | WHY RATIONAL HIGH CREDENCE ISN'T REASONABLE

So much for the general shape. What should we take the content of the norm to be? To get the requisite transparency for defenders, the norm must be sensitive to facts they are usually in a position to know, or at least can track. We might consider something like

> EVIDENCE: A is vulnerable to defensive harm from D when it is reasonable to believe, given D's evidence, that such harm is necessary and proportionate to avert an unjust threat posed by A.

Much depends on how we construe 'D's evidence'. We could invoke familiar idealizing assumptions — stipulating that D must have responsibly gathered evidence, considered all possibilities, conditionalized without error, etc. But there are two immediate problems with this: first, since there is no guarantee that the available evidence accurately tracks the moral facts, this norm does not tie vulnerabilities to A's actions in a fair and predictable way. Consider Dignitary's situation in *Testimony*: given the testimonial evidence from a reliable source, if Dignitary perfectly considers her evidence and conditionalizes on it, she should believe it is highly likely that Alishan is an unjust aggressor. But there is no way for Alishan to avoid this; it is thoroughly unpredictable, and seems the wrong sort of reason to undermine Alishan's complaint against being preemptively harmed by Dignitary.

Second, idealizations render the evidential norm opaque to defenders. We are not cognitively frictionless ideal Bayesian beings, and what evidence we have "is not guided by what [we] are able to compute, but what [we] happen to see at a given moment" (Kahneman 2003, 1469). Rather than entertaining all possibilities and conditionalizing on our evidence across a full partition of logical space, we do and must drastically simplify things. Especially when operating under uncertainty, we must rely on "fast and frugal heuristics" (Gigerenzer and Goldstein 1996),[6] carving up possibility space into a relatively course-grained partition that allows us to take some things as given. We exhibit these behaviors not out of laziness, but because it is too cognitively costly to attempt to operate with Bayesian calculations (or even close approximations). Given the differences between their own reasoning and the ideal, defenders will generally be unable to tell whether an idealized evidential norm permits them to defend themselves in their actual circumstances. Transparency is only achieved if the norm is tractable for agents like us; it cannot require cognitive feats we struggle to perform even under perfect conditions.

But accommodating our cognitive limitations by interpreting EVIDENCE as a simple subjective principle about A's *actual* evidence is also unsatisfactory. It will not allocate risk in a just or predictable way: it will simply count A vulnerable whenever, *given the heuristic that D uses* and the evidence she in fact has, it is likely on D's evidence that A poses a threat. This is a problem even if we restrict what counts as evidence to only D's true beliefs or knowledge. Recall that in *ATM*, Dollars knows that the probability that Avon is a mugger is higher given that he is a black man in this neighborhood. She's also probably nervous, and thus more likely to notice and attend to facts that support the belief that Avon is a mugger than facts suggesting that he just needs to use the ATM. Faced with ambiguous data (like his reaching into a coat pocket), she is likely to interpret it as evidence that he is a genuine threat. Dollars is in this way like many Americans, who (whether consciously or not) when tasked with discerning agents holding threatening objects from those with non-threatening objects, make errors that strongly track the race of the pictured individual.[7] All told, Dollars's belief in *ATM* that Avon is an aggressor may well be supported by her actual evidence. But even if we could be convinced that Dollars does not wrong Avon, these cannot be the reasons justifying such a verdict. Restricting what counts as evidence to just the agent's *true* beliefs won't insulate against biased or haphazard collection of evidence. This sort of standard won't distribute the risks of harm fairly; it instead allows them to concentrate on members of a group disproportionately perceived as threatening.

Summarizing what we can learn from these failures: to reduce false-negative errors a norm must be tractable for actual agents, so it must allow the use of heuristics and proxies. But to distribute risk of suffering false-positive errors fairly, the norm must coordinate which heuristics are used, rule out bad ones, and must also make vulnerability choice-sensitive if possible. This confirms what we already suspected: the notion of 'good reason' invoked by the reasonableness standard can't be rational high credence. But an account of reasonable mistakes must say more than this. For an illuminating account, we'll need (i) an explanation of what makes reasons good enough, if not the role they play in grounding D's rational credences, and (ii) a more general characterization of this class of reasons that is projectable to new cases.

In intuitive cases of reasonably mistaken defense, the thing playing the role of 'good reason' is confidence resulting from having observed A do something that, in a loose sense, conventionally

[6]As the term suggests, a fast and frugal heuristic is a reasoning procedure optimized for minimal cognitive cost. It aims for high enough approximate accuracy based in as little evidence (and hence costly processing) as possible.

