IMPLICIT RACIAL ATTITUDES AND SELF-DEFENSE

STEPHEN P. GARVEY*

INTRODUCTION

Defendant (call him John) shoots victim, intending to cause victim's death. Victim dies. John is charged with murder and pleads self-defense. If the elements of the defense are established,[1] John should be acquitted. One of those elements refers to John's beliefs, according to which John must have believed it was "necessary" to use deadly force to protect himself against death or serious bodily injury. But his belief (or "honest" belief)[2] that the use of deadly force was "necessary" isn't sufficient to establish the defense, even if all the other elements of the defense have been established.[3] His belief must also have been "reasonable." It might have been false, but so long as it was nonetheless reasonable, John is entitled to an acquittal, assuming all the other elements of the defense are established.

For ease of exposition call John's belief that the use of deadly force was necessary to protect himself against the use of deadly force the belief that *p*, where *p* refers to the propositional content of John's belief, i.e., "the use of deadly force is necessary."

Now suppose the person John killed is a black man.[4] Next, enter the following stipulation: John (honestly) believed he needed to use deadly force

---

* A. Robert Noll Professor of Law, Cornell Law School. I'm grateful to Larry Alexander, Kevin Clermont, Joshua Dressler, Russell Osgood, Emily Sherwin, and the faculty of Washington University School of Law for taking time to read and comment on an earlier draft.

1. I ignore for present purposes questions about the burden of production and persuasion with respect to the elements of the defense. According to one treatise, the "burden of production for the defense of self-defense is always on the defendant. The burden of persuasion is almost always on the state, beyond reasonable doubt." *See* PAUL H. ROBINSON, 2 CRIMINAL LAW DEFENSES § 132, at 99–100 (1984).

2. The reference to "honest" belief is routinely found in judicial opinions and academic commentary. That's harmless enough so long as one bears in mind that, so far as I can tell, an honest belief is nothing more than a belief a person in fact possesses or has formed. A person may honestly or dishonestly *report* the content of his beliefs, but simply describing a belief as "honest" adds nothing to the word "belief" standing alone. Perhaps the word "honest" is sometimes added simply as a way to highlight the contrast between a belief and a reasonable belief. An "honest" belief is also sometimes said to be equivalent to a "good faith" belief. If so, then the point just made would apply to the idea of a "good faith" belief just as it applies to the idea of an "honest" belief. I suppose one could imagine a "good faith" belief being distinguished from an "honest" belief in various ways, but I'm not aware of courts or commentators having drawn any such distinction when discussing criminal-law offenses or defenses.

3. When I say "established," I mean proven to the satisfaction of a jury based on whatever standard of proof, placed upon whichever party, pursuant to the law of the particular jurisdiction in which the defendant is being tried.

4. Because John's race is irrelevant to the analysis offered below, I leave it unspecified. I believe John's race is irrelevant because I believe that people who have taken the IAT (or some such test) and who are described as black can also manifest in their performance the presence of the implicit racial attitude on which I focus, and some do. Having reported that belief, I don't provide any citation to the literature as evidence for its truth.

because of that which psychologists call an "implicit racial attitude."[5] In other words, but-for this implicit racial attitude John would not have believed the use of deadly force was necessary, and would not have chosen to use deadly force at the time he used it. John's implicit racial attitude was a but-for cause of his belief that *p:* it was necessary, though not sufficient, for the formation of the belief that *p*.[6]

With that stipulation in place, we reach the question: should the (stipulated) fact that John would not have formed the belief that *p* but for the

---

5. Four observations related to "implicit racial attitudes."

First, I don't discuss the various reasons why some commentators have expressed skepticism about some of the factual and normative claims they believe have been made about implicit attitudes, in both the academic literature and in the popular press, where those claims are often based in turn on factual claims found in the psychological literature on implicit attitudes. For an account of some of these expressions of skepticism, along with a response to each of them, see, for example, Michael Brownstein, *Skepticism About Bias*, *in* AN INTRODUCTION TO IMPLICIT BIAS: KNOWLEDGE, JUSTICE AND THE SOCIAL MIND 57 (Erin Beeghly and Alex Madva eds., Taylor & Francis Group, 2020). *See also* Bertram Gawronski, *Six Lessons for a Cogent Science of Implicit Bias and Its Criticism*, 14 PERSP. ON PSYCH. SCI. 574 (2019). That said, I do offer words of caution (in Part III) about what conclusions can and should be drawn from factual claims found in the psychological literature to the effect that "motivation to control prejudice" are capable of neutralizing, offsetting, rendering causal inert, and so forth the causal influence an implicit attitude would otherwise have had on the beliefs a person forms.

Second, I use the phrase "implicit attitude" rather than other phrases found in the legal and psychological literature to describe the phenomenon at issue—like "implicit bias," "implicit prejudice," "implicit discrimination," "implicit racism," and so on—because I believe these alternative names carry, whether or not intended to carry, a critical or condemnatory moral valence or implication. If so, the reasons why the phenomenon should be understood to possess any such valence or carry any such implication should be made explicit, and not tacitly assumed or implied in the words chosen to name the phenomenon.

Third, I don't use the word "discriminatory" to characterize a belief formed as a result of an implicit racial attitude. What it means to characterize a belief or an action as "discriminatory," when that characterization presupposes that the belief or action in question was wrongful, is a matter of considerable debate and controversy. *See, e.g.*, KASPER LIPPERT-RASMUSSEN, BORN FREE AND EQUAL? A PHILOSOPHICAL INQUIRY INTO THE NATURE OF DISCRIMINATION 2 (2013) ("Not only do people differ over which cases of differential treatment they see as discriminatory, they also disagree about when discrimination is morally wrong and about what makes discrimination morally wrong."); Andrew Altman, *Discrimination*, STAN. ENCYCLOPEDIA PHIL. § 4.1, at 23 (2016) ("Specifying why direct discrimination is wrongful has proved to be a surprisingly controversial and difficult task."). Rather than use the word and beg questions as to its meaning, I avoid its use altogether.

Fourth, although the discussion and argument offered in the text is limited to implicit racial attitudes, I don't see any obvious reason why the conclusion reached would not apply to other implicit attitudes revealed in the various tests used to detect the presence of such attitudes insofar as they are postulated as that which explains a person's performance on these tests.

6. The causal path between the "activation"—a term used in the psychological literature—of an implicit attitude and the belief that *p* might be direct (as when the implicit attitude causes the belief that *p* without causing any mediating belief) or indirect (as when the implicit attitude causes another belief, such as the belief that the victim was holding a gun when in fact he wasn't, which in turn causes the belief that *p*).

causal effect of an implicit attitude suffice to describe his belief that *p* as "unreasonable"? We can turn that normative question into a concrete legal proposal. Suppose a state lawmaker is faced with the following amendment to his or her state's penal code:

> When a defendant is charged with murder and pleads self-defense, any belief he formed that the use of deadly force was necessary to protect himself against deadly force is unreasonable if he would not have formed that belief but for an implicit racial attitude.[7]

---

7. Four additional observations about the Proposal worth noting:

First, the Proposal is analogous to proposals to reform the law of rape or sexual assault pursuant to which evidence sufficient to prove that the victim said "no" is itself made sufficient as a matter of law to establish "non-consent." The Proposal would analogously make evidence sufficient to prove that a belief resulted from the causal influence of an implicit attitude itself sufficient as a matter of law to establish that belief's "unreasonableness."

Second, the Proposal (as well as the argument presented in the text) assumes an implicit attitude (whatever such a thing is) can causally influence (be a cause of) the formation of beliefs, which can in turn causally influence (be a cause of) choices and actions, like the choice to act in a way fairly described as the "use of deadly force." Having said that, I note in connection with this assumption the following observation from a 2020 review on "implicit social cognition."

> There is presently no empirical basis to choose among the three proposed forms of explanation for observed correlations between indirect measures [implicit attitudes] and behavior: (*a*) automatic effects of associations on behavior, (*b*) overlapping influences that (independently but relatedly) produce both associative knowledge and related behaviors, and (*c*) cooperative causation in which automatically activated associations produce conscious judgments that play at least a partial role in guiding judgments and behavior.

Anthony G. Greenwald & Calvin K. Lai, *Implicit Social Cognition*, 71 ANN. REV. PSYCH. 419, 428 (2020).

I don't know precisely what the authors mean when they refer to "cooperative causation" in (*c*) above—a term I can't recall encountering in either the criminal law literature or the philosophical literature with which I'm familiar—but I assume it means an implicit association (whatever it is) can cause (in the sense of but-for cause) a person to form a belief he otherwise would not have formed, where that belief in turn figures into the person's practical reasoning: the reasoning by which a person forms a judgment as to what he should do all things considered. If so, then I assume herein the truth of (*c*), despite (as the quoted language says) having no "empirical basis" to choose it over the other two hypotheses mentioned in the quoted language. Nor can I provide a folk-psychological account explaining how an implicit attitude causally influences the beliefs a person forms. I simply assume implicit attitudes can and do causally influence the beliefs a person forms.

Third, in addition to assuming that implicit attitudes can cause the formation of cognitive attitudes (here, the belief that *p*), I also assume that beliefs, together with other mental states, can cause choices and actions. This assumption is controversial for two reasons:

> 1) The proposition that implicit attitudes can cause actions is controversial. For example, Edouard Machery, in a recent article responding to another recent article by Brownstein *et al.*, writes: "[I]mplicit attitudes are often understood as impacting behavior causally. However, after 30 years of research, there is almost no evidence that indirect measures measure something causally efficient rather than merely epiphenomenal." Edouard Machery, *Anomalies in Implicit Attitudes Research*, WILEY INTERDISC. REV.: COGNITIVE SCI. (2022) (responding to Michael Brownstein et al., *What Do Implicit Measures Measure?*, WILEY INTERDISC. REV.: COGNITIVE SCI. (2019)). The exchange

Call this proposed amendment the Proposal. The Proposal is hypothetical. So far as I know, no legislature has entertained anything like it. But the concept of implicit attitudes has jumped with remarkable speed from the world of academic psychology into the world of politics—an observation I assume needs no string citation—and criminal law is a product of legislative politics. Law review articles proposing legal reforms of one sort of another, based in one way or another, on the psychological literature on implicit attitudes, are now commonplace, with additional proposals appearing with some regularity. Thus, the Proposal, though hypothetical, is worth entertaining. So, should a state lawmaker vote in favor of, or against, the Proposal?

Some academic writers would, I surmise, urge lawmakers to vote in favor.[8] According to these writers (as I understand them), if a person uses deadly force because he believed that *p*, his belief that *p* is (and should as a matter of law be defined as) unreasonable if and because it was the but-for result of an implicit racial attitude (as the Proposal provides). They believe a person like John should therefore be held criminally liable, either for murder or for

---

continued in letters to the editor. *See* Bertram Gawronski et al., *How Should We Think About Implicit Measures and Their Empirical "Anomalies"?*, WILEY INTERDISC. REV.: COGNITIVE SCI. (2022); Edouard Machery, *Anomalies in Implicit Attitudes Research: Not So Easily Dismissed*, WILEY INTERDISC. REV.: COGNITIVE SCI. (2022). For a comment on the exchange, see Chandra Sripada, *Whether Implicit Attitudes Exist is One Question, and Whether We Can Measure Individual Differences Effectively Is Another* WILEY INTERDISC. REV.: COGNITIVE SCI. (2022) (A "verdict," according to which "implicit attitudes exist but we cannot measure individual differences effectively," "is eminently reasonable"). This disagreement among the psychological experts reinforces, I believe, the advice to lawmakers stated on p. 57.

2) The broader proposition that mental states can cause action — known as the "causal theory of action" — is itself controversial. *See generally* CAUSING HUMAN ACTIONS (Jesús H. Aguilar & Andrei A. Buckareff eds., 2010).

Fourth, I am assuming that the point in time at which the defendant forms the reasonable belief that the use of deadly force is necessary to protect himself against deadly force is the point in time at which the law of self-defense permits him to use deadly force, whether or not the defendant chooses to exercise that permission at that time.

8. I base my surmise on two articles in the criminal-law literature, which speak more or less directly to the Proposal: L. Song Richardson & Phillip Atiba Goff, *Self-Defense and the Suspicion Heuristic*, 98 IOWA L. REV. 293 (2012), and Jules Holroyd & Federico Picinali, *Implicit Bias, Self-Defense, and the Reasonable Person*, *in* THE CRIMINAL LAW'S PERSON 165 (Claes Lernestedt & Matt Matravers eds., 2022). My surmise would be strengthened if the Proposal were modified to read:

> When a defendant is charged with murder and pleads self-defense, any belief he formed that the use of deadly force was necessary to protect himself against deadly force is unreasonable if he would not have formed that belief but for an implicit racial attitude, provided that any defendant who believed, reasonably or not, that the use of deadly force was necessary to protect himself against deadly force shall be convicted of manslaughter.

What I understand to be the main argument made in the two articles cited above is discussed in more detail in Appendix A.I. I don't discuss these articles in the text because doing so would, I think, distract from the main line of argument I wish to offer. Other articles more-or-less directly relevant to the Proposal may exist of which I'm unaware.

manslaughter: for manslaughter, if imperfect self-defense is available as a partial defense to a charge of murder; for murder, if not.[9]

This line of thought moves quickly. Indeed, I believe it moves too quickly. It moves from the premise that a belief that *p* was the result of an implicit attitude to the conclusion that the belief that *p* was necessarily unreasonable and should therefore be defined as a matter of law to be unreasonable. Nonetheless, I believe more needs to be said before one can get from the premise and the conclusion.

The space between the premise and the conclusion can, as I see it, be bridged with three additional premises. One of those premises is conceptual (about the meaning of the word "unreasonable" as used in the law of self-defense); one is normative (about what should make a belief "unreasonable" for purposes of the law of self-defense); one is metaphysical (about the nature of implicit attitudes).

With these additional premises in place the argument leading to the Proposal would look like this:

> 1. An unreasonable belief that *p* is a belief a person is culpable for having formed (i.e., what it means to say that a belief that *p* was "unreasonable," or "unreasonably formed," is that the person was "culpable" for having formed it). This is a *conceptual* premise inasmuch as it constitutes a proposition about the meaning of the word "reasonable" in the law of self-defense. "Reasonable" means "non-culpable," and "unreasonable" means "culpable."
>
> 2. A person is culpable for having formed the belief that *p* if he would not have formed it if he'd had "sufficient concern" for the life of the victim, where having sufficient concern for the life of the victim means the life of the victim had as much weight as the person's own life in the person's (theoretical) reasoning (i.e., that capacity or power by means of which a person forms beliefs about the world).[10] This is a *normative* premise inasmuch as it constitutes a proposition about what makes the formation of a belief culpable (and thus unreasonable).
>
> 3. Sufficient concern for the life of the victim can and would have neutralized the causal influence an implicit racial attitude would otherwise have had on a person's formation of the belief that *p*. This is a metaphysical premise inasmuch as it constitutes a proposition

9. "Imperfect self-defense" is a doctrine pursuant to which (in one of its forms) a defendant who unreasonably (but honestly) believes that *p* is liable for manslaughter, not murder. Because the doctrine mitigates what would otherwise be murder to manslaughter, the doctrine is commonly described as a "partial defense." *See* JOSHUA DRESSLER, UNDERSTANDING CRIMINAL LAW § 18.03, at 222 (8th ed. 2018).

10. This premise is a premise in an argument about what it takes for a person's *belief* that *p* is to be described as "reasonable" or not; it isn't a premise in an argument about the moral status—permissible or impermissible, excused or not excused—of a person's *choice* to use deadly force to protect against the use of deadly force, *given* that a person's belief that *p* is "reasonable" or "unreasonable." I don't address this latter argument herein, although I mention it *infra* pp. ___-___.

about the nature of implicit attitudes; specifically, a proposition stating that the causal effect of an implicit attitude on the formation of a belief is itself causally dependent on "concern."

From these three premises, several conclusions can be drawn, leading to the conclusion that the Proposal should be adopted into law.

4. Therefore, a person who in fact forms the belief that *p* as a result of an implicit racial attitude had insufficient concern for the life of the victim, because if he'd had sufficient concern for the life of the victim, that concern would have neutralized the causal force of the implicit racial attitude, and he would not have formed the belief that *p*.

5. Therefore, a person who in fact forms the belief that *p* as a result of an implicit racial attitude formed that belief culpably (i.e., as a result of insufficient concern).

6. Therefore, a person who in fact forms the belief that *p* as a result of an implicit racial attitude has formed that belief unreasonably (i.e., culpably).

