Crim Law and Philos (2018) 12:107–124
https://doi.org/10.1007/s11572-017-9414-0

ORIGINAL PAPER

Ignorance Lost: A Reply to Yaffe on the Culpability of Willful Ignorance

Alexander Sarch[1]

Published online: 23 March 2017

Abstract In a recent paper in this journal, Gideon Yaffe provides an expected utility model of culpability in order to explain why willfully ignorant misconduct sometimes is just as culpable as knowing misconduct. Although promising, I argue here that challenges remain for Yaffe's view. First, I argue that Yaffe's proof of the equal culpability of willful ignorance and knowledge is not watertight in certain realistic cases. Next, I argue that Yaffe's view of culpability is motive-sensitive in a way that sits uncomfortably with criminal law doctrine, and I show that his view has difficulty with unjustified actions that are nonetheless privileged. Perhaps these problems can be solved by modifying Yaffe's account using the notion of legally recognized reasons. However, I argue that difficulties remain when it comes to implementing this solution into Yaffe's mathematical model. Finally, I raise concerns about Yaffe's account of willful ignorance in particular. While his view initially seems to have a major advantage over the additive picture of willful ignorance I've defended, this advantage does not stand up under scrutiny. In fact, Yaffe likely relies (albeit covertly) on an additive metaphysical picture of willful ignorance as well.

Keywords Willful ignorance · Culpability · Knowledge · Recklessness · *Mens rea* · Motives

1 Introduction

Suppose the commissioner of the NFL suspects that football players may be suffering brain damage from the concussions they receive during gameplay. His advisers tell him they could carry out a major study to determine the extent to which this is the case. But he tells them to not to bother—the money should go to finishing up a new stadium in Los Angeles

✉ Alexander Sarch
a.sarch@surrey.ac.uk

[1] School of Law, University of Surrey, Guildford, Surrey GU2 7XH, UK

instead. Moreover, he orders them not to talk to him about this again, and tells them they should actively avoid discussing the matter with any scientist or doctor who might have any expertise on the issue. Suppose it turns out that the concussions really are causing brain damage. In this case, the commissioner has been willfully ignorant. He has deliberately preserved his ignorance of the fact that concussions sustained playing football cause brain damage.[1]

In the criminal law, willful ignorance serves as a substitute for actual knowledge on the theory that doing a crime with the former mental state is just as culpable as doing it with the latter.[2] When and why this thesis might be true are questions that have received far too little attention.

In a recent paper in this journal, Gideon Yaffe provides an intriguing answer.[3] He offers a novel account of the culpability of criminal actions, premised on an expected utility model of the motivations of criminal actors. This model can be used to draw inferences about the actor's level of regard for others in particular cases—including cases of willful ignorance. Yaffe goes on to demonstrate that, in at least some cases, we can be quite sure that willfully ignorant actors are just as culpable as knowing wrongdoers. Yaffe's model is compelling and insightful, and sheds light on many aspects of the way in which *mens rea* concepts function in the criminal law.

In this paper, however, I raise a number of concerns about Yaffe's account. One is a technical problem for Yaffe's demonstration that some willfully ignorant crimes are just as culpable as the analogous knowing misconduct. In particular, I point out that Yaffe's proof is not watertight in certain realistic cases of willful ignorance, where investigation is costly to the agent. The second problem is that Yaffe's view is motive-sensitive in a way that sits uncomfortably with criminal law doctrine. While this may not be a decisive normative objection, it raises questions about the ability of Yaffe's model to account elegantly for posited law. Third, I suggest that Yaffe's account as stated has difficulty with unjustified actions that are nonetheless privileged. I proceed to discuss a way to solve both of these problems that invokes the idea of legally recognized reasons, but I show that difficulties remain in implementing this solution using Yaffe's mathematical model.

Finally, I close with some observations about Yaffe's account as applied to willful ignorance. His approach at first seems to have a significant advantage over other recent accounts—in particular, additive views of the sort I have defended.[4] However, this *prima facie* advantage does not stand up to scrutiny. In investigating the metaphysical picture on which Yaffe's view relies, it emerges that he, too, likely must appeal to an additive metaphysical picture of willful ignorance.

While the posture of this paper is critical, I am on the whole quite sympathetic to Yaffe's view. I think he provides a very illuminating account of the culpability of willful

[1] In its most recent case on willful ignorance, the Supreme Court observed that courts generally take willful ignorance to involve "two basic requirements: (1) the defendant must subjectively believe that there is a high probability that a fact exists and (2) the defendant must take deliberate actions to avoid learning of that fact." Global-Tech Appliances, Inc. v. SEB S.A., 131 S. Ct. 2060, 2070 (2011).

[2] *See, e.g.,* United States v. Jewell, 532 F.2d 697, 700 (9th Cir. 1976); United States v. Heredia, 483 F.3d 913, 917 (9th Cir. 2007) (en banc).

[3] Gideon Yaffe, *The Point of Mens Rea: The Case of Willful Ignorance*, Crim. L. & Philosophy (forthcoming). All page citations to Yaffe's article are to the First Online version available on this journal's website (https://link.springer.com/article/10.1007%2Fs11572-016-9408-3 (accessed March 10 2017)). The pagination of the print article may differ.

[4] *See* Alex Sarch, *Willful Ignorance, Culpability and the Criminal Law*, 88 St. John's L. Rev. 1023 (2014); Alex Sarch, *Beyond Willful Ignorance*, 88 U. Colo. L. Rev. 97 (2016).

ignorance. Thus, my aim here is to push the conversation forward by digging into the details of Yaffe's view and pointing out some places where questions linger and additional work remains.

2 Yaffe on the Culpability of Willful Ignorance

2.1 Yaffe's Formal Model of Culpability

Yaffe's paper develops an account of culpability in general, which he then applies to the case of willful ignorance. I sketch the account in this section before critiquing it in subsequent sections. In a nutshell, Yaffe thinks culpability depends on the strength of one's preference for promoting one's own interests as compared to the interests of others. As he puts it:

> [The agent's] actions manifest the evaluative weight that he gives to his own interests in comparison to the interests of other people. When a tendency to put himself first, to a greater degree than is acceptable, is manifested in his conduct, he is criminally culpable for that conduct.[5]

Yaffe calls this the actor's *social preference*, which is a function of the relative weights attached to one's own interests and the interests of others.