[7]Subjects in several studies (see Correll et al. 2002, 2007, 2011; Kahn and Davies 2010; Sim, Correll, and Sadler 2013) consistently had higher false-positive errors when the pictured individual was black, and highest false-negatives when the pictured individual was white. For more thorough discussion of this point and the challenges it raises for legal determinations of 'reasonable mistakes', see Richardson and Goff (2012).

communicates an aggressive threat. These are the sorts of behaviors that you know—at least on reflection—to avoid, because they tend to make people think you mean to harm them: following a solitary woman at night, aiming a weapon at someone, wearing a ski mask to a bank, etc. The *Assumptive Signals Account* that I construct in §4 analyzes 'good reasons' as signals of aggression which permit agents to assume that the performer is liable to defensive harm, *if* the signals themselves satisfy two stringent moral constraints which I outline and discuss in §5.

4 | SIGNALING AGGRESSION

Tying reasonableness to signals of aggression has the effect of coordinating and stabilizing acceptable heuristics for defensive decisions. In effect, it marks some types of evidence (the assumptive signals) as *privileged*, encouraging defenders to assume that defensive force is permitted when they have evidence of this sort, and to be cautious otherwise. To ensure that vulnerabilities to defensive harm are appropriately choice-sensitive, we must constrain the type of evidence that can be privileged: well-intentioned agents must be able to go about their lives without acting in a way that gives others this evidence. We can represent the general form of this proposal as

> ASSUMPTIVE SIGNALS: D is permitted to impose defensive harm on A if either (i) A ψ s, or (ii) A has acted in a way that gives D evidence e that A ψ s, D has made a reasonable effort to get better evidence, and D lacks undermining evidence.

The first disjunct of this norm is given content by our best theory of the conditions for permissible defense against genuine threats (e.g. we may replace ψ with *incurs a liability by posing an unjust threat*); the second disjunct defines reasonable mistakes, and e is given content by a socially-defined *set of assumptive signals*. Under this norm, mistaken defenders are insulated when they acted on an assumptive signal and lacked undermining evidence, but are wholly responsible for mistakes based on non-privileged evidence. As I have stated it, the norm requires uptake: e licenses D to assume that A is liable only if it is actually part of D's evidence.[8]

Now that we've characterized the role assumptive signals must play, let's get more precise about what they *are*, exactly. A signal is an observable state that carries information about some other state of affairs. Signaling relations can be wholly natural (the way smoke signals fire) or conventional (how nodding signals agreement) or anything in between. Signals become associated with particular contents by the combined effects of reinforcing effective signals and forgetting unsuccessful ones. Assumptive signals are modeled on evolutionary signals, which can be sent by simple organisms, and do not depend on communicative intent for content.[10] They are conventional in that they do not ensure their contents: it is possible to send a signal b without the signaled state X obtaining. But if agents incur costs from relying on false signals, this can only happen so many times before it destabilizes: each occurrence of b that does not co-occur with X introduces signal noise, and the more often that occurs, the weaker the connection

[8]I favor requiring D to be aware of A's ψ ing, permitting her to inflict defensive harm on A only *in response to* A's ψ ing, but it is possible to modify the account to relax this condition, counting D as licensed just if A acts in a way that makes e available, whether or not D is ever aware of it.

[10]I am relying on Skyrms (2010)'s model of communicative signals to provide the scaffolding for this account, rather than Lewisian communicative conventions (Lewis 1969), because the Skyrmsian model is more flexible in many ways, most importantly that signals need neither be arbitrary nor sent intentionally.

between *b* and X becomes, until eventually *b* carries no information about X and ceases to be a signal.[11] I will, in what follows, speak of 'signaling behaviors', but this is slightly metaphorical: one can send a signal by doing something as simple as *instantiating* a conventionally significant property. If you wear a red polo shirt and khakis while wandering through Target, you signal that you are a member of the staff (and may well be approached for help). Signals like this can be canceled when accompanied by sufficient undermining evidence; for instance if you are also carrying shopping bags from several other stores, the 'member of the staff' signal will be drowned out by the much stronger 'household shopper' signal.

As implemented for defensive permissions, candidates for members of the set of 'assumptive signals' are behaviors that are socially marked as correlating strongly with genuine threats, and the relevant content is *that the performer (A) is liable to apparently necessary and proportionate defensive harm*. If this proposition is true, then all else equal D is permitted to impose defensive harm on A. So, when D observes A perform a marked signaling behavior *b*, and D lacks undermining evidence, she has strong evidence that she is permitted to defend herself.