7. Therefore, the Proposal, which defines as unreasonable (as a matter of law) a belief that *p* when that belief is the result of an implicit racial attitude, should be enacted into law, to be included as a definition provision among the provisions defining the elements of self-defense.

I believe propositions (1) (the conceptual premise) and (2) (the normative premise) are true. But the truth of proposition (3) (the metaphysical premise) remains, I believe, in doubt; and if proposition (3) remains in doubt, then so too does the final conclusion stated in proposition (7), along with the intermediate conclusions stated in propositions (4), (5), and (6).

If premise (3) is in doubt—if it turns out that implicit racial attitudes are not the kind of thing the causal force of which sufficient concern can neutralize—then the Proposal, if enacted, would permit the state to punish someone who reasonably (and thus non-culpably) believed the use of deadly force was necessary to protect himself against the use of such force.[11] Because

11. The analysis offered below attempts to identify the conditions under which a person who forms a belief—stipulated to be wrongful (more on that below in Part I.B.1)—has culpably formed that belief, with criminal liability following upon a finding that the belief was culpably formed.

It ignores many other questions (of a more practical nature) the Proposal raises. For example: how is the state to go about proving a defendant possessed an implicit attitude at the moment he chose to kill? Would the defendant be required to take an IAT? Would psychological studies reporting differences in average IAT scores across different demographic populations be admissible evidence to prove the defendant harbored an implicit attitude? How will the state prove the defendant formed the belief that *p* only as a result of an implicit attitude? Will or might jurors in cases in which the evidence available to prove the necessity element is ambiguous or marginal, having been instructed in the language of the Proposal, be more likely to assume an implicit attitude must have caused the defendant to have formed the belief that *p* whenever the victim is black?

These are questions about proof and the psychology of jury decision-making. I leave them for those more versed in those domains. I suspect these questions are not as easy to answer as some

I believe (consistent with prevailing law) that the state should not punish a person unless he unreasonably (and thus culpably) formed the belief that *p*, even when "acting on" that belief results in the death of an innocent person, I would recommend tabling the Proposal until more is confidently known about the metaphysics of implicit attitudes.

Part I isolates the question at hand and distinguishes it from three questions with which it might be confused. The question at hand (once again) is: should a belief that *p* be defined as unreasonable as a matter of law just because a person formed that belief as a result of an implicit racial attitude and would not have formed that belief but for the implicit attitude? Readers familiar with the law of self-defense and who don't believe themselves at risk of confusing this question with others, might consider skipping ahead to Part II.

Part II (specifically, Part II.B) defends the conceptual and normative premises mentioned above. It first makes express my assumption that a belief is unreasonable just in case it was culpably formed. It then introduces an account (or theory) setting forth two conditions under which a belief (including the belief that *p*) should be characterized as culpable (unreasonable), at least when the consequence of so characterizing it is criminal liability, and thus liability to state punishment. This theory of culpable (unreasonable) belief is dubbed the insufficient-concern theory. This theory is fairly well-known in the criminal-law theory literature, though of course it remains the subject of debate and disagreement.

Part III asks if the metaphysical premise mentioned in (3) above—that sufficient concern can and would neutralize the causal influence an implicit attitude would otherwise have had on the beliefs a person forms—is true. So far as I can tell, we don't yet know. I therefore remain agnostic as to its truth.

I add four additional observations before moving to Part I.

First, the discussion in each part proceeds at a brisk pace. Many details—clarifications, elaborations, and so forth, not directly relevant to the narrow question at hand—get relegated to footnotes and appendices in the hope of keeping the main outlines of the argument plain to see and follow. I also try to avoid as much jargon as possible, but some specialized language, whether associated with psychology, philosophy, or law, has been hard to avoid. Nor do I discuss the facts of any recent high-profile case. Saying what one believes to be the "correct" or "right" result in a particular case involving a claim of self-

---

commentators appear to suggest. *See* Holroyd & Picinali, *supra* note 8, at 190 n.74 ("Someone may raise issues concerning the implementation of any proposal that would require ascertaining the role of implicit bias in the defendant's actions.. . . [I]t seems to us that the problems raised by this objection are not qualitatively different from problems that beset other defences with long-standing legal pedigree: consider defences of insanity, loss of control, diminished responsibility and intoxication."). One wonders: aren't some of the facts relevant to the listed defenses plainly manifest in the defendant's observable behavior at the time of the crime, whereas that's not the case for the alleged fact that an implicit attitude caused the defendant to form the belief that *p*? In any event, for a recent discussion describing judicial decisions addressing the admissibility of evidence based on the IAT, see Frank Harty & Haley Hermanson, *Implicit Bias Evidence: A Compendium of Cases and Admissibility Model*, 68 DRAKE L. REV. 1 (2020).

defense depends in no small part on having all the facts, and even then, reasonable minds can disagree when applying the law to those facts.

Second, the John hypothetical is used to facilitate analysis, but its factual stipulations obviously make it remote from any real-world case. No real-world case comes with its facts neatly packaged and sorted into doctrinally or philosophically relevant categories. Some readers may therefore regard the hypothetical as artificial, fanciful, or even silly. That's fair.[12] Still, without some such packaging, any hope for progress—even if limited to identifying points of disagreement—on an answer to the narrow question at hand is apt to get lost amid disagreements over the facts. Stipulations are the price paid to bracket such disagreements.

Third, the argument I make tries to the extent possible to avoid taking positions on controversies surrounding the metaphysics of implicit attitudes. Having said that, I do take one such position. I assume that an implicit attitude, whatever else it might be, is *not* a belief (conscious or unconscious). Specifically, an implicit racial attitude, whatever else it may be, is not a belief the content of which is a race-based generalization having to do with any supposed relationship between being black and being violent, criminal, aggressive, suspicious, and so forth. The reasons I make this assumption are given in Part III below.[13]

Fourth, if a person believes that *p*, but a jury finds his belief unreasonable, he will be found guilty of murder or manslaughter, and having been found guilty of one of those crimes, he will be punished, probably with imprisonment.[14] The consequences for having been found to have believed unreasonably are therefore not insignificant. I believe these consequences should be kept firmly in mind when trying to identify or specify the conditions under which the law of self-defense should ascribe culpability (unreasonableness) to the belief that *p*. If the consequences were different, the conditions might be different as well. What it should take to ascribe culpability for having formed a legally relevant belief should depend, at least in some part, on what follows from it. In criminal law, when it comes to failed claims of self-defense, what usually follows is prison.

---

12. The hypothetical is other-worldly in other way: it ignores any emotion or passion involved when a defendant uses deadly force in self-defense. Of course, rightly or wrongly, the law of self-defense ignores emotion or passion as well, inasmuch as the elements of the defense make no reference to any emotion or passion, or loss of "self-control" arising from them, at the time the defendant chose to use deadly force. *See, e.g.*, Kenneth W. Simons, *Self-Defense: Reasonable Beliefs or Reasonable Self-Control?*, 11 NEW CRIM. L. REV. 51 (2008).

13. *See infra* Part III.

14. Punishment is here understood as a burden or hardship the state intentionally imposes in order to condemn or censure a person for having been found to have culpably committed a crime. *See, e.g.*, Joel Feinberg, *The Expressive Function of Punishment*, *in* DOING AND DESERVING: ESSAYS IN THE THEORY OF RESPONSIBILITY 95, 98 (1970). Other analyses of the concept of punishment have of course been offered and defended, but I believe the just-mentioned analysis is the one most commonly, though not uncontroversially, found in the criminal-law theory literature.

I. ISOLATING THE QUESTION

When a person forms the belief that *p*, but would not have formed that belief but for the causal influence of an implicit racial attitude, has that person formed the belief unreasonably? That's the question. This Part tries to set the stage for answering that question.

First, subpart A discusses elements of the defense not relevant to the question presented. Second, subpart B explains the sense in which the "necessity" element, which is an element relevant to the question presented, is an "objective" element. As an "objective" element, the necessity element is an element about which a defendant may be mistaken, reasonably or unreasonably. A defendant can believe the element exists when "objectively" it doesn't. Finally, subpart C distinguishes the question presented from four other questions with which it can easily be confused.

A. Irrelevant Elements

The law governing the use of private force in self-defense differs from one jurisdiction to the next.[15] To make the present discussion manageable, I rely on a generic statement of what might be called the "core" of the defense: a person has a right (permission) to use deadly force if he *reasonably believes* that the use of such force is *necessary* to protect himself against deadly force. The core of the defense therefore rests most importantly on the elements of necessity and reasonable-belief.

Because my focus is on the necessity and reasonable-belief elements, I ignore other elements one might find in statutory formulations of the defense. Among the elements I set aside:

> 1. Imminence. Some jurisdictions permit the use of deadly force only if the person using such force reasonably believed, not only that the use of deadly force was necessary to protect against the use of deadly force, but also only if he reasonably believed the deadly force against which he used deadly force to protect himself was "imminent." The wisdom of this additional "imminence" requirement, and its relationship to the "necessity" requirement, have been much-discussed and debated.[16]
>
> 2. "Reckless" Belief. The Model Penal Code distinguishes (in a roundabout way) for purposes of self-defense between a "negligent"

15. For an overview, *see*, for example, PAUL H. ROBINSON & TYLER SCOT WILLIAMS, MAPPING AMERICAN CRIMINAL LAW: VARIATIONS ACROSS THE 50 STATES (2018); Victoria Nourse, *Self-Defense, in* THE OXFORD HANDBOOK OF CRIMINAL LAW 607 (Markus D. Dubber & Tatjana Hörnle eds., 2014).

16. *See, e.g.*, DRESSLER, *supra* note 9, § 18.02[D][1], at 220–21. The imminence requirement has been subject to especially strong criticism insofar as it has presented an obstacle to acquittal when self-defense is raised by battered spouses who kill their abusers in a non-confrontational setting. *See, e.g.*, Whitley R.P. Kaufman, *Self-Defense, Imminence, and the Battered Woman, in* CRIMINAL LAW CONVERSATIONS 407 (Paul H. Robinson et al. eds., 2011).

belief that *p* and a "reckless" belief that *p*.[17] Because I believe the Code's definition of "reckless" belief is incoherent,[18] or at least difficult to make sense of, I will ignore it. I assume for present purposes that the law categorizes a person's belief that *p* exclusively as either "reasonable" or "unreasonable" (as I believe most states do).[19]

3. Aggressor Rules. All jurisdictions (so far as I know) withdraw a defendant's right to use deadly force in self-defense if he qualifies as an "aggressor" or "provocateur" who hasn't renounced his aggression or provocation (even if he would otherwise have satisfied the elements of the defense at the moment he used deadly force).[20] What suffices to make a person an unrenounced aggressor or provocateur differs from one jurisdiction to the next. I assume for present purposes that John does not qualify as an unrenounced aggressor or provocateur.

4. Duty to Retreat. A minority of American jurisdictions withdraw a defendant's right to use deadly force in self-defense if he fails to retreat before using deadly force (even if he would otherwise have satisfied the elements of the defense at the moment he used deadly force). A majority of American jurisdictions impose no such duty to retreat.[21] The absence of a duty to retreat is doctrinally distinct from legal provisions popularly grouped under the heading of "Stand Your Ground" laws, although the two doctrines are sometimes erroneously treated as equivalent.[22] I assume for present purposes either that the

17. The Code's distinction between a "negligent" belief and a "reckless" belief is an effort to distinguish between two different cognitive attitudes toward the proposition that *p*, with a reckless belief intended to be a more culpable cognitive attitude compared to a negligent belief. The Code also tries to align the culpability associated with a defendant's mistaken belief that *p* with the homicide offense for which the defendant is ultimately liable. A defendant who made a negligent mistake is liable for negligent homicide (but not murder or manslaughter). A defendant who made a reckless mistake is liable for manslaughter (but not murder). *See* MODEL PENAL CODE §§ 3.04, 3.09; DRESSLER, *supra* note 9, § 18.06[B], at 241–42.

In addition to the Model Penal Code's effort to distinguish between "negligent" beliefs and "reckless" beliefs, beliefs can also be distinguished from one another based on what philosophers describe as "degrees of belief," i.e., how much confidence or credence a person has in the truth of a particular proposition. I briefly discuss this idea, as it figures into proposals to reform the law of self-defense, in Appendix D.

18. *See, e.g.*, Larry Alexander, *Lesser Evils: A Closer Look at the Paradigmatic Justification*, 24 LAW & PHIL. 611, 624–26 (2005).

19. The Model Penal Code "designates" a "reasonable belief" as a "belief which the actor is not reckless or negligent in holding." MODEL PENAL CODE § 1.13(16). Conversely, an "unreasonable belief" is presumably one the actor is either reckless or negligent in "holding."

20. *See, e.g.*, DRESSLER, *supra* note 9, § 18.02[B], at 214–16.

21. *See, e.g.*, *id.* § 18.02[C], at 217 ("Although the original common law rule was that a person was required to retreat . . . , a majority of jurisdictions today do not require a non-aggressor to retreat under the threat of deadly force, even if he could do so in complete safety.").

22. *See, e.g.*, Cynthia V. Ward, *"Stand Your Ground" and Self-Defense*, 90 AM. J. CRIM. L. 89 (2015).

> law imposed on John no duty to retreat, or if it did, that John has discharged that duty.

Again, I ignore the above features of self-defense law because the narrow question I ask doesn't implicate them. It implicates only the necessity and reasonable-belief elements. Any controversy surrounding these other features should be addressed separately and on their own merits. I now turn to the necessity element.

B. Necessity

The necessity element appears to presuppose a fact of the matter. That is, either the use of deadly force was (in fact) necessary, or it wasn't. Or, as the same idea might otherwise be expressed, the necessity element appears to state an "objective" element of the defense. If so, then a jury would in principle need to answer that question before asking about the reasonableness of a defendant's belief with respect to that fact. Only if the use of deadly force was in fact (or "objectively") not necessary, such that the defendant's belief that it was necessary was mistaken, would a jury need to ask about the reasonableness of the defendant's mistaken belief.[23] How should a jury go about answering this "objective" question about the existence or non-existence of "necessity"?

Unfortunately, any effort at a complete answer would take us far afield.[24] For now, I try to make things simple. I assume a set of facts I hope most people would agree would make the defendant's use of deadly force "in fact" unnecessary (on any plausible account explaining what that might mean). Suppose, for example, that what John believed was a gun in the victim's hand at the moment he used deadly force was discovered, after John uses deadly force, to have been a cellphone. On those facts it makes as much sense as it ever will to say that John's choice to use deadly force was "in fact" unnecessary, despite John's belief to the contrary at the moment he used deadly force.

C. Four Questions

In the hypothetical involving John, John chooses to kill, with the intent to cause death, only because he believed that *p*, and he believed that *p* only as a result of the causal influence of an implicit attitude. This hypothetical raises at least four normative questions. The first two address the moral properties of John's "mental act" of forming the belief that *p*. The second two address the moral properties of John's choice (and resulting bodily act), having formed that belief, to use deadly force.

---

23. I ignore cases (known in the literature as cases of unknowing or unwitting justification) in which a defendant uses deadly force, the use of such force was in fact necessary, but the defendant didn't believe the use of such force was necessary. Criminal law theorists have long debated how the criminal law should deal with such cases: should the defendant be convicted of a completed crime or only of an attempt? For a recent addition to this literature, drawing on the author's distinctive thesis as to why resulting harm should matter to criminal liability, see Peter Westen, *Unwitting Justification*, 55 SAN DIEGO L. REV. 419 (2018).

24. I try to explain why in Appendix B.

> Question 1. Did John's formation of the belief that *p* (as a result of the causal influence of an implicit attitude) constitute an epistemic wrong? In other words, did it violate some epistemic obligation governing the beliefs John formed?
>
> Question 2. Assuming John's formation of the belief that *p* (as a result of the causal influence of an implicit attitude) was an epistemic wrong, was John's formation of that wrongful belief nonetheless "reasonable" (non-culpable)?
>
> Question 3. Assuming John's formation of the belief that *p* was an epistemic wrong, but that its formation was nonetheless reasonable, did John's choice to kill as a result of having formed that wrongful but reasonable belief constitute a moral wrong (recognized as a legal wrong)? In other words, did his choice to kill violate some moral (legal) obligation governing John's choices and actions?
>
> Question 4. Assuming John's formation of the belief that *p* was an epistemic wrong, but that its formation was nonetheless reasonable, and assuming his choice to kill constituted a moral wrong (recognized as a legal wrong), was his wrongful choice to kill nonetheless excused (non-culpable)?