Culpability, for Yaffe, doesn't track the social preference one *believes* one possesses (*e.g.*, as when a selfish person unconvincingly asserts that he is really very altruistic). Instead, what matters for culpability is the social preference that is manifested in—*i.e.*, can be inferred from—the actions one actually performs and the mental states with which one does them. In drawing inferences about one's social preferences, Yaffe thinks, we are constrained by certain kinds of presumptions we are obliged to make about our fellow citizens. One is the presumption that the agent is rational.[6] Moreover, we are bound by a principle of lenity, which Yaffe formulates thus:

> The principle of lenity demands that we interpret [the actor's] behavior to manifest a social preference as close as it can be reasonably taken to be to an acceptable one. It demands, that is, that we depart as little as is reasonably possible from the conception of him as someone who cares as much about others as he ought.[7]

Accordingly, Yaffe contends, "[w]e find criminal culpability when, even under [such] constraints, the agent's conduct manifests an unacceptable social preference."[8]

Yaffe seeks to make this picture of culpability precise by using a formal model,[9] premised on the Value Equation he developed in earlier work.[10] It is meant to capture the weights of the applicable reasons that, by the agent's own lights, count in favor of or against a specific action, A. These reasons are cashed out in terms of the expected utility of

[5] Yaffe, *supra* note 3 at 8 (emphasis omitted).

[6] Where this presumption is known not to hold for a particular actor, all bets are off when it comes to drawing inferences about his culpability.

[7] Yaffe, *supra* note 3 at 8.

[8] *Id.*

[9] It involves simplifying assumptions, to be sure, but they can be refined to make the model more realistic if needed.

[10] Gideon Yaffe, *Intoxication, Recklessness, and Negligence*, 9 Ohio St. J. Crim. L. 545 (2012).

doing A—*i.e.*, the probability-weighted benefits and detriments that would accrue both to the agent, and to other people, if he does A. In a nutshell, the value of A equals the net expected value that (by the agent's lights) A would have for himself (abbreviated "EV_{Self} (A)") plus the net expected value that (by the agent's lights) A would have for others (abbreviated "EV_{Others} (A)"). In addition, the latter term—EV_{Others} (A)—is adjusted by two factors: (1) the degree to which the agent *attends* to the effects of A on others (abbreviated ("a"),[11] and (2) the agent's social preference, *i.e.*, the degree to which he *cares* about the effects of his conduct on others as compared with its effects on himself (abbreviated "s"). Thus, assuming A is a criminal action that would have positive value for the actor but would cause harm (*i.e.*, have net negative value) for others, we can state the Value Equation thus:

$$\textbf{Value Equation}: \quad \mathrm{Value(A)} = \mathrm{EV_{Self}(A)} - \mathrm{a} * \mathrm{s} * \mathrm{EV_{Others}(A)}$$

Since the actor is assumed to perform the action, he must regard the overall value of A as positive. Moreover, for simplicity, in what follows, I will assume that the agent pays full attention to the effects of A on others, so that a = 1. Thus, it can be dropped from the analysis for present purposes. From the Value Equation together with the facts of a given case, we can draw inferences about the actor's social preference by solving for s, which culpability tracks.

To make the Value Equation easier to apply, I introduce the following abbreviations:

- B = The benefit of A to oneself broadly understood (*i.e.*, which includes benefits to the personal projects one desires to complete)
- p1 = The probability that, by the agent's lights, B has of resulting conditional on the agent's doing A
- H = The harm to others that would result from the agent's doing A
- p2 = The probability that, by the agent's lights, H has of resulting conditional on the agent's doing A
- s = The agent's social preference

Finally, on Yaffe's model, s = 0 when the agent doesn't care at all about the effects of his conduct on others, and s = 1 when the agent cares just as much about the effects of his conduct on others as he does about its effects on himself. One is culpable when s < 1 (or below some other communally established threshold). With these abbreviations—and omitting for now the attention factor, a—the Value Equation can be schematically stated as follows:

$$\textbf{Value Equation}: \quad \mathrm{Value\ (A)} = \mathrm{B} * \mathrm{p1} - \mathrm{s} * \mathrm{H} * \mathrm{p2}$$

To show how we can infer culpability from the Value Equation, consider two examples: knowledge and recklessness. As Yaffe shows, it will emerge that knowingly causing harm, all else equal, is more culpable than recklessly causing the same harm.

2.1.1 Knowing Misconduct

Start with a case of knowing misconduct. Suppose that the act in question, A, is the act of burning down a building in return for $1000 despite knowing that a person is inside and will die as a result. (Thus, B = the benefit of receiving $1000, while H = the disvalue of a person dying.) Let's suppose B is certain if the actor does A, such that p1 = 1. (However,

[11] "The *a* parameter is a measure of the agent's attention to the potential impact of his act on other people." Yaffe, *supra* note 3 at 8.

he won't receive this benefit, or any other, if he doesn't do A.) Moreover, since A is supposed to be an act of knowingly imposing harm—*i.e.*, doing an act that the agent is practically certain will cause the harm—we can also suppose that p2 = 1. (If one is substantially less confident than this that H will occur, such that p2 is much less than 1, then the act would be only reckless.[12]) The last assumption is that the actor actually performs A, meaning we can be confident that he attaches greater overall value to performing A than he does to not doing A. Now we can solve for s:

Derivation of the actor's s-value in a case of knowing misconduct:

$$\begin{aligned}
&\text{Value (A)} > \text{Value } (\sim \text{A})\\
&\text{B} * 1 - \text{s} * \text{H} * 1 > \text{B} * 0 - \text{s} * \text{H} * 0\\
&\text{B} - \text{s} * \text{H} > 0\\
&- \text{s} * \text{H} > -B\\
&\text{s} * \text{H} < B\\
&\text{s} < \text{B}/\text{H}
\end{aligned}$$

Thus, when the defendant does A, we know his s-value (which his culpability tracks) is below B/H. This reflects his exchange rate as between units of value for himself and units of value for others. Just as a currency trader would rather keep 100 dollars than 100 yen, the actor here would rather keep $1000 than keep the person alive. This clearly is a deplorable way to be. This actor's exchange rate as between units of value for himself as compared to units of value for others is seriously screwed up. Hence, he is very culpable.[13]

2.1.2 Reckless Misconduct

Now compare the exactly analogous reckless act. Let *A'* be an act that is the same as before—*i.e.*, burning down a building—except that now the agent is only aware of a substantial (and unjustified) *risk* that a person is inside and will die in the fire. For the reckless act A', all the variables are as before, except that now p2 is less than 1. Let's say p2 = p, a number somewhat greater than 0 but somewhat less than 1. Thus, we get:

Derivation of the actor's s-value in a case of reckless misconduct:

$$\begin{aligned}
&\text{Value } (\text{A}') > \text{Value } (\sim \text{A}')\\
&\text{B} * 1 - \text{s} * \text{H} * \text{p} > \text{B} * 0 - \text{s} * \text{H} * 0\\
&\text{B} - \text{s} * \text{H} * \text{p} > 0\\
&- \text{s} * \text{H} * \text{p} > - \text{B}\\
&\text{s} * \text{H} * \text{p} < \text{B}\\
&\text{s} < \text{B}/\text{H} * \text{p}
\end{aligned}$$

[12] Model Penal Code ("MPC") § 2.02(2)(c) (a "person acts recklessly with respect to a material element of an offense when he consciously disregards a substantial and unjustifiable risk that the material element exists or will result from his conduct").