Since genuine aggressors cannot be relied on to cooperatively signal their liability, marked signaling behaviors will have to be ones that aggressors have sufficient independent reason to perform, that are easy to recognize, and that occur early enough to give defenders a better chance than simply waiting to be sure whether the threat is genuine. A behavior performed *only* by aggressors would be the strongest signal, indicating with certainty that the performer was an aggressor. But signal strength has to be balanced against efficiency, which is itself a trade-off between the difficulty of recognizing the behavior and the range of cases for which a receiver can rely on a signal. The more fine-grained the signals, the more signals agents must encode, hence the less efficient the system. Similarly if signals are difficult to recognize, the cognitive costs involved in determining whether a signal has been sent count against the efficiency of the system. A likely equilibrium will mark easily recognizable behaviors that strongly correlate (or are believed to strongly correlate) with genuine threats, but occasionally occur in their absence, like aiming a weapon at or stalking someone. In non-ideal societies, some bad signals will also appear as candidates, for instance wearing a hoodie, or being Black in a predominately white neighborhood. Not all candidate signals can do the moral work of assumptive signals; we will need stringent selection constraints.

5 | SIGNAL SELECTION CONSTRAINTS

Basing something as significant as a vulnerability to be physically harmed on a possibly unintentional signaling behavior is only permissible when doing so is the fairest way of apportioning costs. This is a more restrictive requirement than it may appear at first. It generates two major constraints on the selection and use of assumptive signals:

1. *Assumptive signals can only be used when agents either cannot get better evidence, or it is unreasonably costly to do so*, given the stakes. Otherwise, A has a strong complaint against relying on the signal: D could at very little cost remove the risk entirely. It is therefore

[11]Signals can be individuated in somewhat fine-grained way, partially determined by background context. For instance, in some parts of the American Midwest a particularly obnoxious siren noise on Tuesdays around mid-morning means the tornado alert system has been tested and is working; on any other day or time, it means there's been a tornado sighting. As a result, the occurrence of a siren noise on Tuesdays (without a sighting) does not undermine it as a signal for tornado sightings when it occurs at other times; the two event types are different signals, and carry different contents.

unreasonable to ask A to bear the risk simply to save D the small hassle of gathering easily accessible evidence.[12]

2. *The behaviors that license the defender to assume that A is liable must be reasonably avoidable*. This constraint is necessary both in order for it to be intelligible to hold agents responsible for their signaling behaviors, and in order to ensure that the norm distributes costs fairly. If we suppose instead that D's permission stems from a behavior that a well-intentioned A could not avoid, or could but only at unfair cost, it would no longer be true that the norm fairly distributes risk of harm. Reasonable avoidability is a normatively laden notion, rather than a simple measure of difficulty; a signal behavior type *b* is reasonably avoidable only if it satisfies two conditions:
   - PUBLICITY: It is public knowledge in the community that performance of *b* makes the performer vulnerable to defensive harm.
   - BURDENS: Performing *b* is something that every member of the community can typically avoid without undue costs.[13]

If a signal fails either condition, using it as an assumptive signal would not distribute the risk of harm fairly, and so would not be a means of fulfilling the duty to cooperate to fairly redistribute the risk.

The fair costs of avoidance for a signal roughly track what sacrifices one can reasonably be demanded to make for the sake of the coordination gains secured by the norm. Plausibly, the threshold of costliness that constitutes 'undue costs' should be sensitive to background facts about the history of the society; members of groups who came to enjoy positions of relative privilege through imposing unjust costs on other groups may be asked to bear greater costs of avoidance, all else equal, than those in the groups that have been historically unjustly disadvantaged. As a result, the constraint against undue costs of avoidance may take a different form as introduced in a stipulatively just society than as introduced in an actual society with a substantively unjust history.

Still, we can identify an upper bound on costs that we can reasonably demand agents to bear. The ability to select for oneself which goals to pursue and what to value is at the core of valuable exercise of agency; thus we cannot reasonably demand that agents give up a normatively permissible project in which they are already invested, which is central to their self-understanding.[14] Candidate signals tied to permissible sub-cultural expressions fall afoul this constraint, because to avoid them an agent must give up a cultural identity that is not in itself objectionable, which they are entitled to retain. For instance, even if it is not difficult to always wear a suit rather than jeans and a hoodie, or to speak with the diction and slang of General American English rather than Black English Vernacular, doing so involves sacrificing participation in a common Black working-class

[12]It's worth noting, though, that the stronger evidence may just be a significantly more reliable signal. Signal strength is a function of how probable the content is, given the presence of the signal. Suppose that *b* is a moderately strong signal, but has a non-negligible chance of being sent unintentionally and without X obtaining. If there is an alternative signaling behavior *y* that is a *very* strong signal of X (it's *possible*, but *highly* unlikely that *y* would be sent unintentionally and without X obtaining), then *y*'s easy availability precludes permissible reliance on *b*.

[13]It is possible for the costs of avoidance to be uniformly high but not undue because they are fairly distributed among the members of the community. There is a natural ceiling to how high even such symmetric costs can rise, compared to the benefits secured by acting in accordance with the norm, without being undue. If the costs of signal avoidance are so great as to undermine the value of greater security against suffering aggressive harm, then agents cannot be reasonably demanded to accept the costs, and they fail the BURDENS constraint.