The second question is the one I ask, and to which I offer an answer, in Part II. In answer to the first question, I will stipulate, for reasons given below, that John's formation of the belief that *p* was indeed wrongful. I do so in order to reach the second question, since the first question analytically precedes the second. I also discuss below the third and fourth questions, which are analytically posterior to the second, but I offer no opinion on how they should be answered.

1. Question One

The first question, to which I will stipulate an answer, is:

> Question 1. Did John's formation of the belief that *p* (as a result of the causal influence of an implicit attitude) constitute an epistemic wrong? In other words, did it violate an epistemic obligation governing the beliefs John formed?

Again, this question is analytically prior to the second. As I explain in Part II, the theory of reasonable belief developed in that Part identifies an unreasonable belief with a culpable belief (or culpably formed belief), and it identifies a reasonable belief with a non-culpable belief (or non-culpably formed belief).[25] Or, to put the same point differently: an unreasonable belief is a belief a person should not be excused for having formed (because it was culpably formed), and a reasonable belief is a belief a person should be excused for having formed (because it was non-culpably formed).

---

25. It may be that one should say that a non-culpably formed belief is not an unreasonable belief, rather than say that a non-culpably formed belief is a reasonable belief. I'm not sure. In any event, I will in the text continue to equate a non-culpably formed belief with a reasonable belief.

Insofar as a reasonable belief just is a belief a person non-culpably (or excusably) formed, a person who formed the belief that *p* presumably needs to plead an excuse for having formed it only if his formation of it was (in some sense) wrongful. I assume a belief (including the belief that *p*) is wrongful if and only if its formation constitutes a violation or breach of an epistemic obligation: a rule governing the beliefs a person is, due to the obligation, obligated to form. If a person's formation of the belief that *p* was not wrongful—not in violation of any epistemic obligation—then nothing exists to excuse.

Assuming we have epistemic obligations—an assumption to which not everyone would assent—the same logic governing our moral obligations would seem to apply (*mutatis mutandis*) to our epistemic obligations.[26] If a reasonable belief is a belief the formation of which should be excused, because it was non-culpably formed (and if an unreasonable belief is a belief the formation of which should not be excused, because it was culpably formed), then asking about the reasonableness or not of a belief presupposes the belief in question was wrongfully formed: formed in violation of an epistemic obligation. So, assuming we have epistemic obligations, what exactly are they?

People disagree. For example, one might say we have an obligation to form true beliefs and not to form false beliefs. If so, then insofar as a person believes that *p*, and *p* is false, he violates an epistemic obligation: His false belief is in that sense, and for that reason, a "wrongful" belief. Or, one might say a person has an epistemic obligation to form beliefs based on (or conforming to or supported by) his available evidence.[27] If so, then we need to know more. For example, what counts as "evidence"? What does it mean to say that evidence is "available"? What does it mean to say that a belief is "based on" (or "conforms to" or is "supported by") a person's available evidence?[28] And so on.

Accounts such as these have at least one thing in common. They assume our moral obligations (obligations governing what we choose and do) are one thing and our epistemic obligations (obligations governing the beliefs we form) are another. Our epistemic obligations depend on how the world is in fact, or on the evidence available to us about how the world is in fact. We are epistemically obligated to form true beliefs, or beliefs that conform to our available evidence. They don't depend on any account describing how the

26. Some philosophers have argued it makes no sense to claim we have epistemic obligations. As they see it, obligations govern the choices we make, but not the beliefs we form. It makes sense to speak of obligations to choose to do something or to choose not to do something because we ordinarily have "control" over our choices, but it makes no sense to speak of obligations to believe this or that because we lack "control" over the beliefs we form. If "ought implies can," and if we can't choose to conform or not to conform to (putative) obligations to believe this or that, then no such obligations exist. For a brief response to this argument, see *infra* Part II.B.1.

27. This thesis in known in the philosophical literature as evidentialism, inasmuch as it claims a person is epistemically obligated to form beliefs based on and in proportion to his available evidence. *See* Andrew Chignell, *The Ethics of Belief*, STAN. ENCYCLOPEDIA PHIL. § 4, at 23–27 (2018).

28. *See id.*

world should be. In other words, the content of any epistemic obligations we have doesn't (and shouldn't) depend on the content of any moral obligations we have.

Others disagree. They believe our epistemic obligations do (and should) depend on our moral obligations. They believe our epistemic obligations neither are (nor should be) independent of our moral obligations. What we're obligated to believe doesn't depend exclusively on how the world is or on our available evidence as to how the world is. It also depends on how the world should be, or on the morally-relevant consequences said to result from forming or not forming this or that belief. In other words, what we're obligated to believe depends, at least in part, on the "moral stakes" involved in believing one way or the other.

Philosophers who endorse this alternative theory believe our moral obligations do or should "encroach" (the word used in the philosophical literature) on our epistemic obligations.[29] This theory can take different forms. An especially strong form might tell us that John is epistemically obligated, under certain conditions, not to believe that *p* even if *p* is true, or if his available evidence supports the formation of the belief that *p*. For example, one might think that even if John's use of deadly force is "in fact" necessary to protect himself against deadly force, or even if his available evidence supports the formation of the belief that the use of deadly force is necessary to protect himself against deadly force, John is nonetheless epistemically obligated to believe he's not about to be killed, just in case that belief and John's consequent use of deadly force would, for instance, perpetuate or foster in some way what has variously been called "institutional racism," "systemic racism," and so forth.

Fortunately, I can set this debate aside for present purposes. Whatever epistemic obligation one believes John has, I will assume for present purposes that John violated it when he formed the belief that *p*. Because John's formation of the belief that *p* violated an assumed epistemic obligation, John's formation of the belief that *p* was, for that reason, wrongful. Again, I make this assumption to reach the second question: was John's wrongful belief that *p* reasonable or unreasonable (non-culpable or culpable) insofar as John would not have formed that belief but for the causal influence of an implicit attitude?

---

29. This thesis is known in the philosophical literature as "moral encroachment" or "practical encroachment," inasmuch as moral or practical reasons are said to "encroach" upon—be relevant to specifying the content of—our epistemic obligations. *See, e.g.*, Renée Jorgensen Bolinger, *Varieties of Moral Encroachment*, 34 PHIL. PERSP. 5 (2020); Rima Basu, *The Specter of Normative Conflict: Does Fairness Require Inaccuracy?*, *in* AN INTRODUCTION TO IMPLICIT BIAS: KNOWLEDGE, JUSTICE AND THE SOCIAL MIND 191, 191 (Erin Beeghly & Alex Madva eds., 2020) ("Moral considerations can change how we epistemically should respond to the evidence."); Deborah Hellman, *The Epistemic Commitments of Nondiscrimination*, *in* 4 OXFORD STUDIES IN PHILOSOPHY OF LAW 156, 168–76 (John Gardner et al. eds., 2021) (discussing "philosophical debate about whether pragmatic and moral considerations properly play a role in what we believe and how we form our beliefs."). The text states a strong form of the encroachment thesis. A weaker form is discussed in Appendix A.II.

2. Questions Three and Four

I will address the second question in Part II, where I offer two conditions for ascribing "reasonableness" or "unreasonableness" to John's (assumedly wrongful) formation of the belief that *p*. If one assumes that John's belief that *p* was not only (epistemically) wrongful, but also unreasonable, then John's plea of self-defense would fail under the generic formulation of self-defense with which we are working. He would therefore be guilty of murder or manslaughter. The third and fourth questions would thus become irrelevant.

In contrast, if one assumes John's belief that *p* was, though (epistemically) wrongful, nonetheless reasonable, we might then ask the third and fourth questions:

> Question 3. Assuming John's formation of the belief that *p* was an epistemic wrong, but that its formation was nonetheless reasonable, did John's choice to kill as a result of having formed that wrongful but reasonable belief constitute a moral wrong (recognized as a legal wrong)? In other words, did his choice to kill violate a moral (legal) obligation governing John's choices and actions?
>
> Question 4. Assuming John's formation of the belief that *p* was an epistemic wrong, but that its formation as nonetheless reasonable, and assuming his choice to kill constituted a moral wrong (recognized as a legal wrong), was his wrongful choice to kill nonetheless excused (nonculpable)?

If John's wrongful belief that *p* was reasonable—as questions three and four both presuppose—John will be acquitted under existing law. The third and fourth questions ask *why* the law acquits him. John isn't guilty of murder (or manslaughter) under existing law, but why not? This is a theoretical question, not a doctrinal one. John will be acquitted under existing doctrine no matter how one answers it.

One might expect the law itself would explain why it acquits someone who, like John, kills having wrongfully but reasonably believed that the use of deadly force was necessary to protect himself against such force. That would be incorrect. The law does not explain itself. The criminal-law literature (but not the criminal law itself) offers two different interpretations of existing doctrine to explain why the doctrine acquits someone who kills having wrongfully but reasonably believed that *p*.

According to one line of thought, a defendant who caused the death of an innocent, but did so only because he reasonably believed that *p*, has done nothing wrong so far as the law is concerned. He believed as a reasonable person would have believed, and acted as a reasonable person (who believed as the defendant reasonably believed) would have acted, inasmuch as he acted in self-defense. The law shouldn't regard causing an innocent's death, under these circumstances, as having violated any legally-recognized moral obligation, and thus the law should regard the defendant's use of deadly force as permissible (not wrongful). Call this the no-wrong explanation for the defendant's acquittal.

According to a competing line of thought, a defendant who caused the death of an innocent has (for that reason alone) done something wrong. His use

of deadly force was therefore impermissible (wrongful). But because he caused the death of an innocent only because he believed as a reasonable person would have believed, the law should regard his wrongdoing as (completely) excused. Call this the wrongful-but-excused explanation for the defendant's acquittal.

The scholarly debate between these two lines of thought is long-standing and on-going (although my impression is that the wrongful-but-excused camp has more adherents nowadays than does the no-wrong camp). Fortunately, because I address only the second question, this debate is one I need not enter for present purposes.

Which brings us back to the second question. Was John's formation of the belief that *p*, now assumed to have been an epistemic wrong, "reasonable" or "unreasonable"? An answer to that question presupposes some account describing the conditions under which "reasonable" or "unreasonable" should be ascribed to John's formation of the belief that *p* for purposes of self-defense law. The next Part offers such an account.

II. BELIEVING "REASONABLY" OR "UNREASONABLY"

This Part offers what might be described as a "theory" of unreasonable belief (and by implication a theory of reasonable belief, or at least a theory of not-unreasonable belief). Assuming John's formation of the belief that *p* (as a result of the causal influence of an implicit attitude) was an epistemic wrong, the theory identifies two conditions the joint satisfaction of which make it permissible for the state to ascribe "unreasonableness" to John's formation of that wrongful belief. Conversely, the failure to satisfy either condition makes it impermissible for the state to ascribe "unreasonableness" to that belief. The two conditions I offer constitute the "insufficient-concern" theory. Defending this theory entails defending the conceptual and normative claims mentioned in the Introduction. (The metaphysical claim mentioned there will be discussed in Part III.)

A. The "Standard" Theory

Before presenting the insufficient-concern theory, I should say a few words about a competing "theory," which is commonly encountered in the criminal-law literature. Call it the "standard" theory. If first-year law students are told anything about how to go about figuring out what makes a belief "reasonable" or "unreasonable" when the criminal law's rules describe a belief in that way, the standard theory is probably what they're told.

The standard theory is mainly an academic theory: a theory found and used mainly in academic writing. I won't describe the law's theory of reasonable belief, which would presumably be found in penal codes or judicial opinions, because those sources of law (so far as I know) don't really contain anything amounting to a well-developed theory of reasonable belief,[30] though some

30. *See, e.g.*, Paul H. Robinson & Lindsay Holcomb, *Individualizing Criminal Law's Justice Judgments: Shortcomings in the Doctrines of Culpability, Mitigation, and Excuse*, 67 VILL. L. REV.

judicial opinions do read as if their authors are following, or at least conforming to, the standard theory's methodology.

When a defendant pleads self-defense, jurors are told they must decide if (among other things) the defendant reasonably believed that *p*. But standard jury instructions, so far as I know, don't tell them what is meant when the word "reasonable" is used. Jurors may be told a defendant's belief that *p* is "reasonable" if a "reasonable person" in the actor's "situation" would have formed that belief, and "unreasonable" if a "reasonable person" in the actor's "situation" would not have formed that belief. But if that's not a tautology, it comes very close. At any rate, it tells one nothing of much, if any, substance. Its application depends entirely on what beliefs a "reasonable" person in the defendant's "situation" would have formed, but says nothing about who that person is.

Enter the standard theory. So far as one can tell, academics have developed this theory, which various Commentaries to the Model Penal Code encourage,[31] ostensibly to help them (and presumably jurors if instructed on the theory) decide if a reasonable person in the actor's situation would have formed the belief that *p*, and thus whether the defendant, having formed the belief that *p*, did so reasonably or not. According to the standard theory, one decides if a defendant's belief was reasonable or not as follows.

> Step 1. Start with an abstract idea or concept known as the "reasonable person."[32]

273, 313 (2022) ("[T]he law has no guiding rules by which such partial-individualization judgments can be made in a given case.").

31. Although many statements in the Code Commentaries appear to adhere to the standard theory (as when they talk about this or that characteristic or feature of the defendant being imputed or not to the reasonable person), other statements, though less frequent, can perhaps be read as at least gesturing in the direction of the insufficient concern theory. *See, e.g.*, MODEL PENAL CODE AND COMMENTARIES § 210.3, at 73 (1985) (describing the jury's function when asked to decide if a "reasonable explanation or excuse" exists for a defendant's "extreme mental or emotional disturbance" as "assessing the relative depravity of the defendant").

32. This abstract idea might represent an epistemic ideal insofar as the reasonable person is someone who always conforms to his or her epistemic obligations. If so, then what a reasonable person—without any facts about the defendant attributed or imputed to the reasonable person—would have believed could function as a test for the *wrongfulness* of any belief the defendant formed. If so, then it presumably could not (or should not) function at the same time as a test for the *culpability* of any belief he formed.

Having said that, perhaps some of the confusion and controversy surrounding the analytical work the reasonable person test is supposed to do is attributable to possibility that the test is indeed being used to perform both functions at the same time. That is, confusion and controversy may arise because the test is being used to establish if a defendant has breached some epistemic obligation governing the beliefs he formed (when no facts about the defendant are to be imputed to the reasonable person), and at the same time to establish if the defendant (assuming he breached some epistemic obligation) has breached that obligation culpably (when some—but not all—facts about the defendant are to be imputed to the reasonable person).

> Step 2. Decide which facts about the defendant to "impute" to this abstraction. (This step is commonly known as "subjectivizing" or "individualizing," to some extent, the reasonable person).

If the reasonable person *without more* (without any subjectification) is an ideal of some undefined sort,[33] then this step presupposes the reasonable person (with some, but not complete, subjectification), i.e., the standard against which the defendant's belief-formation will be judged, falls short of that ideal. The reasonable person emerging from Step 2 may still represent an ideal of some sort, but it must be an imperfect ideal, at least compared to whatever standard of perfection the "unsubjectivized" reasonable person embodies.

Of course, the more facts about, or true of, the defendant one decides for whatever reason to impute to the reasonable person, the more the reasonable person becomes like the defendant himself. At the extreme, if every fact about the defendant is attributed or imputed to the reasonable person, the reasonable person becomes the defendant. The reasonable person and the defendant become one and the same. In that case, the defendant would end up being judged against himself, which would mean the defendant's belief that *p* was necessarily reasonable because the reasonable person just is the defendant, and the defendant believed that *p*.

That being so, self-defense law's reasonable-belief requirement, understood as the standard theory would have us understand it, can serve as a standard for judging or evaluating the defendant's belief that *p* only if some facts, but not all facts, about the defendant are imputed to the reasonable person. Once one has settled on which facts about the defendant, but not all facts about the defendant, are to be imputed to the reasonable person, the analysis proceeds to the next and final step.

> Step 3. Decide what this hypothetical partly-subjectivized reasonable person would have believed and then compare what he or she would have believed to what the defendant believed.

Despite being partly-subjectivized, this imagined person is nonetheless supposed to continue to be a reasonable person. Thus, if this partly-subjectivized, but-still-reasonable, person would have formed the belief that *p*, then the defendant's formation of that same belief would, for that reason, be counted as "reasonable." Conversely, if this partly-subjectivized, but-still-reasonable, person would not have believed that *p*, then the defendant's belief that *p* would, for that reason, count as "unreasonable."