[13] One might worry that in doing a criminal action one might seek benefits that don't accrue to oneself—*e.g.*, when one steals $5000 from an ATM in order to be able to send one's child to a better private school, or to benefit a Save the Rainforest NGO. Are these really benefits for *oneself*? On closer inspection, however, this isn't really a problem. The simplest way to deal with it is to say that these benefits belong under EV_{Self} construed broadly so as to include any project or goal one personally seeks to promote in acting.

Now we're in a position to compare the s-values of reckless and knowing wrongdoers. The threshold we know the reckless wrongdoer's s-value to be below is B/H*p. By contrast, we saw that the threshold that the knowing wrongdoer's s-value is below is something lower: B/H. (After all, B/H is a smaller number than B/H*p. The factor p is a number less than 1. Thus, if $p = 0.5$, then while B/H might be, say, 1/10, B/H*p would be $1/(10*0.5) = 1/5$.) Thus, the threshold we know the knowing wrongdoer's s-value to be below is *lower* than the threshold we know the reckless actor's s-value to be below. Hence, under the principle of lenity, more culpability (a lower s) can be inferred from a knowing wrong than from the analogous reckless wrong. As Yaffe concludes:

> Given a presumption of lenity—we hold agents responsible for being the best they can be consistent with what we know of their behavior and their mental states—knowing agents manifest in their behavior worse social preferences than do reckless agents.[14]

2.2 The Value Equation Applied to Willful Ignorance

These examples demonstrate how Yaffe's view of culpability can be applied to reach conclusions about culpability. My main concern here is willful ignorance, so I conclude my exposition of Yaffe's view by applying it to the two main types of case of willful ignorance.

2.2.1 The Core Case of Willful Ignorance

Here is how Yaffe understands the core case of willful ignorance. Suppose a criminal defendant suspects that the action he is thinking about doing at time t2 would cause harm to others, but he doesn't investigate in readily available ways, at an earlier time t1, whether the action at t2 really will cause the suspected harm. For example, suppose I plan to burn down a building while suspecting that a person is inside. I could investigate the matter, but I decide not to. In the core case, we assume that investigating would be costless. For instance, perhaps I can quickly look inside before lighting the fire. However, for this to be a core case, we need to add one more feature: suppose that if I knew there was a person inside (*i.e.*, if I was practically certain that the suspected risk would materialize), then I would be unable to go through with the arson. I'd chicken out, perhaps, or be overcome by conscience. Thus, the only way I can bring myself to complete the crime (thereby earning myself a handsome fee) is by preventing myself from obtaining full knowledge of the risk in question. Thus, I actually *need* to preserve my ignorance of whether the harm will occur in order to be able to go through with the act. Investigating whether someone is inside would jeopardize the perceived benefits to me of committing the arson. Hence, as Yaffe describes it, for a case of willful ignorance to be core, we must suppose:

> inquiry was costless and perfect (the method yields neither false positives nor false negatives), and [the actor] failed to inquire at t1 because inquiry would have risked the possibility that he would discover X [*i.e.*, the risk he suspects his later act would impose], a discovery that would prompt him to refrain from acting at t2, and would thus preclude the possibility of his receiving the benefit of such action.[15]

So here's how this plays out. *X* designates the fact that my subsequent action will cause harm to others. In a core case, if I investigate at t1, then I will learn either X or not-X—*i.e.*, that H will result from my contemplated act of arson, A, or that it won't. If I learn X, then

[14] Yaffe, *supra* note 3 at 10.

[15] *Id.* at 11.

my conscience will prevent me from doing A, such that I lose out on the benefit to myself of A. Only if I learn not-X will I go ahead and do A, thereby obtaining the benefit. In this way, investigating jeopardizes my receipt of the benefit to myself of A.

To determine my culpability for lighting the building on fire while willfully ignorant of whether a person is inside, we need to apply the Value Equation and draw an inference about my s-value. Here's how to do it. As before, we must assume that the value I attach to doing A in a state of willful ignorance is greater than the value I attach to not doing so. Thus, we get:

$$\text{Value (not}-\text{inquire and then do A)} > \text{Value (inquire and then maybe}-\text{or}-\text{maybe}-\text{not do A)}$$

Applying the same abbreviations from above, we can flesh this out as follows:

$$B * 1-s * H * p > B * (1-p)-s * H * 0$$

I assume it's obvious how we got the results on the left side of the inequality. But how did we get the results on the right? Well, on the right side of the inequality (where I inquire), I actually get the benefit only if I find out that harm H (the death of a person inside the building) will not ensue. The chance of someone being in the building (*i.e.*, of H occurring) is p, and I'll only light the fire if I find out that no one is in the building. Thus, my chance of getting B is 1—p. By contrast, still on the right side of the inequality, when I inquire, what's the probability of H occurring? Zero. After all, if I find out it's true that H will occur, then I won't do A. So if I investigate, I definitely will prevent H no matter what. In actuality, there is just a 1−p chance that I'll end up being able to do A to get B for myself. So that's how we get the right side of the inequality.

Now we can solve for s to figure out how culpable I am for doing A in willful ignorance:

Derivation of the actor's s-value in a core case of willful ignorance:

$$\text{Value (not}-\text{inquire and then do A)} > \text{Value (inquire and then maybe}-\text{or}-\text{maybe}-\text{not do A)}$$

$$B*1-s*H*p > B*(1-p)-s*H*0$$

$$B-s*H*p > B*(1-p)$$

$$B-s*H*p > B-B*p$$

$$-s*H*p > -B*p$$

$$s*H*p < B*p$$

$$s < B*p/H*p$$

$$s < B/H$$

This is an intriguing result. In the core case of willful ignorance, we can be sure that my s-value is below the very same threshold that we earlier saw the s-value for a knowing wrongdoer is below: namely, B/H. Thus, when I don't inquire and then do act A, and if it is a core case, then we can be sure that my s-value is below the *very same threshold* as it would be if I knowingly did A. In both cases, it's below B/H. This is Yaffe's proof that sometimes acting in willful ignorance can be as culpable as acting knowingly. It is a major contribution to our understanding of willful ignorance, and I nothing I say below will call it into question for the core case.[16]

[16] So far, I've assumed that the only harm in my doing A would be the death of a person—as would be the case if the building in question belonged to me. But if the building belongs to another, there would be an additional harm, H', in doing A—i.e. property damage. If I inquire, learn the building is empty and do A,

2.2.2 *Willful Ignorance with Costly Investigation*

Before proceeding, note how Yaffe's model can be extended beyond the core case. As we just saw, the core case involved simplifying assumptions, and by relaxing these assumptions, we can capture more and more cases. Consider one that will be particularly important below: costly inquiry.

In the core case, investigating the inculpatory proposition—*i.e.*, whether the suspected harm will actually ensue—was assumed to entail no cost whatsoever for the actor. But that's not always realistic, of course. In real life, investigating will generally cost something (even if just a bit of time and effort). So how do we model a case of willful ignorance where investigation is costly?