[14]We can understand 'normatively permissible' here as a relatively thin requirement that the project not presuppose violating or threatening to violate other agents' rights. This restriction ensures that we may discount the costs of giving up a way of life incompatible with respecting the rights of others (e.g. membership in the Ku Klux Klan).

culture. While it is permissible for an agent to freely choose to do so, such a sacrifice cannot be demanded of them.

You may find the restriction of the BURDENS constraint to *typical* costs of avoidance unappealing: if we want vulnerability to track culpability, surely the actual costs for an agent of avoiding a *token* signaling behavior are more relevant. Let me start by granting that it's possible for it to be quite costly on a specific occasion for A to avoid performing *b*, but for a reason that does not undercut the signal. This could happen if A is coerced into *b* ing by Villain, who threatens to kill A's family otherwise. My account implies that in performing *b* (and thus signaling that he is an aggressive threat to an innocent D), A incurs at least partial responsibility for the costs if D reacts by making a defensive error. A acted in a way that is *typically* easy for him to avoid, and prior to this particular event, could not have reasonably complained that attaching vulnerability to *b* ing was too demanding or unfair. Of course, D may not impose defensive harm if she *knows* that A is *b* ing only because he is coerced, and poses no genuine threat; such knowledge would be undermining evidence.

Though A is terribly unlucky, his predicament does not show that BURDENS should track token rather than *typical* costs. The key here is that the Assumptive Signaling Account does not purport to tie vulnerabilities to culpability; it aims only to fairly distribute the risk of suffering harm.[15] The account in fact departs from culpability in a second way, as it does not require that A *intend* to send a signal for his signaling behavior to ground a vulnerability. Both departures from culpability are justified by the distributive role of assumptive signals. To fulfill their coordinating function, signals need to be individuated exclusively by observable features; receivers must be able to generally tell whether they've received a license to act defensively. Agents cannot track facts about others' internal mental states or intentions, and so cannot use signals that essentially reference such things. Similarly, setting the costs of avoidance to tokens rather than types would render signals fine-grained in a way that leaves agents unable to recognize license-granting signals. This would make it impossible for the signals to play a communicative or coordinating function, and thus render them ineffective in fairly distributing risk.

That said, the token-level facts about A's situation or intentions still play an important role on my account. While they do not alter D's permissions, they *do* make a difference to A's culpability for violating the obligation to avoid creating the appearance of threat, and so make a difference to A's counterdefensive permissions. Assumptive signals are in this way parallel to slurs: agents are morally answerable for their slurring speech acts, even if they didn't intend to cause offense, demean, or

[15]In that I take agents who perform conventionally threatening behaviors to lack rights against defensive harm, even if the threat is merely *apparent*, the account I've developed is similar to Kimberly Ferzan (2017)'s 'forfeiture by insincere performance' proposal. On her account, agents who culpably appear to be aggressors incur a liability to suffer apparently necessary defensive harm. The two views agree in a range of cases, but diverge in motivation, scope, and the severity of moral costs they assign to misleading performances. While I justify vulnerabilities as a consequence of the fair allocation of risk, Ferzan motivates forfeiture as a way to prevent culpably misleading agents from getting away with 'normative land-grabs.' My account counts negligent performers among the vulnerable, while for Ferzan culpability requires at least *intentional* awareness and dismissal of the relevant risks, and at most extends to reckless agents. I also allow that vulnerable agents may retain defensive permissions, while on a standard interpretation, liable agents are prohibited to counter-defend.

belittle, though these facts do affect the speaker's blameworthiness.[16] Admittedly, the costs incurred for ignorant or unintentional slurring are significantly lower than those incurred by appearing threatening; one might happily hold agents accountable for the former while wincing at the latter. Still, these more substantial costs are justified by the greater value of the social practice preserved. The increased predictability and security gained through the signaling norm cannot be retained if we attempt to finesse the permissions to match individuals' unpredictable ignorance or undetectable difficulty in avoidance.

There is still a significant difference between the way inadvertent signaling, on the one hand, and the reasons Dignitary has for believing Alishan is a threat in *Testimony*, on the other, relate to an individual's exercise of agency. Even if not intending *to signal*, in acting as he has in *Parking Garage* Aston intentionally performs a behavior that signals aggressive threat. His action is an exercise of his responsible agency, in a context where it is reasonable for others to assume that he is aware of the signaling conventions, and to demand that he exercise caution. In *Testimony*, by contrast, the reasons for Dignitary's belief aren't connected to Alishan's decisions at all. Even if we modify the case, having the informant tell Dignitary a conditional: '*If* Alishan approaches to shake hands, you must defend yourself or die', there is a deep difference between Alishan's responsibility for shaking hands and Aston's responsibility for following Dodge through the garage. The key to this difference is the avoidability and publicity of the signal: the fact that it is reasonable in general for members of the community to expect each other to know about and avoid a given signal is what makes it permissible to hold each other responsible for it. The publicity and avoidability conditions make it possible to strike a balance between tying vulnerabilities to A's responsible exercise of agency, on the one hand, and behaviors that are within D's epistemic grasp, on the other.