The main problem with the standard theory is well known: It says nothing about what to do at Step 2. It doesn't say anything about which facts about the defendant to impute to the reasonable person and which ones not to impute. Writers tend to say things like: "The reasonable person is a normative ideal, and not just a 'typical' person." Or, "The reasonable person is a normative ideal, and not just an 'average' person." Or, "The reasonable person is a normative ideal, and not just an 'ordinary' person." That's saying something, but not an

33. George Fletcher described, many years ago, the "reasonable person" as "mischievous" and "mythical." GEORGE FLETCHER, RETHINKING CRIMINAL LAW § 4.2.1, at 250 (1978).

awful lot. What, exactly, is "ideal" about this normative ideal? Ideal in what way?

So far as I can tell, the standard theory doesn't say, or doesn't say very much, that's very helpful. It requires imputing some facts about the defendant to the reasonable person, but never offers anything amounting to a principle for deciding which ones get imputed and which ones don't. A skeptic can be forgiven for thinking the reasonable person is, on the standard theory, little more than an empty vessel into which anyone applying it to the facts of a particular case can pour his or her intuitions, which means the theory can lead to whatever verdict those intuitions support. Indeed, that might be the point. The reasonable person standard might be little or nothing more than the law's way of permitting a jury to judge the reasonableness of a defendant's belief according to the collective intuitions of its members, whatever they happen to be, without any meaningful help or guidance from the law.[34]

Perhaps, but I find the standard theory unpersuasive, or at least incomplete. Because the standard theory, at least as I've portrayed it, doesn't say very much about which facts about the defendant should be imputed to the reasonable person, let alone try to identify the conditions under which a belief should be classified as "reasonable" or "unreasonable," the standard theory's reasonable person risks being all things to all people.[35] Yet, insofar as a theory of reasonable belief qualifies as a theory only if it identifies (or at least tries to identify), with some degree of specificity, the conditions under which "reasonableness" or "unreasonableness" should be ascribed to the formation of a belief, the standard theory isn't much of a theory at all. It's perhaps a non-

34. The standard theory, given how it works, can be used to support the following line of thought: Whatever characteristics should be imputed to the reasonable person, one characteristic that should not be imputed is "being a racist." A person who possesses an implicit racial attitude is, for that reason alone, a "racist." If "racism" is not imputed to the reasonable person, as it should not be on this line of thought, then the reasonable person in the actor's situation would not have formed the belief that *p*, because the reasonable person does not have or possess an implicit racial attitude, and that attitude was *ex hypothesi* causally necessary to the formation of the actor's belief that *p*. Therefore, the defendant's belief that *p* was unreasonable. Of course, that line of thought presupposes that "being a racist" can be ascribed to a person just because he possesses an implicit racial attitude.

Someone who simply possesses an implicit racial attitude—as revealed by his performance on some test or task psychologists use to detect the presence of such attitudes—can of course be described as a "racist" just because he possesses such an attitude. But a person to whom "racism" is ascribed just because he possesses an implicit racist attitude is nonetheless different from someone to whom "racism" is ascribed because he harbors ill will, animus, malice and so forth toward the welfare of people who are black, or who's indifferent to their welfare. Describing someone as a "racist" in the latter sense is presumably to issue a stronger or more forceful criticism or condemnation than describing someone as a "racist" in the former sense. For more on this question, see Neil Levy, *Am I a Racist? Implicit Bias and the Ascription of Racism*, 67 PHIL. Q. 534 (2017).

35. *See, e.g.*, LARRY ALEXANDER & KIMBERLY KESSLER FERZAN, CRIME AND CULPABILITY: A THEORY OF CRIMINAL LAW 85 (2009) ("The reasonable person is neither the actual nor the omniscient god, but some construct that lies in between. Because there is no principled way to determine the composition of this construct, punishment for negligence is morally arbitrary.").

theory: a grant of discretion to the jury, subject to appellate review for insufficiency if the jury convicts.

B. The "Insufficient-Concern" Theory

Let me now offer an entirely non-original alternative to the standard theory of reasonable belief.[36] This alternative, which I'll call the insufficient-concern theory, starts with the conceptual claim mentioned in the Introduction; namely, that a reasonable belief is a belief a person has non-culpably formed, and an unreasonable belief is a belief a person has culpably formed. In abbreviated form: a reasonable belief is a non-culpable belief, and an unreasonable belief is a culpable belief.

I hope this claim is uncontroversial. Ascribing "reasonableness" to a belief appears to function in criminal law, or at least can function, as a way of saying that the belief in question, as an element of a statutorily defined offense or defense, was not culpably formed: that the person who formed it is not fairly or properly to be blamed (is not culpable) for having formed it. Likewise, ascribing "unreasonableness" to a belief appears to function in criminal law, or at least can function, as a way of saying that the belief in question was culpably formed: that the person who formed it is fairly or appropriately to be blamed (is culpable) for having formed it. These claims are meant to be conceptual: claims about the meaning of the word "reasonable" when it appears in statutory formulations of, among other legal rules, the rules defining the defense of self-defense. If so, then "reasonable" ("unreasonable") and "non-culpable" ("culpable") can be used interchangeably.

The insufficient-concern theory has two conditions the joint satisfaction of which suffice (I claim) to ascribe culpability to a person for having formed the belief at issue. Because these conditions are individually necessary and jointly sufficient to ascribe culpability, the absence of either condition will defeat the ascription of culpability for having formed that belief. The theory is intended as an all-purpose theory for assessing the culpability of beliefs for purposes of criminal liability, but for present purposes I state the theory's two conditions on the assumption that the belief in question is the belief that *p*, which is the belief at the core of a claim of self-defense. The second condition constitutes the normative claim mentioned in the Introduction.

According to the insufficient-concern theory, an "unreasonable" belief that *p* is a culpably formed belief that *p*, and a belief that *p* was culpably formed if and only if:

> Condition 1. The person who formed the belief that *p* could have believed otherwise than he did (i.e., could have formed the belief that not-*p*) at the time he formed it; and

36. This alternative is entirely non-original because it can be found in, or at least plausibly extracted from, the work (as I read it) of Antony Duff, Ken Simons, and Peter Westen, among others. Larry Alexander and Kim Ferzan remain the staunchest critics of this theory.

> Condition 2. The person who formed the belief that *p* would not have formed that belief but for a lack of sufficient concern for the life of the victim.[37]

If an appeal to some more abstract idea or ideal is needed to make these two conditions more persuasive, it would, I suppose, be "fairness." It would be unfair, the idea would be, for the state to blame someone who believed that *p* if he could not have believed otherwise and, in that sense, lacked the freedom to believe otherwise. If a person could not have believed otherwise than that *p*, then we can say he was "compelled" to believe that *p*; and compulsion is typically taken to preclude the ascription of culpability, for both acts and beliefs.

Likewise, it would be unfair, the idea would be, for the state to blame someone who believed that *p* if his belief that *p* did not result from (and in that sense neither "expressed" nor "manifested") any lack of sufficient concern for the life of the victim. If a person could have believed otherwise than that *p*, but his belief that *p* didn't express any lack of sufficient concern for the life of his victim, his belief shouldn't, in fairness, count as having been culpably formed. If a defendant's formation of the belief that *p* expressed no lack of sufficient concern for the victim's life, then for what, one might ask rhetorically, would the state fairly blame him for having formed that belief?

Condition 1 (as we'll see) will almost always be satisfied. Consequently, it does very little work sorting those who culpably believed that *p* from those who non-culpably believed that *p*. That leaves condition 2 to do almost all the heavy lifting, sorting culpably-formed beliefs from non-culpably formed beliefs. Indeed, I give the theory the name I do because condition 2 is its driving force. But before getting to condition 2, I should say a little about condition 1, and why it will rarely render a person non-culpable for having formed the belief that *p*.

---

37. The second condition derives from, or is at least consistent with, theories of moral responsibility known variously in the philosophical literature as "attributionist" or "quality of will" theories. Generally speaking, these theories claim that responsibility should be ascribed to a choice or belief if the choice or belief expresses or manifests insufficient concern for, or care about, the moral status of another person. I apply the second condition to the facts in John's case below. *See infra* Part III.

I would note here that Michael Brownstein, a philosopher whose work deals almost exclusively with implicit attitudes, has recently "proposed a view of what it means to care about something and what it means for an attitude or action to reflect upon that care . . . [a]nd . . . that in paradigmatic cases, BEIB [the behavioral expression of implicit bias] reflect upon agents' cares." Michael Brownstein, *Attributionism and Moral Responsibility for Implicit Bias*, 7 REV. PHIL. PSYCH. 765, 782 (2016). According to this view, a person is "morally responsible" for an action when it results from the causal influence of an implicit attitude. However, this doesn't mean Brownstein would endorse holding someone like John *criminally* responsible (and thus liable to punishment). On the contrary, Brownstein expressly acknowledges that "responsibility admits of kinds," and the kind in which he's interested isn't the kind associated with criminal liability. *See id.* at 782–83.

1. Condition 1

Condition 1 provides that a person who believed that *p* but who could not have believed otherwise—which is meant to be equivalent to saying the person "lacked the capacity" to believe otherwise than that *p*, or was "compelled" to believe that *p*—is not culpable for having formed the belief that *p*, and therefore, his belief that *p* wasn't "unreasonable."

My analysis of condition 1 starts with an objection to it. According to this objection, condition 1 is, in some sense, "too strong." The objection begins from the premise that at any given moment in time no one has the capacity to believe otherwise than he in fact believes. Or, as it might less precisely but more dramatically be put: No one has the "will" to believe otherwise than he does believe. If I look out my window and see a bird at the feeder, I'll form the belief that a bird is at the feeder, no matter how much I want it to be true (or "will" it to be true) that no bird is at the feeder. Our beliefs respond to, or answer to, the world, not to our will.[38] We can choose to do otherwise than we do,[39] but we can't believe otherwise than we do in fact believe.

If this objection is persuasive (as some believe it is), then no one will ever satisfy condition 1, which is a necessary condition for ascribing culpability to any belief. That would be bad news for the insufficient-concern theory, which purports to be a theory by which to sort culpable beliefs from non-culpable ones. For, if the objection succeeds, all beliefs would get sorted into the non-culpable category. No one would ever be culpable for any belief he formed, which would mean either all beliefs would be allocated to the "reasonable" category, and none to the "unreasonable" category.

Fortunately, the insufficient-concern theory can offer a reply to this objection, as follows. True, we can't "will" to form beliefs in the same way we can and do "will" to form intentions and volitions (which cause our bodies to move). But if one analyzes capacity statements as counterfactual statements (as is common, if not uncontroversial),[40] then sense can be made of the proposition that we do have the capacity to believe otherwise. Indeed, if we analyze capacities into counterfactuals, then far from never being satisfied, condition 1 is almost always satisfied, at least on one specification of the relevant counterfactual. On this specification, human beings almost always have the

38. A person can of course do or fail to do things with the intent to indirectly influence, or with the realization that doing or failing to do something will or might indirectly influence, the beliefs he forms. But this kind of diachronic influence or (if you prefer) control over the beliefs a person forms should be kept distinct from the synchronic influence or control discussed in the text. I address the distinction between diachronic and synchronic influence or control over belief-formation in Appendix C.

39. Accepting this position presupposes, of course, either a libertarian or compatibilist account as to the relationship between "freedom" and "determinism." Some philosophers—usually known as "hard incompatibilists"—believe neither of these accounts is true, and thus believe that we lack the capacity to choose or act otherwise than we in fact do choose or act. One consequence of this belief is that no one is morally responsible for his or her choices or actions, at least insofar as being morally responsible entails being liable to reactive emotions like resentment or indignation.

40. *See, e.g.*, Michael S. Moore, *Compatibilism(s) for Neuroscientists*, *in* LAW AND THE PHILOSOPHY OF ACTION 1, 29 (Enrique Villaneuva ed., 2014).

capacity at any moment in time to believe otherwise than they actually do believe.

To see why, start with the capacity to choose (or act) otherwise, rather with the capacity to believe otherwise. If statements about capacities are, as I will suppose, elliptical counterfactual statements, then when we say a person could have chosen (or acted) otherwise, we mean he would have chosen otherwise in some counterfactual world.

For example, when we say a person who stole $100 in merchandise from the store could have chosen otherwise, we mean he would have chosen otherwise in some yet-to-be specified counterfactual world. The hard and controversial part is to describe the relevant counterfactual. Suppose the relevant counterfactual world is described as one in which the person would suffer an immediate and dramatic consequence if he stole: say, life imprisonment.[41] If we agree the person would not have stolen under those circumstances in this counterfactual world, then we'd say he had the capacity to choose not to steal in the real world in which he did in fact steal. He had the capacity not to steal, but chose not to exercise that capacity. Instead, he chose to exercise his capacity to steal, and so he stole.

What goes for a person's capacity to choose otherwise than he does choose goes (*mutatis mutandis*) for his capacity to believe otherwise than he does believe. Or so I would argue. When we say a person could have believed otherwise—had the "freedom" to believe otherwise, had "control" (in one sense of the word) over his beliefs, and so on—we should be understood (or at least can be understood) to mean he would have believed otherwise in some counterfactual world. Going back to John, who believed that *p* because he believed, among other things, that the victim had a gun in hand when in fact he had an iPhone, condition 1 tells us to ask: did John have the capacity to believe otherwise than that *p*? Answering that question entails providing some description of a counterfactual world in which we test John's capacity to form beliefs. If he passes, then he would have believed otherwise; if he fails, then he couldn't have believed otherwise.

For example, suppose the counterfactual world in which we test John's capacity to form beliefs is one in which he sees an iPhone and not a gun in the victim's hand and thus forms the belief that the victim had an iPhone and not a gun. If in that counterfactual world John would not have formed the belief that *p*—and we have no reason to suppose he would in that counterfactual world have formed the belief that *p*—then we'd say John had the capacity to believe otherwise than that *p* in the actual world in which he did in fact form the belief that *p*. If so, then we'd say he could have believed otherwise than that *p* at the moment he formed the belief that *p*, which would mean his formation of the belief that *p* in the actual world satisfies condition 1.[42]

---

41. I don't explain in the text why I find this counterfactual world a plausible one to choose, but I do discuss why it might be plausible in STEPHEN P. GARVEY, GUILTY ACTS, GUILTY MINDS 127–31 (2020).

42. Although I believe most people most of the time have the capacity to choose otherwise and to believe otherwise than they do choose and believe, I also believe the experience or

If this analysis of condition 1 is correct, then the only people who will fail to satisfy it are those who were "compelled" to believe that *p*. But the only people who are truly compelled to believe that *p* are people who would continue to believe that *p* even if they had what would be described as "compelling evidence" available to them at the time they formed the belief that *p* that *p* was false. Such a belief (i.e., a belief that persists even when a person has available to him compelling contrary evidence) can fairly be called a delusion. If that's so, then very few people will fail to satisfy condition 1. Not many people delusionally believe the use of deadly force is necessary to protect themselves against deadly force, and in any event, the legal avenue for relief in those rare cases would likely be a plea sounding in insanity, not self-defense.

2. Condition 2

That brings us to the main event: condition 2. Whereas condition 1 asks if a person *could have* believed otherwise than he in fact believed, condition 2 asks *why* he believed as he in fact believed.

If we ask a person why he believed as he did, chances are he'd reply with an account or recitation of the "evidence" available to him supporting his belief. Any such account or recitation would be unlikely to refer to his wants or desires. We don't usually say we believe something because we want to believe it. But why we believe what we believe, and why we say we believe what we believe, are not necessarily the same thing. For an agent who cares only about believing that which is true (think, perhaps, Spock),[43] the evidence available to him at the time he formed a belief would presumably be the only thing in fact causing him to form the beliefs he formed. The only desire figuring into any such causal explanation would be a desire to form true beliefs and an aversion to forming false ones.