To do it, we need to add another term to the Value Equation on the right side of the inequality. Let *Q* refer to the costs to the actor himself of inquiring.[17] In this case, the right hand side of the inequality—*i.e.*, the value of inquiring and then maybe or maybe not doing act A (arson)—would be this: = B*(1 − p) − s*H*0 − Q. Here, I've just added the variable Q on at the end as an additional cost to the actor. Thus, we can derive the actor's s-value in a case of willful ignorance where investigation is costly in the following way:

Derivation of the actor's s-value in a costly-inquiry case of willful ignorance:

$$\text{Value (not}-\text{inquire and then do A)} > \text{Value (inquire and then maybe}-\text{or}-\text{maybe}-\text{not do A)}$$
$$B*1-s*H*p > B*(1-p)-s*H*0-Q$$
$$B-s*H*p > B-B*p-Q$$
$$-s*H*p > -B*p-Q$$
$$s*H*p < B*p+Q$$
$$s < (B*p+Q)/H*p$$
$$s < B*p/H*p+Q/H*p$$
$$s < B/H+Q/H*p$$

Note that what we end up with for my s-value here is a bigger number than it was in either the case of knowing misconduct or the core case of willful ignorance (where we could infer

Footnote 16 continued
then while the death, H, would not result, the property damage, H', still would. I omitted this from the analysis for simplicity and because Yaffe also seems to. But it changes nothing of substance. The proof still goes through if we include H'. Start with knowingly causing both the death and property damage, i.e. H and H': Value (A) > Value (~ A) = B*1–s* (H*1+H'*1) > B*0–s*(H*0+H'*0). Solving for s, we get s < B/(H + H'). Now consider doing A while willfully ignorant of whether H (the death) will occur, but while knowing that H' (the property damage) will occur. As before, if I don't investigate before doing A, I am certain to receive benefit B and there is a p chance that H will occur. But now there is also a probability of 1 that H' will occur. Thus, Value(not inquire and do A) = B*1–s*(H*p+ H'*1). Furthermore, upon investigating, I will do A only if I learn that no one is in the building. Thus, the probability of H occurring is 0, while the probability is (1-p) both that I'll receive benefit B and that the property damage H' will occur. Thus, Value(inquire and then maybe-or-maybe not do A) = B*(1-p)–s*(H*0+H'(1-p)). Since I don't investigate and do A, we know that Value(not-inquire and do A) > Value(inquire and then maybe-or-maybe not do A) = B*1–s*(H*p+H'*1) > B*(1-p)–s*(H*0+H'(1-p)). Solving for s, we again get s < B/(H + H')—just as in the case of knowingly causing H and H'. So Yaffe's proof goes through even taking account of the additional property damage.

[17] What about costs to others of inquiring? Perhaps it might hurt others to inquire in some cases. Presumably this should be allowed to affect the culpability calculus as well. It would work in an analogous way to cases where inquiry is costly to oneself.

my s-value was below B/H). Thus, we end up with the intuitively correct result that it is *less* culpable to fail to inquire when doing so costs something. Accordingly, we have a nice illustration of how relaxing the assumptions of the core case in various ways can yield interesting results.

3 A Technical Problem

Now on to some criticisms. My aim is not to refute Yaffe's view. On the contrary, he makes a major contribution to our understanding of culpability and willful ignorance. Instead, I highlight some worries that remain around the edges of his view. I begin with the narrowest and most technical criticism.

Return to where we just left off: the case of willful ignorance where investigation is costly. Call this *Costly Inquiry WI*. The worry is that Yaffe's argument does not provide a watertight proof that Costly Inquiry WI is just as culpable as the core case of willful ignorance—which I dub *Core WI*—and thus as culpable as knowing misconduct—which I call *K*.

Here is why. For Core WI, as with K, we can be sure that the actor's s-value is below B/H. However, for Costly Inquiry WI, we can only be sure that the actor's s-value is below a higher threshold, which at the end of the last section we saw was B/H + Q/H*p. This means that in Costly Inquiry WI, the actor's s-value is only guaranteed to be below a *higher* threshold than the threshold it's going to be below in Core WI and K.

This, of course, does not mean that *no* cases of Costly Inquiry WI are as culpable as the analogous K or Core WI cases. Of course, it's possible that there are cases where this is true. However, the trouble is that Yaffe's proof does not give us any *guarantee* that there will be Costly Inquiry WI cases that are as culpable (*i.e.*, where the actor's s-value is as low) as in Core WI or K. All we know is that in Core WI and K cases, the actor's s-value is below B/H, while in Costly Inquiry WI cases, the actor's s-value is below B/H + Q/H*p. However, consistent with this, it remains logically possible that the actor's s-value is lower in all Core WI and K cases than it is in all Costly Inquiry WI cases. That would mean all Core WI and K cases could in principle be more culpable than all Costly Inquiry WI cases. Thus, we don't have a watertight proof that Costly Inquiry WI is guaranteed to be as culpable as Core WI or K. Of course, in reality, it is very unlikely that this logical possibility would obtain. But the trouble is that for all Yaffe's proof establishes, this possibility is not *ruled out*. For that reason, the proof provides no watertight *guarantee* that Costly Inquiry WI is just as culpable Core WI and K.

Yaffe seems aware of this worry, and he tries to block it with the following line of reasoning:

> [I]magine that it would cost D $5 to inquire, that he thinks that the probability of [the inculpatory proposition] X is 0.1, and that his A-ing will cost others $1000, if X. The question, then, is this: would an agent who knowingly A'd (A'd while knowing X) be diminished in his culpability had he done it, in part, to avoid paying $5/($1000 * 0.1) = five cents? The answer is ''no'' and so we learn that the agent who A's at t2 without inquiring at t1 in such circumstances is properly treated as though he knowingly A'd at t2.[18]

[18] Yaffe, *supra* note 3 at 22–23.

This does not seem to plug the gap in the argument, however. Granted, the person Yaffe envisions here—who does the crime knowing that it will cause $1000 of harm to others, where part of his motivation is to avoid incurring a $5 cost to himself—still seems very culpable. But the Value Equation still allows us to infer that this person is *slightly* less culpable than someone who did the crime knowing it would cause $1000 of harm to others in order to avoid, say, a $1 cost to himself. In other words, when the Value Equation is strictly applied, the threshold we can be sure the actor's s-value is below is slightly *higher* for those who knowingly impose $1000 of damage to avoid $5 than it is for those who do the same thing to avoid a $1 cost to himself.

What this means is that, for all Yaffe says, we still have no *guarantee* that Costly Inquiry WI will be as culpable as Core WI, and thus by extension as culpable as K. After all, under the Value Equation, we can only be sure that doing the crime after deciding not to investigate in costly ways (*i.e.*, Costly Inquiry WI) entails a slightly less bad s-value than doing the crime after deciding not to investigate in totally cost-free ways (*i.e.*, Core WI). The burdens of investigating are slightly greater in the former case than in the latter. So the s-value manifested, on Yaffe's view, is slightly higher for Costly Inquiry WI than for Core WI. Thus, the culpability level we can infer in the former case is slightly less bad than it is in the latter. Thus, we don't have a watertight proof that Costly Inquiry WI is guaranteed to be as culpable as Core WI and K.