6 | WHITHER REASONABLENESS?

On the final account, a mistake is reasonable when:

- D acts in response to A's performance of an assumptive signal of aggression *b*, D has made a reasonable effort to get better evidence, D lacks undermining evidence, *and*
- *b* satisfies PUBLICITY (it is public knowledge in the community that *b* is an assumptive signal of aggression) and BURDENS (performing *b* is something every member can typically avoid without undue costs), *and*
- The harm D imposes on A would be permissible had the signaled threat been genuine: it would satisfy the necessity and proportionality constraints, and would not be unjust.

[16]See discussions in Camp (2013); Bolinger (2017b); Jeshion (2013); Croom (2011). The signaling convention I've proposed has another parallel with the accountability practices seen in slurs: agents are accountable to the conventions and practices of the signaling community they happen to be in, *even if* the agent is unaware of those conventions, *so long as it is reasonable for the rest of the community to think that the agent is aware*. It is in general a foreigner's obligation to acquaint herself with the local conventions, even though doing so takes more effort for her than for an average local. She can insulate herself from accidentally incurring vulnerabilities only by making it clear that she is ignorant. A foreigner who uses a potent slur while mistakenly believing it to be inoffensive still licenses offense: unless his ignorance is clear to hearers, his use of the term signals that he endorses the objectionable attitudes associated with the slur and makes the speaker vulnerable to censure. Likewise, if Alexis, an agent from an isolated community completely unfamiliar with guns, travels to Boston and points a gun at Daniel in a way that any reasonable Bostonian in Daniel's shoes would interpret as an act of deadly aggression, Alexis incurs a vulnerability—unless it is clear to Daniel that Alexis does not understand the conventions and does not pose a genuine threat. If Daniel can discern *that*, then he has undermining evidence and is not permitted to impose defensive harm on Alexis. If he lacks it, Daniel is in a position structurally similar to Dahlia, and Alexis's responsible agency put him in that position.

This requires judgments about (1) what is a *reasonable effort* to get better evidence, (2) when a signal is sufficiently public for it to be *reasonable to expect* A to be aware of it, and (3) whether a cost is undue, such that we cannot *reasonably demand* that agents make the sacrifice for the sake of the signaling relation. Are these judgments any more tractable, or less susceptible to bias, than directly intuitively judging whether a mistake is reasonable?

I contend that they are. What counts as a reasonable effort to get evidence can be analyzed in terms of the value of information, which admits of rigorous theorizing.[17] It is harder to say exactly when a signal is sufficiently public to make it reasonable to expect that community members are aware of it, but parallel questions are a core concern for public law, and progress made there can be used to illuminate this question.[18] The third question—whether a cost is undue—still requires a balancing of moral interests, but re-casts the problem: rather than having to make highly individualized judgments between the interests of particular pairs of mistaken defenders and apparent aggressors, we can seek general principles about the forms of sacrifices which can and cannot be demanded. So, though my account still requires judgments about reasonableness, it is a significant improvement over the hazy intuitive notion with which we began, and gives verdicts that are projectable to new cases.

There is a second important way in which this improves over a no-theory theory. Assumptive signals are not just a good way for cognitively bounded agents to manage risk *in theory*. There is good evidence to suggest that, at least in the United States, judgments about the reasonableness of mistaken self-defense already implicitly appeal to accepted signals of aggression.[19] Making the appeal explicit, as the Assumptive Signaling Account does, allows us to insist that the signals satisfy the rigorous moral constraints necessary to justify the moral conclusions they are used to support. It thus yields resources for reforming our actual determinations of reasonableness.

7 | EXPLAINING THE CASES

If the constraints are met, the ASSUMPTIVE SIGNALS norm has the right structure to address the distributive problem. It holds agents responsible to avoid performing marked behaviors that conventionally signal aggression, and licenses them to act defensively when others perform them. Because performing a marked behavior is a public, observable event, D *can* in general know whether she is licensed in a context to defend herself. Because the marked behaviors are known to be marked, agents can predict and easily avoid behaviors which would make them appear to be aggressors. This allows agents to coordinate to reduce the overall risk of defensive mistakes. If an agent performs a marked behavior that satisfies the PUBLICITY and BURDENS conditions, they cannot reasonably demand that others destabilize the norm in order to make a special exemption; they thus lack a complaint against suffering a defensive mistake resulting from their signaling performance.