Mere mortals aren't like Spock. The beliefs we form depend not only on our available evidence, but on our motivations, desires, aversions, intentions, and other such mental states, besides (one hopes) a desire for truth and an aversion to falsehood. Philosophers typically call these kinds of mental states conative or motivational states. These states reveal what a person cares about (what his concerns are). The overall configuration of a person's motivating cares or concerns (which some people call his "will") can be said to represent

---

phenomenology involved when those capacities are exercised differ from one another. When a person exercises his or her capacity to choose, he or she almost always, though not always, experiences a sense of agency: he or she experiences himself or herself as the source or agent doing the choosing, or as being "in control" of the choice he or she makes. In contrast, when a person exercises his or her capacity to believe, he or she typically does not experience himself or herself as the source or agent doing the believing, or as being "in control" of the beliefs he or she forms. Instead, we experience beliefs as "just happening" to us, or as "coming to us unbidden," and so forth. This difference may explain why some people believe, incorrectly in my view, that we lack the capacity to believe otherwise. For a more extended discussion on what I mean when I refer to a "sense of agency," see GARVEY, *supra* note 41, at 233–35.

43. Spock was (if memory serves) part-Vulcan and part-human. For present purposes I assume he was at least Vulcan enough such that no desire other than the desire to form true beliefs and an aversion to forming false ones causally influenced the beliefs he formed.

his attitude (what some people call his "quality of will, as in "ill will," "indifference," or "good will.") When a person cares not at all about someone or something, or less for someone or something than he should care, his will is configured with "insufficient concern" towards that someone or something.

A person's cares and concerns (among other things) cause or motivate the choices he makes and the actions he performs (thereby "manifesting" in those choices the quality of his will at the moment of choice). They can also cause or motivate the beliefs he forms (thereby "manifesting" in those beliefs the quality of his will at the moment he formed them). For example, if a person doesn't care about the welfare of others, he's less apt to believe a person is in distress or jeopardy, compared to someone who does care about others. A person's cares and concerns can (in an assortment of different ways) alter the causal force his available evidence would otherwise have had on the beliefs he forms. Among other things, they can cause some evidence to become more salient than it would otherwise have been, and other evidence to become less salient, thereby causing beliefs to form when they otherwise would not have, or not to form when they otherwise would have.

Pursuant to condition 2, a person who forms the belief that *p* is culpable for having formed that belief (assuming condition 1 is also satisfied) if that belief was the result of "insufficient concern" for the life of the victim. Spelling that proposition out in more detail, a person who forms the belief that *p* is culpable for having formed it if and because his formation of it was a result of (and thereby "expressed" or "manifested") insufficient concern for the victim's life, not directly in any choice he made (because he used deadly force only because he believed its use was necessary), but indirectly in the beliefs he formed.

The "test" for such "insufficient concern," indirectly expressed or manifested in the beliefs a person forms, is a counterfactual inquiry:

> Holding everything about the defendant and his situation constant at the time he formed the belief that *p*—i.e., taking the defendant and his situation as they were at the time he formed the belief that *p*—would he have formed the belief that *p* if he'd had sufficient concern for the life of his victim?

This test flips on its head the question the standard theory asks. The standard theory starts with an abstraction called the "reasonable person" and asks which particular facts about the defendant should be "imputed" to that abstraction—without providing any principled guidance as to how to go about that task. The test associated with the insufficient-concern theory, in contrast, starts with the defendant (as he is) and then asks the above-described counterfactual question: Would the defendant (as he otherwise is) have formed the belief that *p* if he'd had sufficient concern? If yes, then the belief that *p* wasn't the result of insufficient concern (and was thus non-culpably formed, and was thus not unreasonable). If no, then the belief that *p* was the result of insufficient concern (and was thus culpably formed, and thus unreasonable).

III. IMPLICIT RACIAL ATTITUDES

The last task is to apply the insufficient-concern theory and its associated counterfactual test to the facts described at the outset, in which John formed the belief that *p*—that the use of deadly force was necessary to protect himself against deadly force—but (by stipulation) would not have formed that belief but for the causal influence of an implicit racial attitude.[44]

I apply the theory step-by-step to the facts in John's case as follows. I eventually reach the crux of the matter in Step 5.

> Step 1. Start with John (subject to the stipulations in Steps 2 through 4) and his situation as they were at the moment he chose to exercise deadly force, which is presumably the moment he formed the belief that *p*.

This opening description of John and his situation is co-extensive with all the facts about John and his situation at that moment he formed the belief that *p*. Put another way: all the facts about John and his situation are held constant or fixed. This step is important. The insufficient-concern theory rests on the thesis that a defendant's formation of the belief that *p* is culpable only if it resulted from insufficient concern, not from any other fact or feature of the defendant or his situation. Everything about the defendant and his situation is therefore held constant—at the outset. (The only change made to the John's folk psychology will come in Step 5).

> Step 2. Stipulate that John is not a "racist," by which I mean that he does not in any way believe (e.g., consciously or unconsciously, occurrently or dispositionally) that any race-based generalization is true. His formation of the belief that *p* therefore was not and could not have been the result of any such belief.

This stipulation is important. It tells us to assume, for purposes of Step 1, that John, at the time he formed the belief that *p*, did not in any way believe any race-based generalization (which some might describe as a "stereotype"), e.g., he does not believe in any way that black people as a group or category are any

---

44. The literature discussing the conditions under which a person is morally responsible for a wrongful action resulting from an implicit attitude (though not necessarily criminally responsible, insofar as the conditions for ascribing criminal responsibility differ from those for ascribing moral responsibility) is large and growing. *See, e.g.*, Noel Dominguez, *Moral Responsibility for Implicit Biases: Examining the Options*, *in* AN INTRODUCTION TO IMPLICIT BIAS: KNOWLEDGE, JUSTICE, AND THE SOCIAL MIND (Erin Beeghly & Alex Madva eds., 2020); Luc Faucher, *Revisionism and Moral Responsibility for Implicit Attitudes*, *in* 2 IMPLICIT BIAS & PHILOSOPHY: MORAL RESPONSIBILITY, STRUCTURAL INJUSTICE, AND ETHICS 115 (Michael Brownstein & Jennifer Saul eds., 2016); Joshua Glasgow, *Alienation and Responsibility*, *in* 2 IMPLICIT BIAS & PHILOSOPHY, *supra*, at 37; Jules Holroyd, *Responsibility for Implicit Bias*, 43 J. SOC. PHIL. 274 (2012); Neil Levy, *Consciousness, Implicit Attitudes and Moral Responsibility*, 48 NOUS 21 (2014); Neil Levy, *Implicit Bias and Moral Responsibility: Probing the Data*, 94 PHIL. & PHENOMENOLOGICAL RSCH. 3 (2017); Natalia Washington & Daniel Kelly, *Who's Responsible for This? Moral Responsibility, Externalism, and Knowledge About Implicit Bias*, *in* 2 IMPLICIT BIAS & PHILOSOPHY, *supra*, at 11; Robin Zheng, *Attributability, Accountability, and Implicit Bias*, *in* 2 IMPLICIT BIAS & PHILOSOPHY, *supra*, at 62.

more dangerous or violent than white people as a group.[45] Whatever the causal antecedents of John's belief that *p*, not included among them was any belief the content of which was a race-based generalization.

This stipulation rests on an assumption about the nature (metaphysics) of implicit attitudes. Because the nature of implicit attitudes is, so far as I can tell, contested, I avoid to the extent possible making any assumptions as to what implicit attitudes are and what they are not. Having said that, I do make the stipulation described Step 2, which presupposes that whatever an implicit racial attitude turns out to be, an implicit racial attitude is not an unconscious belief the content of which is a race-based generalization.

I make that assumption for the following simple reason. If an implicit racial attitude is nothing more than an unconscious belief the content of which is a race-based generalization (as a good deal of the legal literature seems to me to suppose), then implicit racial attitudes are (so far as I can tell) nothing new. Yet, in part because of all the attention implicit attitudes have received, inside the academy and out, I assume that implicit attitudes must be *something* new; specifically, that they must be something new insofar as they must in some way be metaphysically distinct from unconscious beliefs the content of which is a race-based generalization. If not, if "implicit racial attitude" is just another name for an unconscious belief the content of which is some race-based generalization, then an implicit racial attitude is something much less novel than I am, for present purposes, supposing it to be.

For example, Charles Lawrence famously wrote about "unconscious discrimination"—presumably meaning (at least in part) unconscious beliefs the content of which was a race-based generalization—in 1987.[46] That was well before 1995. The year 1995 is noteworthy because, according to one recent review of the literature, that was the year in which Anthony Greenwald and

---

45. One often-cited article on race and self-defense limits its discussion to cases in which the defendant formed the belief that *p* as a result of a background belief the content of which is a race-based generalization, not an implicit racial attitude. *See* Jody D. Armour, *Race Ipsa Loquitur: Of Reasonable Racists, Intelligent Bayesians, and Involuntary Negrophobes*, 46 STAN. L. REV. 781 (1994). As noted in the text, a case involving a defendant who formed the belief that *p* as a result of a belief the content of which is a race-based generalization would raise a question different from the one addressed here.

A large literature exists on what might be called the "morality of stereotyping," far too much to collect in a footnote, with different authors using the word "stereotype" to mean different things. Having said that, the reader might find the following helpful places to start: Erin Beeghly, *What is a Stereotype? What is Stereotyping?*, 30 HYPATIA 675 (2015); Lawrence Blum, *Stereotypes and Stereotyping: A Moral Analysis*, 33 PHIL. PAPERS 251 (2004); Renée Jorgensen Bolinger, *The Rational Impermissibility of Accepting (Some) Racial Generalizations*, SYNTHESE (forthcoming) (published online May 23, 2018); Katherine Puddifoot, *Stereotyping: The Multifactorial View*, 45 PHIL. TOPICS 137 (2017). For a discussion of "demographic statistics," specifically as they relate to the use of defensive force, see Renée Jorgensen Bolinger, *Demographic Statistics in Defensive Decisions*, 198 SYNTHESE 4833 (2021).

46. Charles R. Lawrence III, *The Id, the Ego, and Equal Protection: Reckoning with Unconscious Racism*, 39 STAN. L. REV. 317, 322 (1987) ("[A] large part of the behavior that produces racial discrimination is influenced by unconscious racial motivation," where "motivation" includes "beliefs, desires, and wishes").

Mahzarin Banaji published "Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes" in the journal *Psychological Review*, which has been said to mark the start of the "implicit revolution."[47] That revolution has resulted (as of the end of 2018) in over 1,485 peer-reviewed articles in the PsycNET database in which the words "implicit," "social" and "cognition" all appear in the indexing fields.[48] But if "implicit attitude" is just a synonym for "unconscious belief," then the existence of implicit attitudes has been known and talked about long before the "implicit revolution"—at least since Freud, probably before.[49]

Of course, if implicit attitudes turn out to be the same thing as unconscious beliefs, then the question I should be asking would be different from the one I am asking. I should be asking about the culpability of a person for killing another because he formed the belief that *p* when that belief resulted from another belief (a conscious or unconscious belief the content of which is a race-based generalization).[50] I would no longer be asking the question I am asking, which is about the culpability of a person for killing another because he formed the belief that *p* when that belief resulted from an implicit racial attitude.[51]

> Step 3. Stipulate that John believed that *p* and would not have chosen to use deadly force but for having formed that belief.
>
> Step 4. Stipulate that John would not have formed the belief that *p* but for the causal influence of an implicit racial attitude.

Psychologists have developed several different tasks designed to detect the presence of implicit attitudes, with the Implicit Association Test (IAT)

---

47. Greenwald & Lai, *supra* note 7, at 420; *see also* Anthony G. Greenwald & Mahzarin R. Banaji, *The Implicit Revolution: Reconceiving the Relation Between Conscious and Unconscious*, 72 AM. PSYCH. 861, 865 (2017).

48. *See* Greenwald & Lai, *supra* note 7, at 420.

49. *See, e.g.*, Michael S. Moore, *Responsibility and the Unconscious*, 53 S. CAL. L. REV. 1563, 1563 (1980) ("Even before Freud, novelists, historians, lawyers, and others have long felt free to disregard an actor's apparently sincere avowal of his reasons for a given action, and to divine within his unconscious the 'real' reasons for it."). Lawrence also invokes Freud: "Freudian theory states that the human mind defends itself against the discomfort of guilt by denying or refusing to recognize those ideas, wishes, and beliefs that conflict with what the individual has learned is good and right." Lawrence, *supra* note 46, at 322.

50. I discuss a case in which a defendant who asserts self-defense to a charge of murder is stipulated to have formed the belief that *p* as a result of a belief the content of which is a race-based generalization in Stephen P. Garvey, *Self-Defense and the Mistaken Racist*, 11 NEW CRIM. L. REV. 119, 149–54 (2008). Because the analysis there offered did not rely on the insufficient concern theory, which I'm presently persuaded is the best theory to date by which to determine the reasonableness of a statutorily relevant belief, the analysis I'd offer now of the case there discussed would no doubt differ from the analysis I then offered.

51. Suppose a person forms the belief that *p* in the following way: 1) An implicit racial attitude is "activated"; 2) this activated implicit attitude causes the person to form a belief the content of which is a race-based generalization; 3) this belief in turn causes the formation of the belief that *p*; and 4) without the "activation" of the implicit racial attitude, the belief the content of which is a race-based generalization would not in turn have caused the formation of the belief that *p*. Would this more complicated causal chain between the implicit racial attitude and the belief that *p* (with a belief the content of which is a race-based generalization intervening between the two) change the analysis offered in Part III? Not so far as I can see.

probably being the most well-known.[52] But the test most relevant for present purposes is probably the test used in what are known as "shooter studies." I trust the reasons why this test is most relevant for present purposes needs no elaboration.

Some (many) readers may believe these studies support the inference that (all else being equal) people (or at least study participants) are on average more likely to shoot (or more likely to shoot more quickly, and so on) when the person they see is black compared to when the person they see is white. My review of the literature suggests a more complicated picture, or at one less unequivocal. So far as I can tell, the literature contains claims a non-expert (and maybe even an expert) will find difficult to reconcile.[53] Nonetheless, for present purposes I stipulate that John believed that *p* due to the causal influence of an implicit attitude, and would not have believed that *p* without that causal influence.

With that stipulation, we come at last to the crux of the matter. The next step is the crux because it asks the counterfactual question on which John's culpability for having formed the belief that *p* turns, according to the insufficient-concern theory.

> Step 5. Would John have formed the belief that *p* if he'd had sufficient concern for the life of the victim?

That, of course, isn't an easy question to answer. For one thing, although the question might be understood to presuppose that John lacked sufficient concern, maybe that wasn't so. Maybe John had sufficient concern, yet formed the belief that *p* anyway. Or, assuming John lacked sufficient concern, maybe such concern, if he'd had it, wouldn't have made any difference. Maybe he would, once again, have formed the belief that *p* anyway. Or, on the contrary, maybe such concern would have made a difference. Maybe, had John had it, such concern would somehow have managed to neutralize the causal influence attributable to the implicit racial attitude he's stipulated to have had, such that he wouldn't have formed the belief that *p*.

Of course, asking if John would or would not have formed the belief that *p* if he'd had sufficient concern presupposes something about the metaphysics of implicit attitudes. It presupposes that an implicit attitude's causal force *can* be neutralized by sufficient concern. But is that true? Are implicit attitudes the type of thing sufficient concern is capable of rendering causally inert? Can "good will" (i.e., sufficient concern) neutralize the causal influence an implicit attitude would otherwise exercise on the beliefs a person forms, or do implicit attitudes exercise their causal influence no matter what the quality of a person's will?

---

52. So far as one can tell, psychologists disagree about what exactly the "score" a person obtains on the IAT tells the person about him- or herself. For example, a recent article concludes with this sentence: "Conflating implicit processes in the measurement of implicit attitudes with implicit personality constructs has created a lot of confusion. It is time to end this confusion. The IAT is an implicit measure of attitudes with varying degrees of validity. It is not a window into people's unconscious feelings, cognitions, or attitudes." Ulrich Schimmack, *The Implicit Association Test: A Method in Search of a Construct*, 16 PERSPS. ON PSYCH. SCI. 396, 412 (2021).

53. *See* Appendix D.