4 Two Substantive Problems and a Partial Fix

This was only a technical problem, since in reality we can be confident that many cases of Costly Inquiry WI actually will have very low s-values—indeed, roughly as low as in Core WI or K cases. Thus, it is going to be quite safe in numerous real life scenarios to treat Costly Inquiry WI as roughly as culpable as the core case of willful ignorance, and thus as the analogous case of knowing misconduct. However, there are more substantive worries about Yaffe's view of culpability, as he formulates it in this paper. Let me present two—one concerning the relevance of motives, and the other concerning unjustified but privileged actions—before discussing a possible solution to both problems. The solution relies on the notion of legally recognized reasons. Nonetheless, I'll argue that difficulties remain for this proposed solution when it comes to implementing it in Yaffe's mathematical model in a coherent way. Accordingly, the solution is not yet a complete fix for the problems Yaffe's view faces.

4.1 Objection 1: Motive Sensitivity

The first substantive objection is that on Yaffe's view, the s-value is motive-sensitive in a way that criminal culpability generally is assumed not to be—at least as a matter of posited law. Wayne LaFave's well-known treatise observes that "motive, if narrowly defined to exclude recognized defenses and the 'specific intent' requirements of some crimes, is not relevant on the substantive side of the criminal law."[19] Thus, a thief is not guilty of a more serious offense because he is motivated by hatred of the victim rather than wanting to be able to afford to send his kid to a better school.[20] Affirmative defenses function similarly: "when an individual finds himself in a position where the law grants him the right to kill

[19] Wayne Lafave, 1 Subst. Crim. L. § 5.3 (2d ed.).

[20] *Id.*

another in his own defense, it makes no difference whether his dominant motive is other than self-preservation."[21] Of course, some offenses are defined to make a bad motive an element (as in treason,[22] or hate crimes[23]). Perhaps motives also influence sentencing proceedings.[24] Nonetheless, the *general rule* is that motives are not relevant to substantive criminal law doctrine—which, I take it, is the main thing that should be informed by and match up with criminal culpability. One might think the law is normatively flawed in not taking motives to be relevant. But as a matter of posited law, the default rule is that motives usually don't matter to criminal law doctrine applicable at the guilt stage.

Yaffe's account of how to calculate the s-value, however, does not sit comfortably with the default rule that motives usually don't matter to criminal culpability. Here is why. Take a case of knowing misconduct: someone offers me $1000 to burn down a building with a person inside, which I know will kill him. Thus, let *H* stand for the disvalue of the death of a person. The term *B* will refer to the benefit I seek in doing the act, *i.e.*, $1000. Thus, applying the Value Equation, my s-value would be below $1000/H. But now compare a case where someone offers me $2000 to do the same job. Now the benefit I seek is twice as great. Thus, all we know is that my s-value now is below a *higher* threshold than in the first case. That is, now we know my s-value is below $2000/H—a higher threshold than when I was seeking $1000. As a result, I manifest less culpability in the second case than in the first. More precisely, we can only infer that my s-value is below a higher threshold in the second case than we can in the first.[25]

The upshot is that on Yaffe's view, the specific benefit one seeks in performing a criminal action—which is to say, one's motive in breaking the law—will always have a direct impact on one's inferable s-value, and thus on one's criminal culpability. This conflicts with the default rule that motives—*i.e.*, one's particular aims in acting—don't directly bear on one's degree of criminal culpability. At least, they don't directly bear on the sort of culpability that matters at the guilt stage of the trial (even if motives might matter at sentencing). As a result, Yaffe's view suggests that culpability generally is motive-sensitive. So it has a hard time explaining why substantive criminal law doctrine, as a general rule, is motive-*in*sensitive.

This is admittedly not a decisive objection to Yaffe. It is open to him to argue that his motive-sensitive view of culpability is correct as a normative matter, and that the motive-insensitivity of posited criminal law is a mistake. Nonetheless, all else equal, a view of culpability that fits with the general contours of posited criminal law is to be preferred over views that sit uncomfortably with important features of the criminal law. Moreover, Yaffe's view has an additional explanatory burden to discharge. If the normatively correct view is that criminal culpability should be motive-sensitive (as Yaffe's view suggests),

[21] *Id.*

[22] Treason requires purpose to aid an enemy of the state, and mere knowledge that this will result does not suffice. Haupt v. United States, 330 U.S. 631, 641–42 (1947); Wayne LaFave, 1 Subst. Crim. L. § 5.2 (note 9).

[23] Hate crimes carry harsher punishments if one acted "because of the actual or perceived race, color, religion, or national origin of any person." 18 U.S.C. § 249.

[24] *See* Carissa Byrne Hessick, *Motive's Role in Criminal Punishment*, 80 S. Cal. L. Rev. 89, 90 (2006).

[25] We saw in the previous section that a related point follows from the Value Equation as applied to Costly Inquiry WI. Specifically, we saw that the more costly it is to investigate, the less culpable you are for deciding not to do so before performing the underlying risky act. In other words, for Costly Inquiry WI, we know that $s < B/H + Q/H*p$. Thus, as the cost of investigating—Q—goes up, so does the threshold below which we know your s-value must fall.

how did actual criminal law get it so wrong as to adopt motive-*in*sensitivity as the default rule?[26]

4.2 Objection 2: Unjustified but Privileged Actions

A related worry about Yaffe's view concerns actions that are unjustified but privileged. On one common view, some cases of self-defense can be understood this way. Consider an act of self-defense that isn't justified on lesser evils grounds. For instance, suppose you are about to inflict grievous bodily harm on me—*e.g.*, breaking my leg—so I use lethal force to avert that harm. Suppose this is the only way I could avoid suffering this imminent harm. Under the Model Penal Code ("MPC") § 3.04(2)(b), I could use the defense of self-defense, which (roughly) permits me to use lethal force if I believe it necessary to avert imminent "serious bodily injury."[27] Although the MPC says my act would be "justified" here, I will stick to the standard philosophical usage of that term, which takes it that, for an otherwise wrongful act to be justified, its benefits must outweigh the harm or disvalue it carries with it. That is not what we have in this scenario. Thus, a plausible way to describe my action here is that it is unjustified (in the philosophical sense), but I have a privilege or permission to use lethal force to prevent the lesser harm to myself.