The norm outlined affects both the relative costs of mistake and the strength of evidence available to deliberating defensive agents: false-positive mistakes made in response to an assumptive signal are low-cost, because the norm insulates the defender, but those made for any other reason are high-cost. And, since well-intentioned agents know to avoid the signaling behaviors, observing

[17]See, e.g., Gersbach (1997). For an argument that the value of increasing awareness can be given a parallel analysis, see Quiggin (2016).

[18]See, for example, discussions of Lon Fuller (1958)'s explication of the requirement that law be public in Luban (2002), Murphy (2005), among others.

[19]For an overview, see Bolinger (2017a).

a signal gives defenders strong evidence that she is facing a genuine threat, reducing her chances of making a false-positive mistake. This encourages cautious defenders to act decisively when they observe a marked signal. Meanwhile, the effects on the relative costs of error encourage reactive defenders to be more cautious about relying on their own suspicions, and so decrease the disproportionate risk faced by members of 'suspicious-looking' groups. So, adopting a norm like Assumptive Signals is a good candidate for how agents in a society should cooperate to address the unjust risks imposed by their collective, uncoordinated defensive activity. What does this imply about our cases?

Stalker is straightforwardly a reasonable mistake: Alex intentionally signaled an aggressive threat, and could easily have avoided doing so. He thus has no complaint against Dahlia's imposition of the defensive harm he caused her to believe was necessary and proportionate. This is not the case in *Testimony*. Though (given her evidence) Dignitary is rational in having a high credence that Alishan is a threat, the reasons for Dignitary's belief are not based on anything Alishan has done or could foresee would create the appearance of a threat. In the absence of an assumptive signal, Alishan retains his rights against being harmed by Dignitary; her mistake is rational, but wrongs him.

So far so good; we have justified the clear intuitive verdicts. But can the account illuminate the murkier, everyday cases? In *Parking Garage*, Aston need not intend to appear threatening. Perhaps he is wholly focused on taking the fastest route to his car, without regard for others' discomfort, and is having trouble remembering where he parked. Nevertheless, he decides to act in a way that foreseeably grounds a rational fear of aggression. Even if Aston thoughtlessly did not notice this, it is reasonable to expect him to be aware that following a solitary woman through a garage is threatening. Alex and Aston are not equally culpable—Alex intentionally caused the fear, while Aston did so merely foreseeably—but both could easily have avoided it altogether. If this is right, then Dodge can legitimately demand that Aston be situationally aware in contexts like these, and avoid behaviors that predictably signal aggression in them. There is a limit to the costs he can be fairly expected to bear: plausibly he cannot be required to avoid parking garages altogether, or to use them only in the company of others. But provided that it would not generally be unduly costly for Aston to avoid following solitary women through parking garages (e.g. by changing his pace or route, or waiting a bit), the Assumptive Signals Account predicts that Dodge's mistake is reasonable.

What of *ATM*? We've stipulated that the statistics in this neighborhood support Dollars' belief that Avon is more likely to be a mugger than any other demographic. Does this fact obligate Avon to go out of his way to avoid appearing threatening, and if so, what does that require him to do? Speaking generally, distributive inequalities drive statistics like this: when muggings disproportionately involve members of a particular social group, it is often because the group is disproportionately economically disadvantaged. In most actual societies, these inequalities are themselves a result of antecedent distributive injustice. Even if the statistics make it rational for individuals to be fearful in a context,[20] using the consequences of overburdening a social group as grounds for demanding that they accept higher burdens of avoidance is suspect at best. Members of an unjustly disadvantaged class are under less, not more, obligation to accept costs to avoid signaling.

Second, it's difficult to identify a signal that could be active in *ATM* that satisfies even an ordinary standard for avoidability. Suppose we construe the signal as being 'a black male in this neighborhood *approaching the ATM*.' It is no easy thing for Avon to avoid appearing threatening, given this setup; he would just have to avoid neighborhood ATMs altogether. The net effect would be to exclude him from a significant range of public spaces others enjoy, which runs afoul of the upper-bound on costs

[20]Though there is reason to doubt that statistics of this kind make it rational to be noticeably more suspicious of specific demographic groups; see Bolinger (2019), Gardiner (2018), and Armour (1997).

that significantly restrict scope for agents' autonomy.[21] Nor can we reasonably demand that Avon avoid appearing Black (either by passing as white or adopting the trappings of 'white culture'). So long as the agent's identity and projects are themselves normatively permissible, we cannot fairly demand the sacrifice of his entitlement to personal and cultural identity. Insisting that Avon refrain from reaching into his pockets while at the ATM won't do, either: that's an activity necessary for *using* the ATM. If there is no adequately avoidable public signal, there is nothing in *ATM* to license Dollars to assume that Avon is liable, and thus nothing to make her mistake reasonable; it is at best a rational mistake that wrongs Avon.