The amount of data psychologists have gathered using tools like the IAT, and the number of publications reporting that data, is breathtaking (and difficult for anyone whose professional life is not dedicated to the subject to master and fully understand). Different people (mainly professional philosophers) looking at this data have offered different theories meant to capture or describe what they believe is the unique metaphysical profile or status of implicit attitudes. These theories explain how (if at all) the metaphysics of implicit attitudes differ from the metaphysics of folk psychological mental states more familiar to the criminal law, like beliefs, desires, and intentions.[54] So far as I can tell, no single account of the metaphysics of implicit attitudes has managed to achieve a consensus among philosophers.[55]

According to one account, implicit attitudes are nothing more than unconscious beliefs the content of which is some generalization. I've already mentioned this possibility and have offered my reasons for setting it aside.[56] Most accounts likewise presuppose that the phenomenon reflected in data gathered from tests like the IAT is metaphysically distinct from (conscious or unconscious) beliefs, desires, or intentions. They presuppose that implicit attitudes are *sui generis* and as such deserve a name all their own to distinguish them from beliefs, desires, and intentions. Philosophers have given various names to the thing psychologists have discovered (whatever names psychologists have given it). These names include: "associations,"[57] "aliefs,"[58] "in-between beliefs,"[59] "patchy endorsements,"[60] and "mental imagery."[61] Nor should one rule out the possibility that the thing psychologist have discovered isn't any one thing at all. Maybe the thing is "really" a bunch of different things.[62]

---

54. *See generally* Daniel Hutto & Ian Ravenscroft, *Folk Psychology as a Theory*, STAN. ENCYCLOPEDIA PHIL. (2021).

55. Psychologists offer their own accounts using concepts and categories, where those concepts and categories are presumably useful to doing whatever it is that they understand themselves to be doing. *See, e.g.*, Michael Brownstein, *Implicit Bias*, STAN. ENCYCLOPEDIA PHIL. §§ 2.1–2.2, at 13–15 (2019) (discussing psychological "models" with names like MODE ("Motivation and Opportunity as Determinants"), MCM ("Meta-Cognitive Model"), RIM ("Reflective-Impulsive Model"), and APE ("Associative-Propositional Evaluation")).

56. Having said that, I note that the work of one philosopher is associated with the thesis that implicit attitudes are beliefs. *See* Eric Mandelbaum, *Attitude, Inference, Association: On the Propositional Structure of Implicit Bias*, 50 NOUS 629 (2016).

57. The IAT (Implicit Association Test) is a test designed to detect "*associations"* between concepts. A common analogy is the "association" most people make between salt and pepper.

58. Tamar Gendler, *Alief and Belief*, 105 J. PHIL. 634 (2008).

59. Eric Schwitzgebel, *In-Between Believing*, 51 PHIL. Q. 76 (2001).

60. Neil Levy, *Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements*, 48 NOUS 800 (2015).

61. Bence Nanay, *Implicit Bias as Mental Imagery*, 7 J. AM. PHIL. ASSOC. 329 (2021).

62. Jules Holroyd et al., *What is Implicit Bias?*, 12 PHIL. COMPASS 1, 13 (2017) (suggesting the possibility that "there is no such psychological kind and therefore no such account that attempts to characterize implicit bias as a particular mental state or psychological kind will succeed").

For present purposes I need not enter this debate. I don't need to know which of these metaphysical accounts best explains all the data on implicit attitudes. All I need to know for now is whether or not implicit attitudes, whatever else is true of them, are in principle responsive to or sensitive to cares or concerns, such that sufficient concern is capable of neutralizing the causal effect an implicit attitude would otherwise have had on the beliefs a person forms. Are implicit attitudes the kind of thing whose causal force sufficient concern or goodwill can defeat, or at least diminish, or do implicit attitudes exercise whatever causal force they have whatever concern, or lack thereof, a person has for the lives of others?

At this point the reader might expect an extended discussion of the psychological literature, undertaken in the hope of unearthing an answer to this question. Alas, any such undertaking would require far more familiarity with the empirical literature, and a far greater capacity truly to understand that literature, and how it relates to the folk psychological concepts with which the criminal law formulates its rules of liability, than I can honestly say I have. Empirical study may be the only way mere mortals have to discover truths about the natural world, of which I've been assuming human beings (and their implicit attitudes) are a part. Yet broad and unequivocal claims about what "studies show," claims that are common in law review articles, should make skeptical minds skittish, especially when the study's object is itself the human mind, and more so when the field of study is as young (in the grand scheme of things) as the study of what might be called the implicit human mind.[63]

I nonetheless made a good faith attempt to review the relevant literature (as of February 2021)[64] in search of a clear-cut, definitive, or even consensus answer. Perhaps the sought-after answer is sitting somewhere in plain view, but if it is, I missed it. I therefore have no such answer, based on the work of social psychologists, to report or defend. Having said that, I would enter two observations bearing on the question, which is (to repeat): are implicit attitudes the kind of thing whose causal force sufficient concern or "goodwill" can defeat or diminish?

First, one can surely find empirical studies saying things like those who "espouse egalitarian values" do better on the IAT, or those with "implicit motivation to control prejudice" do better on the shooter task. Moreover, one could fairly infer from such statements that if a person is a "good" person—here understood as a person with sufficient concern for others—then even if he harbors an implicit racial attitude, that attitude won't cause him to form other beliefs, or would be less likely to cause him to form other beliefs, like the belief that *p*. In other words, one can find studies apparently supporting an affirmative answer to the metaphysical question to which I seek an answer.

---

63. *See generally* THE POLITICS OF SOCIAL PSYCHOLOGY (Jarret T. Crawford & Lee Jussim eds., 2018); Gregory Mitchell & Philip E. Tetlock, *Popularity as a Poor Proxy for Utility: The Case of Implicit Prejudice*, *in* PSYCHOLOGICAL SCIENCE UNDER SCRUTINY: RECENT CHALLENGES AND PROPOSED SOLUTIONS 164 (Scott O. Lilienfeld & Irwin D. Waldman eds., 2017).

64. The article was completed in early 2021. The manuscript was sent to 212 law reviews at various points in time thereafter, but was only accepted for publication, thanks to the editors of the *Notre Dame Journal of Law, Ethics and Public Policy*, in April 2022.

For example, one article from which this inference might be drawn has (as of October 1, 2020) been cited in thirty-one law-review articles contained in Westlaw's "Secondary Sources—Law Reviews & Journals" database. The article, Jack Glaser & Eric D. Knowles, "Implicit Motivation to Control Prejudice," 44 *Journal of Experimental Social Psychology* 44 (2008), has been cited for the following propositions (among others):

> "[P]eople who are highly motivated to behave fairly in interracial interactions or are primed with situational reminders of their antiracist values tend not to act upon their biases."[65]
>
> "Once . . . biases are identified to members of the public, it then creates an implicit motivation to control prejudice."[66]
>
> "[R]are individuals with a white preference on the IAT and who are highly motivated to control prejudice were able to avoid the shooter bias."[67]

I won't recapitulate or try to evaluate the above-mentioned study's methodology. I'm not competent to do so in any event. I note only that the study deployed a shooter simulation, and as such, its results might be thought directly relevant to the question at hand, which is another reason, beyond its apparent influence among legal academics, for spending a few moments discussing it.

The prudent reader can and should consult and evaluate the article's methodology for himself or herself. All I do here is highlight one sentence from the article's "Conclusion," which is the place (I would guess) many non-experts turn (first, if not only) to discover what the study claims to have found. The sentence states: "The evidence indicates that those high in an implicit negative attitude toward prejudice show less influence of implicit stereotypes on automatic discrimination."[68] Again, I can see how someone might infer from a sentence like this that a "good" person—a person with sufficient concern for others—won't form the belief that *p* as a result of an implicit racial attitude, assuming he possesses such an attitude.

But I would hesitate to draw that inference. The reason is simple. The sense in which the Glaser-Knowles article uses the word "motivation" in its title differs from the sense of the word "motivation" as that word figures into the insufficient-concern theory described above. The study described in the Glaser-Knowles article uses the results from an IAT to "operationalize" the concept of motivation. The study therefore uses the word "motivation," so far as I can tell, to refer to the causal effect of one implicit attitude (an "implicit negative attitude toward prejudice") on another implicit attitude ("implicit racial attitudes"),

---

65. Kim Shayo Buchanan & Phillip Atiba Goff, *Racist Stereotype Threat in Civil Rights Law*, 67 UCLA L. REV. 316, 369 (2020).

66. David Fontana & Donald Braman, *Judicial Backlash or Just Backlash? Evidence From a National Experiment*, 112 COLUM. L. REV. 731, 779 (2012).

67. Jeffrey J. Rachlinski et al., *Does Unconscious Racial Bias Affect Trial Judges*?, 84 NOTRE DAME L. REV. 1195, 1204 (2009).

68. Jack Glaser & Eric D. Knowles, *Implicit Motivation to Control Prejudice*, 44 J. EXPERIMENTAL SOC. PSYCH. 164, 171 (2008).

whereas the insufficient-concern theory uses the word "motivation" to refer to the causal effect of some configuration of desires and other conative states on an implicit attitude.

The desires and aversions (and conative states more generally) that determine the quality of a person's will at any given moment in time don't belong in the same metaphysical camp to which implicit attitudes belong. The metaphysics of implicit attitudes (whatever they turn out to be) presumably differ from the metaphysics of desires (and motivational or conative states more generally), just like the metaphysics of implicit attitudes presumably differ from the metaphysics of beliefs (and cognitive states more generally). Just as I've assumed implicit attitudes differ metaphysically from unconscious beliefs, I likewise assume they differ metaphysically from unconscious desires. Consequently, if one concluded from the Glaser-Knowles study that goodwill (sufficient concern) can neutralize the causal effect of an implicit attitude, one would risk drawing a conclusion about apples from a premise about oranges.

Second, I note the title of a 2013 book on implicit attitudes: *Blindspot: Hidden Biases of Good People*.[69] That book was a *New York Times* bestseller, which once had its own website.[70] I highlight the title's use of the word "good." According to the book's authors (Mahzarin R. Banaji and Anthony G. Greenwald), the "'GOOD PEOPLE' of [the] book's title are people who, along with other good traits, have no conscious race preferences," but "nevertheless obtain an automatic White preference result on the Race IAT."[71] The question I ask is this: is the class of "good people" so defined co-extensive with the class of people having sufficient concern for others?

I don't know. But if the two classes are indeed co-extensive, and if having sufficient concern isn't sufficient to prevent "good people" from performing as they do on the IAT, then presumably neither would it be sufficient to prevent whatever implicit racial attitudes they do have from causally influencing the formation of the belief that *p*. Of course, I could be wrong. Something may get lost in translation from the concepts and categories psychologists like Banaji and Greenwald use to the folk-psychological concepts the insufficient-concern theory (and the criminal law more generally) uses. Maybe Banaji and Greenwald have in mind people who are in some sense "good" despite their lack of concern for others. Again, I don't know.

CONCLUSION

Michael Brownstein is a philosopher, but so far as one can tell from his published work, he understands the psychological literature on implicit attitudes as well as anyone.[72] Writing in a book chapter published in 2020, he observed

---

69. MAHZARIN R. BANAJI & ANTHONY G. GREENWALD, BLINDSPOT: HIDDEN BIASES OF GOOD PEOPLE (2013).

70. HARVARD, https://web.archive.org/web/20220418015119/http://blindspot.fas.harvard.edu/.

71. BANAJI & GREENWALD, *supra* note 69, at 158.

72. Brownstein is the author of a 2018 monograph entitled *The Implicit Mind: Cognitive Architecture, the Self, and Ethics* (2018).

that the "ongoing refinement of theories of the nature of implicit bias, and the ongoing debate about the proper explanatory scope of tools like the IAT, is simply the regular process of (often slow) scientific progress at work."[73] That looks like a fair-minded statement about the nature of progress in science, including the science of the mind. Brownstein goes on to say: "[u]ndoubtedly, there are challenges, fundamental open questions, problematic assumptions, and conflicting data in research on implicit social cognition."[74]

Anthony Greenwald is a psychologist. Along with Mahzarin Banaji, Greenwald has been credited with founding the modern-day study of implicit attitudes.[75] The abstract of an article Greenwald (together with Calvin Lai) published in the 2020 *Annual Review of Psychology*, entitled "Implicit Social Cognition," states: "Although this rapidly growing body of research [on implicit social cognition] offers prospects of useful societal applications, the theory needed to confidently guide those applications remains insufficiently developed."[76]

Observations such as these, which highlight "open questions," "problematic assumptions," "conflicting data," and "insufficiently developed" theory, leave one wondering. As the science of implicit attitudes develops, what should those in charge of the criminal law—the institution by and through which the state condemns and punishes—do when faced with proposed reforms of the substantive criminal law predicated on how that science tells us the human mind works? Should they vote to enact them into law? More specifically, should they vote to enact the Proposal into law?

Reasonable minds are apt to disagree. One might believe that anyone who believes he's about to be killed, but would not have formed that belief but for an implicit racial attitude, has for that reason and that reason alone formed that belief unreasonably. Or, one might believe the science of implicit attitudes has progressed far enough to believe that sufficient concern (as that concept is used in the insufficient-concern theory) is indeed capable of neutralizing the causal effect an implicit attitude would otherwise have on the beliefs a person forms. Or, one might even believe the state is permitted to punish a person who forms the belief that *p* even if that belief wasn't culpably formed. Any of these beliefs would give one reason to support enacting the Proposal into law.

My conclusion is more cautious. For reasons given above, I believe we should not assume that a person who believes he's about to be killed, but would not have formed that belief but for an implicit racial attitude, has necessarily formed that belief unreasonably (culpably). I decline that assumption for two related reasons. First, the insufficient-concern theory, as a way of distinguishing

---

73. Brownstein, *supra* note 5, at 68.

74. *Id.* at 71.

75. For Greenwald, the word "implicit," which has carried different meanings in the psychology literature, should be understood as meaning nothing more than "indirectly-measured." Greenwald believes that this definition would avoid any "theoretical stances" on the "debated terms (e.g., implicit, unconscious, and automatic)." Greenwald & Lai, *supra* note 7, at 421 ("Defining 'implicit' as meaning 'indirectly measured' is therefore a theory-uncommitted definition, allowing research to proceed without need for debate about the conceptual understanding of 'implicit.'").

76. *See id.* at 419.

reasonable beliefs from unreasonable ones, is the theory I (presently) find most persuasive. Second, I don't believe the science of implicit attitudes, in its present state, can confidently say that anyone who believes he's about to be killed, but who forms that belief only as a result of an implicit racial attitude, has necessarily manifested insufficient concern for the victim's life.

The Proposal limits the scope of self-defense and thereby expands (however marginally) the scope of criminal liability. Any reform expanding the criminal law should to my mind bear the burden of proof. Because I don't believe that burden has yet been carried, I would urge lawmakers to table the Proposal.

APPENDIX A — PRIOR SCHOLARSHIP

I group prior scholarship on race and self-defense into two categories. Scholarship in the first group addresses a question similar to the one raised in the text: what result should a finder of fact reach when applying existing self-defense doctrine to a set of facts in which a defendant forms a false belief that *p* as a result of an implicit racial attitude?

Scholarship in the second group addresses a different question: should the law of self-defense be changed, and if so how, given a range of costs and consequences said to arise when a defendant kills a black person and successfully interposes a claim of self-defense? Scholarship in this group proposes reforms to the law of self-defense, reforms that narrow the conditions under which the law would afford a right or permission to use deadly force in self-defense, with the aim of reducing or avoiding those costs and consequences.

I. Applying Existing Doctrine

Only two articles (so far as I know) directly address a question similar to the one asked here. I would frame the question those articles address as follows: when a person forms the belief that *p*, but would not have formed that but for the causal influence of an implicit racial attitude, what should the verdict be?

The first article, published in 2012, is L. Song Richardson & Phillip Atiba Goff, "Self-Defense and the Suspicion Heuristic," 98 *Iowa L. Rev.* 293 (2012). The second article is Jules Holroyd & Federico Picinali, "Implicit Bias, Self-Defense, and the Reasonable Person," in *The Criminal Law's Person* 167 (Claes Lernestedt & Matt Matravers ed., 2022).

Both articles emphasize a range of consequences said to result if and when others believe a person who believed that *p*, who formed that belief as a result of an implicit racial attitude, and who kills an innocent black person because he believed that *p*, is acquitted on grounds of self-defense. These consequences play a central role in the argument the articles make on behalf of the conclusion they propose. I set forth that argument (as I understand it) step-by-step as follows:

> 1. A belief that *p* resulting from an implicit racial attitude is unreasonable just because a reasonable person does not have or possess any such attitude.[77]

77. Richardson and Goff write: "[M]istakes based upon the suspicion heuristic would be unreasonable *because* the ideal person would not have or act on these mistaken judgments." Richardson & Goff, *supra* note 8, at 320 (emphasis added). What Richardson and Goff describe as the "suspicion heuristic" seems to me materially indistinguishable from what I'm calling an implicit racial attitude. Two comments: first, so far as one can tell, the argument for this claim relies on what I've referred to in the text as the standard theory, i.e., it relies on the standard theory as the way in which to decide if a statutorily relevant belief is "reasonable" or "unreasonable." The text explains why I find that approach unpersuasive. Second, one might resist the claim that an "ideal" person "would not . . . act on" the basis of a mistaken judgment that the use of deadly force was necessary to defend against deadly force. Would any person, ideal or not, "not act" on a "mistaken judgment"

2. Under existing doctrine, a person who unreasonably believes that *p* will be convicted of murder (if the jurisdiction does not recognize imperfect self-defense) or manslaughter (if it does).