To make this more precise, suppose my act, A, would kill you, my attacker, thus bringing about 500 units of harm. Moreover, suppose it provides me no affirmative benefits—*i.e.*, it just avoids the harm that would befall me if I refrained from doing A. By contrast, if I don't do A, then I'd suffer 300 units of harm from your breaking my leg. Suppose these consequences are certain, such that p1 and p2 = 1. Plugging this into the Value Equation, we get the following:

Derivation of the actor's s-value in a case of unjustified but privileged harm:

$$\begin{aligned}
&\text{Value}(\text{A}) > \text{Value}(\sim\text{A})\\
&\text{EV}_{\text{Self}}(\text{A}) + \text{EV}_{\text{Others}}(\text{A}) > \text{EV}_{\text{Self}}(\sim\text{A}) + \text{EV}_{\text{Others}}(\sim\text{A})\\
&0 - \text{s} * 500 * 1 > -300 + 0\\
&\text{s} < 300/500
\end{aligned}$$

Thus, what follows on Yaffe's view is that we have reason to infer that my act of self-defense is somewhat culpable. After all, given the (I think not implausible) numerical assumptions just made, we can infer that my s-value is less than 1, as required for being culpable. Nonetheless, I submit that this is incorrect. By hypothesis, I am *privileged* to inflict this amount of harm in self-defense. Some might insist that I'm a bit morally blameworthy because I failed to do what I knew would maximize value in the world. But that does not seem to be the law's view, at least.

[26] One might offer a practical explanation based on the idea that it is simpler and more cost-effective for the criminal law not to care about motives at the guilt stage. However, this is not satisfying since even when it is easy and costless to identify the defendant's motives, criminal law doctrine remains unaltered in its indifference to one's motives for breaking the law (provided they don't amount to a defense). Thus, a more principled explanation of the motive-insensitivity of substantive criminal law doctrine is preferable. Yaffe's view as stated doesn't provide one.

[27] MPC § 3.04(2)(b) ("The use of deadly force is not justifiable under this Section unless the actor believes that such force is necessary to protect himself against death, serious bodily injury, kidnapping or sexual intercourse compelled by force or threat.").

4.3 A Possible Solution: Restricting the Culpability Calculus to Legally Recognized Reasons

Let me now explore a way to amend Yaffe's view that promises to help with both of the problems just presented. Ultimately, however, I think difficult questions remain for how to implement this fix in Yaffe's mathematical model in a coherent way.

Begin with the problem of motive-sensitivity. Recall the case of burning down a building despite knowing it will cause a death in order to receive \$1000 versus doing so in order to receive \$2000. The solution I propose is to take it that the benefit sought here—whether it be \$1000 or \$2000—does not count as a *legally recognized* benefit. That is, since the law does not regard my receipt of any amount of money for personal ends as giving me any—even partial—justification for burning down a building with a person inside, I would have no *legally recognized* reason under the circumstances to burn down the building. Thus, regardless of whether my motive for doing this criminal action is to obtain \$1000 or \$2000, from the law's perspective, the amount of culpability that should be attributed to me is the same. In general, then, the idea is that rather than taking culpability to be a function of the *actual* balance of the probability-weighted burdens and benefits to oneself and to others that one expects to result from the action, culpability should instead be a function of these burdens and benefits *to the extent they are legally recognized*.[28] Thus, to find one's culpability, we start with the benefits and burdens to oneself and others that one believes may result from one's action, but then exclude those that are not legally recognized.

In the arson example, we can model this in Yaffe's view by setting the legally recognized benefit that I seek, viz. B, as equal to 0 regardless of how much money I believe I would obtain from setting fire to the building with a person inside. Hence, regardless of whether I seek \$1000 or \$2000 for doing the crime, my culpability would be calculated as follows:

Derivation of my "cleaned up" s-value in a case of knowing misconduct:

$$\begin{aligned}
&\mathrm{Value(A)} > \mathrm{Value}(\sim \mathrm{A})\\
&\mathrm{B} * 1 - \mathrm{s} * \mathrm{H} * 1 > \mathrm{B} * 0 - \mathrm{s} * \mathrm{H} * 0\\
&0 - \mathrm{s} * \mathrm{H} * 1 > 0 - 0\\
&-\mathrm{s} * \mathrm{H} > 0\\
&\mathrm{s} < 0/\mathrm{H}\\
&\mathrm{s} < 0
\end{aligned}$$

While initially promising, problems with this solution remain. Specifically, an interpretive question arises here about what it means for my s-value to be below 0. Strictly speaking, an s-value of 0 would mean that the agent is willing to impose serious harm on others (equal to death) in return for *no benefit whatsoever for himself*. After all, we saw above that the s-value is meant to capture the exchange rate between a unit of benefit for oneself and a unit of harm for others. However, it is unrealistic to suppose that criminal actors would routinely have an s-value that is *below* 0. Such an s-value would mean they are willing to impose serious harm on others even if it would *cost* them something—*i.e.*, even if they have to pay to do it. Criminals may well possess insufficient regard for others, but it is the

[28] This is similar to the view of criminal culpability I have defended elsewhere. *See* Alex Sarch, *Who Cares What You Think? Criminal Culpability and the Irrelevance of Unmanifested Mental States* (manuscript on file with author).

rare case where such an actor would be willing to cause serious harm and even death for no reason whatsoever—and indeed even if it costs the actor something to bring about the death. It would be surprisingly harsh even for our imperfect criminal law to routinely treat criminal actors as if they had negative s-values. Nonetheless, this is what follows on Yaffe's view—under the proposed amendment.

The natural way to avoid this worry, then, is to maintain that the law doesn't treat the receipt of, say, $1000 as *no reason at all* in favor of burning down the building with a person inside, but rather as a very small one. To capture this in Yaffe's model, we would have to introduce a new factor, w, that adjusts the weight of the benefits sought in performing a criminal act. Here, the legally recognized weight of the reason provided by seeking $1000 in return for burning down a building would be something extremely small. Thus, this factor, w, would have to be something extremely small too: perhaps 0.001. This would still make my s-value rather low, but it wouldn't be literally below zero. (We could adjust the weight of factor w to generate as plausible results as we can.) Thus, the arsonist's culpability would amount to the following (where H is the disvalue of a death and B is the value of the monetary gain sought from the crime):

Amended derivation of the actor's "cleaned up" s-value in a case of knowing misconduct:

$$\begin{aligned}
&\text{Value}(A) > \text{Value}\ (\sim A)\\
&w * B * 1 - s * H * 1 > w * B * 0 - s * H * 0\\
&0.001 * B - s * H\ > 0 - 0\\
&-s * H > -0.001 * B\\
&s < 0.001 * B/H
\end{aligned}$$

Although this expression 0.001*B/H is very close to 0, it doesn't quite equal 0.

Unfortunately, this further amendment is not ideal either. It reintroduces the very same motive-sensitivity we wanted to avoid in the first place. If we introduce the w factor as indicated, then the value of the benefit one seeks in acting again gets to matter to the level of culpability imputed to one. Thus, seeking a benefit in doing the crime that is twice as great—say, $2000 instead of $1000—means that one's s-value is now below a threshold that is twice as high. That is, one's s-value would be beneath the higher threshold of w*2*B/H rather than the lower threshold of w*B/H. Thus, one would be less culpable if one does the crime to get $2000 rather than to get $1000. That is the very sort of motive-sensitivity we aimed to avoid from the outset.