8 | CONCLUSION

An especially salutary feature of the account I have outlined here is that, while it makes reasonableness judgments explicitly involve balancing agent and patient interests in defensive contexts, it does not restrict our theory choice for explaining the permissibility of defense against genuine threats. All accounts face the same puzzle about mistaken defense; all need some story to tell about why *Stalker* does not pattern together with *Testimony*. The Assumptive Signaling Account delivers such an explanation, with no strings attached.

Appropriately constrained, the ASSUMPTIVE SIGNALS norm solves the distributive problem with which we began. It identifies particular observable behaviors that license defenders to assume that A is liable to harm. Abiding by this norm does not require cognitively costly calculations, nor does it subject agents to the idiosyncrasies or biases of others. It minimizes the incidence of unpredictable gambles by coordinating the heuristics used by a community, letting agents know which behaviors to avoid to minimize their risk of being mistaken for an aggressor. Because it would fulfill the duty agents owe to members of their community on whom they would otherwise impose unjustly disproportionate risk of harm, it is plausibly obligatory to adopt and abide by such a norm.

When an assumptive signaling norm is in place, an agent who has sent a marked signal of aggression which she could easily have avoided cannot reasonably demand that others refrain from defending themselves, and so does not have a complaint against their doing so. This explains the asymmetry between the moral evaluations of defensive mistakes that are reasonable and those that are merely rational. It yields an articulate explanation of when and why the defender's epistemic limitations are relevant to whether she wrongs anyone in making a defensive error, not only vindicating the intuitive judgments about clear cases, but illuminating what we ought to say about murkier, more everyday defensive mistakes.

ORCID

Renée Jorgensen Bolinger https://orcid.org/0000-0002-1351-1892

REFERENCES

Alexander, L. (1993). Self-defense, justification and excuse. *Philosophy and Public Affairs*, *22*(1), 53–66.

Anderson, E. (2010). *The Imperative of Integration*. Princeton, NJ: Princeton University Press.

Armour, J. (1997). *Negrophobia and reasonable racism: The Hidden costs of Being Black in America*, New York: New York University Press.

Bolinger, R. J. (2017a). The pragmatics of slurs. *Nous*, *51*(3), 439–462. https://doi.org/10.1111/nous.12090

Bolinger, R. J. (2017b). Reasonable mistakes and regulative norms: Racial bias in defensive harm. *Journal of Political Philosophy*, *25*(2), 196–217. https://doi.org/10.1111/jopp.12120

[21]For a fuller discussion of this point and of the costs related to exclusion from public spaces, see Anderson (2010).

Bolinger, R. J. (2019). Demographic statistics in defensive decisions. *Synthese*. https://doi.org/10.1007/s11229-019-02372-w

Camp, E. (2013). Slurring perspectives. *Analytic Philosophy*, *54*, 330–349. https://doi.org/10.1111/phib.12022

Correll, J., Wittenbrink, B., Park, B., & Judd, C. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, *83*(6), 1314–1329. https://doi.org/10.1037/0022-3514.83.6.1314

Correll, J., Wittenbrink, B., Park, B., Judd, C., & Goyle, A. (2011). Dangerous enough: Modeling racial bias with contextual threat cues. *Journal of Experimental Social Psychology*, *47*, 184–189. https://doi.org/10.1016/j.jesp.2010.08.017

Correll, J., Wittenbrink, B., Park, B., Judd, C., Sadler, M., & Keesee, T. (2007). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology*, *92*(6), 1006–1023. https://doi.org/10.1037/0022-3514.92.6.1006

Croom, A. (2011). Slurs. *Language Sciences*, *33*(3), 343–358. https://doi.org/10.1016/j.langsci.2010.11.005

Draper, G. (1993). Fairness and self-defense. *Social Theory and Practice*, *19*(1), 73–92. https://doi.org/10.5840/socth eorpract19931913

Draper, K. (2009). Defense. *Philosophical Studies*, *145*(1), 69–88. https://doi.org/10.1007/s11098-009-9387-

Ferzan, K. K. (2005). Justifying self-defense. *Law and Philosophy*, *24*(6), 711–749. https://doi.org/10.1007/s1098 2-005-0833-z

Ferzan, K. K. (2012). Culpable aggression: The basis for moral liability to defensive killing. *Ohio State Journal of Criminal Law*, *9*, 699.