3. A person who forms the belief that *p*, which is unreasonable just because it resulted from an implicit attitude, is nonetheless not culpable for having formed it, for reasons having to do with the nature of implicit attitudes and how they influence the beliefs a person forms.

4. The consequences of not convicting such a defendant—i.e., one whose belief that *p* was unreasonable but who was not culpable for having formed that belief—would be unacceptable, unwarranted, unjustifiable, impermissible, and so forth, because not convicting such a defendant would, for example, "send inappropriate messages about the value of the victim's life and the importance of curbing or eliminating racial biases,"[78] or would "generate[] deeply problematic normative messages which entrench stereotypes and devalue the lives of black citizens."[79]

5. To avoid these consequences such a defendant should be punished. Because his belief that *p* was unreasonable, he should be convicted and punished, although not for murder. He should instead be convicted and punished for manslaughter (pursuant to the doctrine of imperfect self-defense), because his unreasonable belief that *p* was non-culpably formed.[80] A manslaughter conviction is said to

---

that deadly force was necessary to save his life, since presumably he doesn't realize his judgment is mistaken, even if he should so realize?

Holroyd and Picinali do not (so far as I can tell) state unequivocally that a belief that *p* resulting from an implicit racial attitude is for that reason alone unreasonable. Still, when discussing what they refer to as a "palliative solution," according to which a person who forms the belief that *p* as a result of an implicit racial attitude should be convicted of manslaughter (not murder) pursuant to, or akin to, the partial defense of imperfect self-defense, they write: "this partial defense could provide a model for the treatment of self-defence claims in cases of *non-culpably unreasonable* belief." Holroyd & Picinali, *supra* note 8, at 189 (emphasis added). The idea of a "non-culpably unreasonable belief" suggests that a belief that *p* resulting from an implicit attitude is "unreasonable," presumably just because it results from an implicit racial attitude, but is at the same time "non-culpable."

78. Richardson & Goff, *supra* note 8, at 320.

79. Holroyd & Picinali, *supra* note 8, at 185.

80. Richardson and Goff state:

> A partial excuse provides a way to *balance* the concerns raised by punishing the *non-culpable* actor with a murder conviction on the one hand and condoning the use of deadly force precipitated by racial stereotypes though complete exoneration on the other.

Richardson & Goff, *supra* note 8, at 325 (emphasis added).

Holroyd and Picinali state:

> We set out with a hypothetical—but all too likely—scenario where implicit racial biases are implicated in perceptual judgements, on the basis of which beliefs about the imminence and gravity of a threat are formed and acted upon. Is the criminal law's person someone who would form such bias-based beliefs? We have argued that there are

> represent a "balance" or "compromise" between a murder conviction (because his belief that *p* was unreasonable) and no liability (because his belief that *p* was non-culpably formed).

Assuming this reconstruction of the argument offered in the two articles is correct (and it may not be), I offer two points in response.

First, the argument (as interpreted) assumes that what makes a belief "unreasonable" is one thing and what makes a belief "culpable" is something else. If so, then it makes sense to say, as I understand the articles to say, that a belief can at the same time be both "unreasonable" and "non-culpable." Now, one can of course assign meanings to the words "unreasonable" and "non-culpable" such that both properties can intelligibly be ascribed to one and the same belief. That having been said, it would nonetheless have been helpful to know more about what makes a belief "unreasonable" and what makes it "non-culpable," such that it makes sense to say that any one belief can be both at the same time.

The insufficient-concern theory described in the text draws no such distinction: a belief is unreasonable just in case it has been culpably formed, and conversely, a belief is reasonable (or at least not unreasonable) just in case it has been non-culpably formed. Like the reconstructed argument above, the insufficient-concern theory can recognize the doctrine of imperfect self-defense. It can permit a jury to return a verdict of manslaughter if a jury believes the defendant (honestly) believed that *p*, but further believes the defendant's belief that *p* was unreasonable (and thus culpable). What it doesn't recognize, but what the reconstructed argument wants to recognize, is a case in which a person's belief that *p* is at the same time both unreasonable and non-culpable.

Second, the reconstructed argument assumes the state is permitted to punish for manslaughter a person who is admittedly non-culpable for having formed the belief that *p* (where that belief is the result of the causal effect of an implicit racial attitude) if and when doing so is thought necessary to produce some anticipated good consequence or to avoid producing some anticipated bad consequence. Indeed, so far as one can tell, the unreconstructed argument appears to call an admittedly non-culpably formed belief "unreasonable" just in case doing so would result in criminal liability and punishment for the person who forms that belief, provided such punishment would (in someone's judgment) produce more good consequences than it would bad consequences.

That logic seems to me materially indistinguishable from the logic countenancing state punishment for someone who is non-culpable because he has committed no crime, or disproportionately punishing him for some crime he has committed, so long as the anticipated good consequences thereby gained,

---

> costs on either way of settling this question. If the criminal law's reasonable person is not susceptible to implicit biases, a distinctive question of fairness arises. The fact is that such biases may influence all of us despite our best efforts. Under these circumstances it is *unfair* to punish individuals for acting in accordance with bias-based beliefs. Yet if the criminal law's reasonable person is susceptible to such implicit biases, the system embeds racist stereotypes. This communicates problematic normative messages about the legitimacy of deploying racial stereotypes, and devalues black lives.

Holroyd & Picinali, *supra* note 8, at 190 (emphasis added).

or the anticipated bad consequences thereby avoided, are believed to be worth the moral cost involved in punishing the innocent or disproportionately punishing the guilty. The logic is the logic of punishing (and thereby using) a person who does not deserve to be punished, or punishing a person beyond the punishment he deserves, as a means to an end. The language of "balance," "compromise," "palliative" and so forth ought not, I should think, to obscure this logic.

This logic isn't indefensible. For example, if we suppose a person has a (claim-) right not to be held criminally liable and punished unless he has culpably chosen to commit a crime, that right, one might argue, ought nonetheless sometimes (all things considered, and rarely) to be overridden, provided the ends the state intends thereby to achieve are believed to be sufficiently urgent, compelling, weighty, and so forth. Again, such an argument is not to my mind always indefensible, but those who advance arguments such as this should acknowledge that the ends they hope the state will achieve when the non-culpable are punished do not come cost-free: they come at the cost of punishing the admittedly non-culpable. I assume this cost is not one to be borne lightly.

II. Reforming Existing Doctrine

Most articles dealing with self-defense and race don't ask how to apply existing law to a case in which a defendant has formed the belief that *p* as a result of an implicit attitude. Instead, they ask how existing self-defense law should be changed because the consequences resulting when a white person is acquitted based on self-defense for having mistakenly killed an innocent black person are said to be worse compared to the consequences resulting when a white person is acquitted based on self-defense for having mistakenly killed an innocent white person. The thesis is that existing self-defense law does not do enough to reduce or prevent these additional consequences. The question is therefore how existing statutory formulations of self-defense should be changed or (if possible) re-interpreted to prevent these consequences.

Assuming an existing statutory formulation according to which the law does not ascribe criminal liability to a person who causes the death of another if he reasonably believed the use of deadly force was necessary to protect against deadly force, the proposed reforms I have in mind retain the reasonableness requirement but recommend other changes to the language of existing law. These recommended changes differ from one author to another, but so far as one can tell, the argument behind all of them is the same.

According to this argument, the law of self-defense should be judged good or bad based on how it distributes the risk of error: the risk of false-positive error (i.e., the type of error resulting when a person reasonably believed the use of deadly force was necessary when in fact it was not) and risk of false-negative error (i.e., the type of error resulting when a person did not believe the use of deadly force was necessary when in fact it was). The cost of a false-positive error is (as a first approximation) death or serious bodily injury, with that cost falling on the person against whom deadly force was mistakenly used. The cost of a false-negative error is (as a first approximation) death or serious bodily

injury, with that cost falling on the person who mistakenly failed to use deadly force. Proposals to reform the law of self-defense—understood as a way of allocating the risk of these errors—then proceeds as follows.

Start with a generic statement of the law of self-defense as it currently exists: a person is permitted to use deadly force in self-defense only if the use of deadly force was necessary to protect against the use of deadly force, but the unnecessary use of deadly force is nonetheless excused if the person reasonably believed its use was necessary to protect against the use of deadly force.[81]

This rule is a liability-defeating rule since it identifies the conditions under which liability for murder will be completely defeated, either because the use of deadly force was permitted, or because the use of such force was prohibited but excused. This rule allocates the risk of false-positive error and false-negative error in whatever way it happens to do so. One might agree or disagree with the way the rule allocates the stated risks, but the current rule can be seen as setting a baseline against which alternative rules, which allocate the risk differently, can be compared.

Reform proponents begin from the premise that the baseline allocation of risk under existing law should be changed because it counts among the relevant costs only the death of the victim at the hands of a person who mistakenly uses deadly force (in the case of false-positive error), and the death of a person who mistakenly fails to use deadly force at the hands of an aggressor (in the case of false-negative error). Existing law fails to account (but should account) for other consequences associated with false positives when a white person uses deadly force against a black person reasonably but mistakenly believing the use of such force was necessary. These consequences are usually described as perpetuating racism, reinforcing stereotypes, limiting the freedom of movement among black people, and so on. Once these additional costs are taken into account for purposes of deciding how to allocate risk, reformers believe the law should be adjusted to reduce the now-recognized greater costs associated with the risk of false-positive error.

The language of the liability-defeating rule governing the use of deadly force in self-defense can be changed in at least three different ways to attempt to accomplish this aimed-for reduction in the incidence of false-positive error:

> 1. Increase the confidence or credence associated with the cognitive attitude applicable to the necessity element. For example, assuming existing law uses a cognitive attitude described simply as "belief," a different word or locution should (reformers claim) be used to state a higher degree or measure of confidence or credence, such as "believed without doubt" or "believed with high confidence." The resulting liability-defeating rule would thus be: a person is permitted to use deadly force in self-defense only if the use of deadly force was necessary to protect against the use of deadly force, but the

81. This formulation of existing law presupposes the wrongful-but-excusable interpretation discussed in the text *supra* Part I.C.2. Under the no-wrong interpretation discussed on those pages, the unnecessary use of deadly force would be permitted—and not simply excused—if the person reasonably believed its use as necessary to protect against the use of deadly force.

> unnecessary use of deadly force is excused only if the person reasonably believed without doubt (or with high confidence) that its use was necessary to protect against deadly force.[82]
>
> 2. Replace the necessity requirement with an imminence requirement. The resulting liability-defeating rule would thus be: a person is permitted to use deadly force in self-defense only when threatened with the imminent use of deadly force against him, but the use of deadly force against the non-imminent use of deadly force is excused only if the person reasonably believed that the use of deadly force against him was imminent.
>
> 3. Combine proposals (1) and (2). The resulting liability-defeating rule would thus be: a person is permitted to use deadly force in self-defense only when threatened with the imminent use of deadly force against him, but the use of deadly force against the non-imminent use of deadly force is excused only if the person reasonably believed without doubt (or with high confidence) that the use of deadly force against him was imminent.

I express no opinion on the wisdom of any of these proposed reforms or similar proposals.[83] I note only that adopting any of them would continue to

---

82. The general idea that some beliefs should be acted upon only if held with confidence is familiar to criminal lawyers. Taking the obvious example, jurors in criminal trial are told they are to vote to convict only if they believe the state has proven each element of the offense charged beyond reasonable doubt. Indeed, one philosopher who defends the "moral encroachment" thesis, discussed *supra* note 29, illustrates the thesis with the criminal law's requirement of proof beyond reasonable doubt. *See, e.g.*, Basu, *supra* note 29, at 205.

83. Two other reform proposals contained in the literature can be found in Mark Kelman, *Reasonable Evidence of Reasonableness*, 17 CRITICAL INQUIRY 798 (1991), and Renée Jorgenson Bolinger, *Reasonable Mistakes and Regulative Norms: Racial Bias in Defensive Harm*, 25 J. POL. PHIL. 196 (2017).

As I understand it, Kelman's argument does not expressly propose making any change to existing doctrine. Existing doctrine asks only about the reasonableness of a defendant's belief that *p*. Nonetheless, Kelman states: "[A]lthough the *stated* norm in self-defense cases makes reference only to the reasonableness of the defendant's factual perceptions, *we in fact also expect the jury to judge the reasonableness of his decision to use deadly force*…." (emphasis added). Kelman, *supra*, at 800. The word "expect" is ambiguous. Does one "expect the jury to judge the reasonableness of [the defendant's] decision to use deadly force" because that's just what (Kelman believes) the average juror does; or, does one "expect the jury to judge the reasonableness of [the defendant's] decision to use deadly force" because (Kelman believes) that's what jurors should do? In any event, this assumption about what "we in fact also expect" of jurors allows Kelman to argue that jurors asked to return a verdict on a claim of self-defense should take account of the additional adverse consequences arising when a white person kills a black person. *See id.* at 815-16.

Kelman's proposal, so far as I can tell, would make a claim of self-defense turn on a jury's assessment of the costs and benefits of the defendant's decision to use deadly force, such that even if a defendant reasonably believed that *p*, a jury could still deny the defense if it judged that the costs of the defendant's action exceeded the benefits. Moreover, the defendant could be denied the defense on that ground without any inquiry into the defendant's mental state with respect to the balance of costs and benefits. In other words, Kelman appears to want to graft a strict-liability

leave open the question addressed here: what makes a person's belief (with whatever required level of confidence or credence) reasonable for purposes of the law of self-defense (whether the object of that belief is the necessity of using deadly force in self-defense, the imminence of the use of deadly force against the person, or the two combined)? Changing the cognitive attitude the liability-defeating rule requires (from belief to belief without doubt and so forth) or the object of that attitude (from necessity to imminence or both) would leave intact the need to identify the conditions under which that cognitive attitude is said to be "reasonable" or "unreasonable."

APPENDIX B — NECESSITY

The law of self-defense (as I have generically stated it) permits the use of deadly force only when the use of such force is "necessary" to defend against deadly force. But the word "necessary," so far as I know, is not defined in statute or standard jury instructions.[84] Jurors are simply asked to decide if (among other things) the defendant's use of force was "necessary." A few commentators have nonetheless tried to say more about how "necessity" might or should be defined, or at least to identify questions the law's use of the word "necessity" raises.

For example, perhaps "necessity" should be defined using probabilities. Suppose the law provided: the use of deadly force is "necessary" if, at the time

---

element onto the existing law of self-defense, or to believe that such an element already "exists," at least in the sense that jurors already behave as if it exists in law, which it doesn't.

Bolinger offers another way to reform self-defense law. Among other possibilities, she suggests changing the liability-defeating rule such that it apparently states something like the following: a person is not permitted to use deadly force unless he reasonably believed the use of such force was necessary to protect against the use of deadly force, but the use of deadly force is not necessary unless φ, where φ describes with some degree of specificity some action or set of actions the victim performed. *See* Bolinger, *supra*, at 201 ("Under a public evidential norm, a society can treat some behaviors as marked 'signals' of aggression, licensing agents to assume (absent countervailing evidence) that the performer is an aggressor."); *see also* Renée Jorgenson Bolinger, *The Moral Grounds of Reasonably Mistaken Self-Defense*, 103 PHIL. & PHENOMENOLOGICAL RES. 140, 143 (2021) (offering "an account that could work as a moral foundation" for the thesis advanced in Bolinger (2017)).

This proposal changes the defense inasmuch as it makes the liability-defeating rule more rule-like. That change will, inasmuch as it moves the formulation of self-defense toward the rule end of the rule-standard continuum, bring familiar problems associated with under-inclusion and over-inclusion. The proposal is analogous to other proposals according to which a more-or-less vague element of an offense or defense should be statutorily defined with greater precision. *See, e.g.*, Michael S. Moore & Heidi M. Hurd, *Punishing the Awkward, the Stupid, the Weak, and the Selfish: The Culpability of Negligence*, 5 CRIM. L. & PHIL. 147, 186-91 (2011) (discussing proposals to substitute "mini-maxims" the violation of which result in harm in place of a general prohibition on the inadvertent (negligent) creation of risk with harm resulting). This proposal, like the second proposal identified in the text, changes what the law of self-defense takes as the object of the defendant's belief. It therefore leaves open what makes a belief with respect to that object reasonable or not, just as the second proposal discussed in the text does.