Would it therefore be better to revert to the initial version of the solution? Note that we now have two versions of the solution on the table. The first would exclude from the culpability calculus any of the reasons for or against one's action that do not count as legally recognized. Call this the *Exclusion Approach*. The second would give a heavily discounted weight to the reasons that do not count as fully legally recognized. Call this the *Weighting Approach*. The Weighting Approach is in trouble because it reintroduces the motive-sensitivity that we sought to avoid in the first place. So the question is whether we should revert to the Exclusion Approach. One might think its problems are less damning.

However, that would be premature. For the Exclusion Approach seems fatally flawed if we use it to try to solve the second big objection from above, discussed in Section 4.2: namely, the problem of unjustified but privileged actions. Here is how the basic idea would be applied to the self-defense example from earlier. We might say that the lethal harm I impose—the 500 units of disvalue that my attacker's death would have—does not count as legally recognized

because this harm is privileged under the circumstances. Thus, it does not provide me with any *legally recognized* reason to refrain from my lethal action. Accordingly, under the Exclusion Approach, it would not factor into the calculation of my culpability. By contrast, the law *would* recognize the harm you, my attacker, impermissibly seek to inflict on me—*i.e.*, the broken leg. The law would take this to be a genuine cost to my *not* using force (*i.e.*, doing nothing). Thus, this cost would indeed factor into the culpability calculation.

This all sounds good in theory, but trouble arises when we seek to implement this reasoning mathematically using Yaffe's model. Specifically, on the Exclusion Approach, because the harm to you (my attacker) does not count as a reason to abstain from my privileged use of lethal force, H (on the left) is to be given a value not of 500, as in the initial example, but rather of 0, since it is to be excluded. Moreover, there are no overt benefits to me if I do A, so the benefit B is 0. But the damage I'll suffer if I don't do A—call it *D*—equals −300. Accordingly, what we get is this:

Derivation of my "cleaned-up" culpability for the unjustified but privileged harmful act (Exclusion Approach):

$$\begin{aligned}
&\mathrm{Value(A)} > \mathrm{Value}(\sim \mathrm{A})\\
&\mathrm{EV_{Self}(A)} + \mathrm{EV_{Others}(A)} > \mathrm{EV_{Self}}(\sim \mathrm{A}) + \mathrm{EV_{Others}}(\sim \mathrm{A})\\
&\mathrm{B} * 1 - \mathrm{s} * \mathrm{H} * 1 > \mathrm{D} * 1 - \mathrm{s} * \mathrm{H} * 0\\
&0 - \mathrm{s} * 0 * 1 > -300 - 0\\
&\mathrm{s} < 300/0\\
&\mathrm{XXX}
\end{aligned}$$

The trouble, of course, is that 300/0 is undefined. Thus, Yaffe's model does not yield sensible results in such cases if the Exclusion Approach is adopted.

This strongly suggests that the legally unrecognized costs and benefits to acting should not be fully excluded from the culpability calculus, as the Exclusion Approach suggests. Instead, it seems more promising to discount their weight as the Weighting Approach implies. Thus, we'd have to apply a w factor that heavily discounts the weight of the reasons that are not fully legally recognized. Since my lethal action in self-defense is privileged, the w factor must impose a very heavy discount to the disvalue of H. Thus, let's again suppose that $w = 0.001$. So we get:

Derivation of my "cleaned-up" culpability for the unjustified but privileged harmful act (Weighting Approach):

$$\begin{aligned}
&\mathrm{Value(A)} > \mathrm{Value}(\sim \mathrm{A})\\
&\mathrm{EV_{Self}(A)} + \mathrm{EV_{Others}(A)} > \mathrm{EV_{Self}}(\sim \mathrm{A}) + \mathrm{EV_{Others}}(\sim \mathrm{A})\\
&\mathrm{B} * 1 - \mathrm{s} * \mathrm{H} * \mathrm{w} * 1 > \mathrm{D} * 1 - \mathrm{s} * \mathrm{H} * 0\\
&0 - \mathrm{s} * 500 * 0.001 * 1 > -300 - 0\\
&\mathrm{s} < 300/0.5
\end{aligned}$$

Accordingly, on the Weighting Approach, we do not reach absurd results about my culpability. Instead, with these numerical assumptions, my act of self-defense would not reveal that I have any culpability. The action does not show my s-value to be below 1, as is plausibly required for being culpable. The act is compatible with my s-value being extremely high—at least by the law's lights. Thus, my action would manifest no criminal culpability. This is the intuitively correct result, since I have a privilege in this case to inflict the 500-unit harm on you.

When all is said and done, we face a dilemma if we want to use the idea of legally recognized reasons to help solve the two main problems sketched above. If we adopt the Exclusion Approach to revising Yaffe's model, we run into two difficulties. First, in the arson example, this entails that the criminal actor has an s-value below 0, which is an implausibly harsh verdict. Second, this approach yields mathematically troubling results as applied to the self-defense case. By contrast, if we adopt the Weighting Approach to revising Yaffe's model, we run into other difficulties. Most importantly, it reintroduces the very motive-sensitivity into the culpability calculus that we wanted to avoid from the outset. (This approach also may seem worryingly *ad hoc*.[29]) Thus, although I am inclined to think that adopting the concept of legally recognized reasons is a promising direction for Yaffe's theory to take, more work is needed to implement the idea in a coherent and principled way. The question is worth exploring further.

5 Additive versus Non-additive Views

Let me end by returning to willful ignorance. What might at first appear to be an advantage of Yaffe's view of willful ignorance over other accounts, on closer inspection, turns out not to be a genuine advantage after all. To show this, compare the account of willful ignorance I have defended elsewhere.[30]

My account starts from the claim that when one plans to perform an action that one suspects will impose risks of harm on others, there is a *pro tanto* obligation (a weighty reason) to investigate—in reasonably available ways—whether these risks will actually materialize if one proceeds to do the act in question. Failing to do so, I claim, can manifest insufficient regard for others and therefore increase one's culpability. For instance, if one plans to burn down a building, but suspects a person might be inside who would die in the fire, one has a weighty reason to investigate in available ways whether this is so. If one refuses to, this would increase one's culpability beyond what it would be if one had no way of investigating these suspicions and started the fire while merely reckless with respect to the fact that someone is inside. My view claims that the culpability inherent in doing the underlying *actus reus* purely recklessly (*i.e*., lighting the fire while aware of the risk of death) can be *added* to the extra culpability inherent in deciding not to investigate in available ways first. Sometimes, I claim, the extra insufficient regard manifested in deciding not to investigate one's suspicions can be enough to raise one's culpability up from the level of a purely reckless actor (where no investigation is possible) to the level of the exactly analogous knowing actor. In such cases, but only in such cases, the equal culpability thesis would be true: acting in willful ignorance would be at least as culpable as the analogous knowing misconduct. Call this the *Additive View*, since it presupposes that the culpability of refusing to investigate as one ought to can be added to the culpability of doing the underlying risky action in a state of recklessness.[31]

[29] On the Weighting Approach, we also need some non-*ad hoc* way to determine the weight that should be used when applying the discount factor w to the costs and benefits of the action that are not fully legally recognized. I know of no obvious theory or principle that would let us make progress on this question. So the worry that the Weighting Approach would prove disturbingly *ad hoc* also looms large.