Ferzan, K. K. (2017). The bluff: The power of insincere actions. *Legal Theory*, *23*(3), 168–202. https://doi.org/10.1017/S135232521700026X

Frowe, H. (2010). A practical account of self-defence. *Law and Philosophy*, *29*(3), 245–272. https://doi.org/10.1007/s10981-009-9062-1

Fuller, L. L. (1958). Positivism and fidelity to law - a reply to Professor Hart. *Harvard Law Review*, *71*, 630. https://doi.org/10.2307/1338226

Gardiner, G. (2018). Legal burdens of proof and statistical evidence. In J. Chase & D. Coady (Eds.), *Routledge Handbook of Applied Epistemology* (1st ed.). London: Routledge.

Gersbach, H. (1997). Risk and the value of information in irreversible decisions. *Theory and Decision*, *42*(1), 37–51. https://doi.org/10.1023/A:1004911026088

Gigerenzer, G., & Goldstein, D. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669. 110.1037/0033-295X.103.4.650

Gordon-Solmon, K. (2017). What makes a person liable to defensive harm? *Philosophy and Phenomenological Research*, *97*(3), 543–567. https://doi.org/10.1111/phpr.12369

Jeshion, R. (2013). Slurs and stereotypes. *Analytic Philosophy*, *54*(3), 239–314. https://doi.org/10.1111/phib.12021

Kahn, K. B., & Davies, P. (2010). Differentially dangerous? Phenotypic racial stereotypicality increases implicit bias among ingroup and outgroup members. *Group Processes and Intergroup Relations*, *14*(4), 569–580. https://doi.org/10.1177/1368430210374609

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, *93*(5), 1449–1475. https://doi.org/10.1257/000282803322655392

Lazar, S. (2009). Responsibility, risk, and killing in self-defense. *Ethics*, *119*(4), 699–728. https://doi.org/10.1086/605727

Lewis, D. (1969). *Convention: A Philosophical Study*, Cambridge, MA: Harvard University Press.

Luban, D. (2002). The publicity of law and the regulatory state. *Journal of Political Philosophy*, *10*(3), 296–316. https://doi.org/10.1111/1467-9760.00154

McMahan, J. (2005). The basis of moral liability to defensive killing. *Philosophical Issues*, *15*(1), 386–405. https://doi.org/10.1111/j.1533-6077.2005.00073.x

McMahan, J. (2009). *Killing in War*, New York: Oxford University Press.

McMahan, J. (2011). Who is liable to be killed in war? *Analysis*, (71), 544–559. https://doi.org/10.1093/analys/anr072

Murphy, C. (2005). Lon Fuller and the moral value of the rule of law. *Law and Philosophy*, *24*, 239–262. https://doi.org/10.1007/s10982-004-7990-3

Otsuka, M. (1994). Killing the innocent in self-defense. *Philosophy and Public Affairs*, *23*(1), 74–94. https://doi.org/10.1111/j.1088-4963.1994.tb00005.x

Otsuka, M. (2016). The moral responsibility account of liability to defensive killing. In C. Coons & M. Weber (Eds.), *The Ethics of Self-defense* (pp. 51–68). New York: Oxford University Press.

Quiggin, J. (2016). The value of information and the value of awareness. *Theory and Decision*, *80*, 167–185. https://doi.org/10.1007/s11238-015-9496-x

Quong, J. (2009). Killing in self-defense. *Ethics*, *119*, 507–537. https://doi.org/10.1086/597595

Quong, J. (2012). Liability to defensive harm. *Philosophy and Public Affairs*, *40*(1), 45–77. https://doi.org/10.1111/j.1088-4963.2012.01217.x

Quong, J. (2015). Rights against harm. *Proceedings of the Aristotelian Society*, (89), 249–266. https://doi.org/10.1111/j.1467-8349.2015.00252.x

Richardson, L. S., & Goff, P. A. (2012). Self-defense and the suspicion heuristic. *Iowa Law Review*, *98*, 293–336.

Ripstein, A. (1996). Self defense and equal protection. *University of Pittsburgh Law Review*, *57*, 685–724.

Rodin, D. (2003). *War and Self-defense*, Oxford: Oxford University Press.

Sim, J., Correll, J., & Sadler, M. (2013). Understanding police and expert performance: When training attenuates (vs. exacerbates) stereotypic bias in the decision to shoot. *Personality and Social Psychology Bulletin*, *39*(3), 291–304. https://doi.org/10.1177/0146167212473157

Skyrms, B. (2010). *Signals*. Oxford University Press.

Thomson, J. J. (1991). Self-defense. *Philosophy and Public Affairs*, *20*(4), 283–310.

Uniacke, S. (1994). *Permissible killing: The self-defence justification of homicide*, Cambridge: Cambridge University Press.

How to cite this article: Bolinger, R. J. The Moral Grounds of Reasonably Mistaken Self-Defense. *Philos Phenomenol Res*. 2020;00:1–17. https://doi.org/10.1111/phpr.12705