84. Nor does it define "imminence" in those jurisdictions in which imminence is required before a person is permitted to use deadly force in self-defense.

the defendant used deadly force, the probability that the victim would φ (where φ describes some action) was *x*%, and the probability that φ-ing would cause death or serious bodily injury to the defendant was *y*%, where the law further specifies some value for *x* and *y*.

Using probabilities to define when the use of deadly force is "necessary" would make the law of self-defense more precise (or at least appear more precise) but doing so would bring another problem more clearly into view. Judgments about the probability of an event happening are presumably based on beliefs about the facts at one moment in time, together with beliefs about how those facts relate probabilistically to some future event. The probability that an identified event will happen is thus relative to some epistemic point of view from which those predicate beliefs are formed. The problem is how to identify or describe that point of view.

If determinism is true, and if someone knows all the facts and all the laws of nature causally connecting one event to another, then the probability that one event or set of events will cause another will be (and the someone will know it to be) either 0 or 1. In other words, for someone who's omniscient the language of probability would neither be used nor useful.[85] The language of probability is a language for which only epistemically limited beings (like us) have any use. From an omniscient point of view the language of probability wouldn't make much sense.

Beliefs about probabilities formed from an omniscient point of view are one extreme. Beliefs about probabilities formed from some particular and identifiable person's point of view are the opposite extreme. For purposes of self-defense, that particular point of view might be the defendant's point of view. But from the defendant's point of view, the probability that an event will happen is nothing more than whatever the defendant believed was the probability it would happen, at the moment he believed it.

Between these two extremes—the beliefs an omniscient being would have formed and the beliefs the defendant formed—are the beliefs some imagined person would have formed—someone who is neither the defendant nor omniscient—like the "reasonable person" or a person of "common sense" and so forth. If the beliefs a reasonable person (or whatnot) would have formed is used to fix or identify the relevant probabilities—to fix or identify what the relevant probabilities "in fact" were or what they "objectively" were—then any distinction between what those probabilities "in fact" are or what they "objectively" are, and what a reasonable person would have believed them to be, disappears.

As noted in the text, I attempt to bracket all these questions as best I can. As I describe the facts in John's case, I assume he formed the belief that *p*

85. This way of stating the question oversimplifies it. For a more complete and nuanced analysis, see Larry Alexander, *Recipe for a Theory of Self-Defense: The Ingredients, and Some Cooking Suggestions*, *in* THE ETHICS OF SELF-DEFENSE 20 (Christian Coons & Michael Weber eds., 2016); Larry Alexander, *The Need to Attend to Probabilities—For Purposes of Self-Defense and Other Preemptive Actions*, 55 SAN DIEGO L. REV. 223 (2018); Vera Bergelson, *Self-Defense and Risks*, *in* THE ETHICS OF SELF-DEFENSE, *supra*, at 131.

because he believed the victim had in hand a gun, only to discover after the fact (i.e., after killing the victim because he believed that *p*) that the gun was an iPhone. I could also have simply stipulated that the victim never intended to cause John any harm. Including facts such as those make the most unambiguous case I can imagine for saying that John's use of deadly force was "in fact" unnecessary. They are intended to increase the probability that the reader will agree with the conclusion that the probability the victim would have φ'ed was at or near zero, in which case the probability of death or serious bodily injury to the defendant would likewise have been at or near zero.

APPENDIX C — SYNCHRONIC (DIRECT) V. DIACHRONIC (INDIRECT) BELIEF FORMATION

The text asks the following narrowly defined question: has a person who forms the belief that *p*, but who would not have formed that belief but for the causal influence of an implicit attitude, formed that belief unreasonably *at the moment he uses deadly force (which is presumably the same moment at which he forms the belief that p)?*

The italicized language should not be overlooked. It limits the scope of the question. The question asks about the reasonableness (culpability) associated with a belief a person forms at the moment he uses deadly force. It doesn't ask about anything he did or failed to do at any prior moment in time. It asks about the culpability associated with the formation of the belief that *p*, not about any culpability associated with any putatively wrongful act or omission prior to the moment the belief that *p* was formed—even if the actor would not have formed the belief that *p* had he not acted or failed to act in some way identified as wrongful. In other words, the question asks about responsibility (culpability) for synchronic (or direct) belief formation, not about responsibility (culpability) for diachronic (or indirect) belief formation.

Consider the following example. Stipulate (for the sake of argument) that John's formation of the belief that *p*, which he would not have formed but for the causal influence of an implicit attitude, was reasonable (non-culpable). Now suppose you believe John was morally obligated (and should be legally obligated), at some point before he formed the belief that *p*, to ask his victim what he had in his hand, and suppose that if John had performed this simple speech act the victim would have told him the thing in his hand was a cell phone. Next suppose that if the victim had so responded John would have believed him and not formed the belief that *p*.

Now stipulate that John's wrongful failure to inquire was culpable (though more would need to be said to establish his culpability for having failed to inquire). As I understand it, existing law imposes no such duty to inquire, but assume it did. Assume the law of self-defense is changed. The amended law of self-defense provides that because John's failure to inquire was wrongful and culpable, he is no longer permitted to plead self-defense—even though *ex hypothesi* John's formation of the belief that *p* at the moment he formed it was reasonable (non-culpable).

The proposed duty-to-inquire rule would function like another rule. Unlike the proposed rule, this other rule is a familiar feature of the law of self-

defense. According to this rule, when a person's conduct and mental states satisfy a particular jurisdiction's definition of an "aggressor," the person loses his permission to use deadly force in self-defense, unless he effectively "renounces" his initial aggression—even if at the moment he used deadly force he otherwise satisfies all the elements of the defense.[86] This "aggressor rule" is a forfeiture rule: unrenounced aggressors forfeit the legal permission to use deadly force in self-defense. The proposed duty-to-inquire rule would likewise function as a forfeiture rule.

One might propose such a forfeiture rule involving implicit racial attitudes. For example, one might propose a duty to act in such a way as to eliminate, or attempt to eliminate, the attitude altogether from one's mind, such that it could no longer causally influence the beliefs a person forms, including the belief that *p*. Or one might propose a duty to act or not to act so as not to acquire an implicit racial attitude in the first place, such that once again the beliefs one forms, including the belief that *p*, could not causally result from an implicit attitude. If a person no longer possesses an implicit attitude—having never acquired it, or having dispossessed himself of it after having acquired it—then it wouldn't exist to influence the beliefs he forms.[87]

Alternatively, one might propose a duty obligating a person to perform some specified mental act at some point prior to the moment he forms the belief that *p*. For example, one might propose that a person is obligated, upon forming the belief that another person is black, to bring to mind positive images of black men, or to perform a mental act of "race-switching," and so on. The hope would be that discharging some such mental-act obligation would either prevent an implicit attitude from becoming "activated" in the first place, or if such an attitude does become "activated," would somehow prevent the "activated" implicit attitude from causally influencing the beliefs one forms despite its having been activated.

Such proposed duties—all intended to prevent implicit attitudes from causally influencing the beliefs one forms—are mental duties, but that doesn't, so far as I can tell, make them any less diachronic: something a person is obligated to do (albeit with his mind) at some point in time prior to the point in time at which he forms the belief that *p*. The interval between the time John is obligated to discharge these mental duties and his formation of the belief that *p*, if he fails to perform them, may be vanishingly small, but some interval,

---

86. *See, e.g.*, MODEL PENAL CODE § 3.04(2)(b)(i) ("The use of deadly force is not justifiable … if … the actor, with the purpose of causing death or serious bodily injury provoked the use of force against himself in the same encounter…."); DRESSLER, *supra* note 9, § 18.02[B][1], at 226-28.

87. Two common citations in the psychological literature about the effectiveness and durability of various interventions intended to achieve this result appear to be Calvin K. Lai et al., *Reducing Implicit Racial Preferences: I. A Comparative Investigation of 17 Interventions*, 143 J. EXPERIMENTAL PSYCH.: GEN. 1746 (2014), and Calvin K. Lai et al., *Reducing Implicit Racial Preferences: II. Intervention Effectiveness Across Time*, 145 J. EXPERIMENTAL PSYCH.: GEN. 1001 (2016). *But see* Heidi A. Vuletich & B. Keith Payne, *Stability and Change in Implicit Bias*, 30 PSYCH. SCI. 854 (2019) (reanalyzing Lai (2016) and reaching "nearly the opposite of the original interpretation").

however small, must nonetheless exist between the failure to perform the mental act and the formation of the belief that *p*.

Whatever other difficulties such forfeiture rules might encounter, commentators have long objected to them because they typically result in disproportionate punishment. For example, stipulate that John's formation of the belief that *p* was reasonable at the time he formed it. Stipulate further, however, that John had a duty to bring to mind positive images at the moment he formed the belief that the person he saw was black. Finally, stipulate that if John had discharged this duty, he would not have formed the belief that *p*, but that John fails to discharge this duty. Because he fails to discharge this duty, he forms the belief that *p*. His formation of the belief that *p* at the time he formed it nonetheless remains (I'm supposing) reasonable (non-culpable). The only wrongful and (I'll suppose) culpable thing John has done—actually, failed to do—is to bring positive images of black men to mind.[88]

Although John's formation of the belief that *p* was (I'm supposing) otherwise reasonable, John would nonetheless be convicted of murder (or manslaughter) under the proposed forfeiture rule. He would by operation of law forfeit his permission to use deadly force in self-defense when he would otherwise have been entitled to prevail on that defense. John would be convicted of murder (or manslaughter) because he failed (culpably, I'm again supposing) to discharge his stipulated duty to bring positive images to mind, and as a result, caused the death of an innocent. Of course, that might be a rule some commentators would endorse, on the theory that disproportionate punishment is permissible, all things considered, in light of the anticipated good consequences they believe its application will produce.[89]

APPENDIX D – SHOOTER STUDIES

I assume for purposes of the analysis offered in the text that an implicit racial attitude can cause a person to form the belief that *p*, when the person would not otherwise have formed that belief. I mentioned what are known in the psychological literature as "shooter studies." I suspect many readers who've read these studies, or who've seen then cited, believe their results provide strong evidence for the truth of the assumption just mentioned. I expressed some doubt about that assumption in the text. One reason for that doubt is as follows.

On the one hand, in the abstract (which I quote in full) of the only meta-analysis of shooter studies of which I'm aware (as of February 2021),[90]

88. I should also note that the sundry forfeiture rules associated with criminal-law affirmative defenses are not themselves criminal offenses, although conduct instantiating the rule could be. For example, a person who engages in conduct in virtue of which he would be characterized as an aggressor for purposes of self-defense (and who would thus forfeit the defense unless he "renounced") might not, in virtue of that conduct alone, be guilty of any criminal offense. It depends on what the rule says.

89. Compare the argument discussed *supra* App.A.I.

90. For another "overview" of the shooter-bias literature, which is how the authors describe their article in its abstract, see William T.L. Cox & Patricia G. Devine, *Experimental Research on*

published in 2015 in the *Journal of Experimental and Social Psychology*, a peer-reviewed journal, the authors (Yara Mekawi and Konrad Bresin) write, with my emphasis added:

> The longstanding issue of extrajudicial police shootings of racial and ethnic minority members has received unprecedented interest from the general public in the past year. To better understand this issue, researchers have examined racial shooter biases in the laboratory for more than a decade; however, shooter biases have been operationalized in multiple ways in previous studies with mixed results within and across measures. We meta-analyzed 42 studies, investigating five operationalizations of shooter biases (reaction time with/without a gun, false alarms, shooting sensitivity, and shooting threshold) and relevant moderators (e.g., racial prejudice, state level gun laws). *Our results indicated that relative to White targets, participants were quicker to shoot armed Black targets (d_{av} = —.13, 95% CI [—.19, —.06]), slower to not shoot unarmed Black targets (d_{av} = .11, 95% CI [.05, .18]), and more likely to have a liberal shooting threshold for Black targets (d_{av} = —.19, 95% CI [—.37, —.01]).* In addition, we found that in states with permissive (vs. restrictive) gun laws, the false alarm rate for shooting Black targets was higher and the shooting threshold for shooting Black targets was lower than for White targets. These results help provide critical insight into the psychology of race-based shooter decisions, which may have practical implications for intervention (e.g., training police

---

Shooter Bias: Ready (or Relevant) for Application in the Courtroom?, 5 J. APPLIED RSCH. MEMORY & COGNITION 236, 238 (2016). This article's concluding paragraph states:

> It is deeply unsettling that the color of someone's skin could lead to them dying at the hands of a law enforcement officer. When such tragedies happen, we as a society want justice, and those devoted to fighting for social justice point to these tragedies as evidence of deep-seated racial issues in the United States. For these reasons, it often seems readily apparent that shooter bias research should, of course, be applied in the courtroom—the place where justice is served. *Shooter bias research, however, does not yet display a robustness and consistency necessary to make any confident claims related to police officer's vulnerability to showing patterns of shooter bias*.

Id. at 238 (emphasis added).

Yet another review of the research, published in 2014, states in its abstract:

> Experimental work with undergraduate participants reveals a clear pattern of bias (a tendency to shoot Black targets but not Whites), which is associated with stereotypes linking Blacks with the concept of danger. Subsequent work with police officers presents a more complex pattern. Although police are affected by target race in some respects, they generally do *not* show a biased pattern of shooting.

Joshua Correll et al., *The Police Officer's Dilemma: A Decade of Research on Racial Bias in the Decision to Shoot*, 8 SOC. & PERSONALITY PSYCH. COMPASS 201, 201 (2014) (emphasis in original).

> officers) and prevention of the loss of life of racial and ethnic minorities.[91]

On the other hand, in another study, published in 2018 in *Collabra: Psychology*, also a peer-reviewed publication, the authors (Caren M. Rotello, Laura J. Kelly and Evan Heit) write in a section entitled "General Discussion," again with my emphasis added:

> Overall, then, we are reluctant to draw larger conclusions based on the body of all published studies of the weapon identification and first-person shooter tasks. However, here we consider the implications of our own experiments. In four experiments that extended two classic paradigms (Payne, 2001; Correll et al., 2002), we observed little to no difference for race-based effects on gun/non-gun decision accuracy. This was true both when decision accuracy was very high (Experiments 1a) and when it was lower (Experiments 1b, 2a, and 2b). We also observed no evidence that supports a negative bias toward Black agents or primes; the "gun" response rate was never significantly higher for stimuli involving Black individuals than those involving White individuals, and in Experiment 2a the "gun" response rate to Black agents was significantly lower. *Thus, these data are consistent with the conclusions of Mekawi and Besin's (2015) meta-analysis on the effects of race on decision accuracy, but the data contradict their conclusion that participants are generally more likely to "shoot" at Black agents.* An optimistic view of these different empirical outcomes is that societal changes in the 17 years since publication of Payne's initial study have weakened the association between guns and Black agents. A more realistic view might be simply that response rate biases vary readily (e.g., Rotello & Macmillan, 2008). Indeed, Mekawi and Bresin's (2015) meta-analysis included 13 (of 29) individual experiments in which no difference in response bias was observed across agent race (like our Exps. 1a, 1b, and 2b) and another 3 experiments which (like our Exp. 2a) reported a significant bias to "shoot" more at White agents.[92]

What, then, do the shooter studies tell us about the human mind, and how it works? Readers should, before answering, read for themselves the two articles I've quoted, and not just the passages I decided to reproduce, as well as the many other articles to be found in the literature falling under the header "shooter studies." As I mentioned in the text, I assumed that the shooter studies give evidence (reasons) to believe that implicit racial attitudes cause, or can cause, a person (or some average person, perhaps of this or that demographic) to form the belief that *p* when otherwise the person (or some average person, perhaps of this or that demographic) would not have formed that belief. I made

---

91. Yara Mekawi & Konrad Bresin, *Is the Evidence from Racial Bias Shooting Task Studies a Smoking Gun? Results from a Meta-Analysis*, 61 J. EXPERIMENTAL SOC. PSYCH. 120, 120 (2015).

92. Caren M. Rotello et al., *The Shape of ROC Curves in Shooter Tasks: Implications for Best Practices in Analysis*, 4 COLLABRA: PSYCH. 1, 12 (2018).

no effort to assess the truth of that assumption. Readers will need to assess its truth for themselves.