[30] *See* Sarch, *supra* note 4.

[31] Yaffe raises a separate objection to my account in his article. *See* Yaffe, *supra* note 3 at 18 (note 17). He contends that breaching a duty of inquiry is not *necessary* for willfully ignorant misconduct to be as culpable as the analogous knowing wrongdoing, even though on my account it would be. As Yaffe puts it, "[a] case

Some have questioned the sort of addition involved in accounts like mine.[32] That is, one might think there is something fishy about my claim that we can add the culpability of refusing to investigate at t1 together with the culpability of the underlying criminal act at t2. One might question my argument that the sum of these two quanta of culpability can sometimes be taken to equal the amount manifested in doing the same act knowingly, such that the criminal law could treat one's willful ignorance as the functional equivalent of knowledge. After all, what licenses this addition? Aren't these two bits of behavior (the refusal to investigate and the subsequent reckless action) different in nature? Why do they get to be combined to form a single unit of conduct that can fairly be the basis of a single culpability attribution? I have attempted to answer these concerns in other work,[33] but I recognize they are a challenge for me.

At first sight, Yaffe's view might seem to avoid these worries altogether. It might seem that he provides an elegant way to avoid having to rely on this kind of addition. Yaffe shows both that (1) for knowing misconduct, we can be confident that one's s-value will always be below B/H, and (2) in core cases of willful ignorance, we can also be sure that the defendant's s-value is below B/H. We can deduce this from the Value Equation plus the essential features of these cases without having to rely on any addition of culpability. So one might think this is an advantage of Yaffe's account of willful ignorance compared to my Additive View.

On closer inspection, however, I think Yaffe also has to invoke some kind of addition in order to get his view to be a satisfactory explanation of the phenomenon under investigation. In particular, let's focus on the *metaphysical* question of how it could be that the

Footnote 31 continued
can be core [such that the willfully ignorant defendant would be as culpable as a knowing wrongdoer] without the defendant having any duty to inquire." *Id*.

Nonetheless, I think Yaffe's objection does not actually threaten my account—at least if I am clear about how I understand the duty to investigate. All I want to say is that when one is in certain triggering conditions—*i.e*., when one suspects the inculpatory proposition, p, of a crime to be true, and one reasonably can and should inquire whether p—then deciding not to investigate *is a* pro tanto *contributor to one's culpability*. This doesn't mean that deciding not to investigate always makes one on balance more culpable, or that not inquiring would always be wrong. I do not claim that there is a duty to inquire in this strong sense. Rather, all I want to say is that when one is in the relevant conditions and fails to inquire, this *pro tanto* contributes to one's culpability. This implies that there is a duty to inquire only in a very weak sense.

When the notion of a duty to inquire thus is clarified, I think there is actually no substantive disagreement between myself and Yaffe on the role of "duties" of inquiry, understood in the weak sense. Both Yaffe and I think that failing to inquire when one should can render one more culpable, and I simply add the further claim that there's a sense in which one, all else equal, should—*i.e*., has a weak *pro tanto* "duty" to—avoid that source of extra culpability. As a result, I do not see any cases where (1) one would be more culpable for failing to inquire, but (2) there is no "duty" in the weak *pro tanto* sense to inquire (*i.e*., where inquiring isn't something one, all else equal, *should* do to avoid extra culpability). If so, then it seems to me that at least breaching this kind of weak *pro tanto* duty to inquire really is necessary for WI to be as culpable as K. Thus, Yaffe's objection to my account does not succeed.

[32] *See* Paul H. Robinson, *Imputed Criminal Liability*, 93 Yale L.J. 609, 650 (1984) (arguing that "imputation of the elements of a serious offense can be justified on a finding of equivalent culpability reached by aggregating the actor's culpability for two or more less serious offenses," but worrying that the conduct aggregated must be closely related or else we face "an unacceptable precedent: permitting such aggregation in all cases of multiple offenses"); Commonwealth v. Life Care Centers of America, 926 N.E. 2d 206, 213 (Mass. 2010) (arguing that "aggregation of less culpable behavior to create more culpable behavior not only is illogical but also raises due process concerns").

[33] Alex Sarch, *Beyond Willful Ignorance*, 88 U. Colo. L. Rev. 97, 136–138 (2016) (arguing that the components of willfully ignorant action are fairly regarded as one "course of conduct"). *See also* Robinson, *supra* note 31 at 650 (endorsing aggregation on similar grounds).

s-value for the actor in a core case of willful ignorance could be at least as low as the actor's s-value in the analogous case of knowledge. Granted, we *know* this is true thanks to Yaffe's proof that the s-value in both cases is below B/H. But even granting that result, there is still an explanatory question left over about the metaphysics that underwrite this result. That is, what is the *mechanism* that explains why a core case of willful ignorance is as culpable as a case of knowledge? My Additive View is one picture that at least provides an answer to this metaphysical question. However, for Yaffe's view to be a complete explanation of willful ignorance, he, too, must answer this question.

Presumably, Yaffe would want to say the answer to this metaphysical question has something to do with the fact that the defendant in the core case of WI does not want to jeopardize his receiving benefit B through investigating, and that this speaks ill of him. It manifests some form of disrespect or insufficient regard for others. I agree that it does, but note that this *by itself* is not enough to explain why this actor's s-value sinks down to the same level as a knowing wrongdoer. We can only draw this inference when we *also* know that the defendant goes ahead and actually performs the underlying risky act (after not investigating). Yaffe's account does not claim that failing to investigate *by itself* makes one as culpable as a knowing actor; it's only failing to investigate *and then doing the risky act* that makes one as culpable as a knowing actor. As a result, there is reason to think that Yaffe also covertly relies on something like the additive metaphysical picture I've defended. His view, too, appeals to the conjunction of not investigating together with subsequently doing the underlying risky action.

Thus, I doubt that Yaffe's view actually amounts to an alternative *metaphysical* picture that explains why Core WI could be as culpable as K. Indeed, it is not clear what alternative answers to the metaphysical question might look like apart from the Additive View.

This is not to cast aspersions on Yaffe's mathematical proof that the culpability of Core WI equals that of the analogous case of knowing misconduct. That is an interesting and exciting result. In this section, I have merely sought to show some limits of this proof. It establishes a particular culpability result, but does not go all the way to explaining the metaphysical picture that would explain why that result holds. For that, we may have to appeal to the sort of additive picture I have defended, or else formulate an alternative metaphysical picture that can do the job.