RISK AVERSION AND ELITE-GROUP IGNORANCE

David Kinney
Santa Fe Institute

Liam Kofi Bright
London School of Economics and Political Science

Forthcoming in Philosophy and Phenomenological Research

Abstract

Critical race theorists and standpoint epistemologists argue that agents who are members of dominant social groups are often in a state of ignorance about the extent of their social dominance, where this ignorance is explained by these agents' membership in a socially dominant group (e.g., Mills 2007). To illustrate this claim bluntly, it is argued: 1) that many white men do not know the extent of their social dominance, 2) that they remain ignorant as to the extent of their dominant social position even where this information is freely attainable, and 3) that this ignorance is due in part to the fact that they are white men. We argue that on Buchak's (2010, 2013) model of risk averse instrumental rationality, ignorance of one's privileges can be rational. This argument yields a new account of elite-group ignorance, why it may occur, and how it might be alleviated.

1. Introduction

Willful ignorance of social inequality is a frustrating feature of life in highly unequal societies. Despite the ubiquity of available information on the extent of social inequalities, those at the top of social hierarchies regularly display ignorance of the challenges faced by those at the bottom. For instance, a 2016 Pew Research Center poll found that only 50% of white Americans, as compared to 88% of Black Americans, believe that Blacks are treated less fairly than whites in interactions with the police.[1] However, there is a slew of evidence suggesting black people are treated less fairly in their/our interactions with the justice system. To sample just a small part, a comprehensive study of de-identified longitudinal data covering nearly the entire U.S. population from 1989-2015 found that "21% of Black men born to the lowest-income families are incarcerated on a given day, as compared with 6% of white men [born in the same income band]." Further, among men born to parents in the top 1% income band, "only 0.2% of white males were incarcerated, whereas 2.2% of Black males

[1] It remains to be seen whether the 2020 protests against the killings of George Floyd, Breonna Taylor and many others will significantly change this state of affairs.

were incarcerated" (Chetty 2018, p. 23). Relatedly, Ross (2015) finds "evidence of a significant bias in the killing of unarmed Black Americans relative to unarmed white Americans" even when factors such as local crime rate are taken into account. In addition, Lum and Isaac (2016) argue that new techniques of predictive policing used in California are biased so that even with approximately equal rates of drug use between Black and white residents, the police in Oakland, California would still be disproportionately likely to arrest Black citizens for drug use. Finally, Camp et al. (2021) recently produced evidence that police adopt a harsher more confrontational tone when interacting with black as compared to white drivers. Such a pattern of findings suggests that there is a disparity in the way that Blacks and whites are treated in the U.S. criminal justice system that is not explained by other potentially salient factors like income or even crime rates. Nevertheless, half of white Americans apparently do not believe that such disparities exist. Thus, a case can be made that although many whites benefit from disproportionately positive treatment from the criminal justice system, they do not believe that they possess this privilege. Similarly, Kraus et al. (2019) have found that white Americans systematically underestimate the extent of the racial wealth gap in the United States, with 89.3% of white Americans overestimating the amount of wealth held by Black Americans by at least twenty percentage points.

In addition to being frustrating, the fact that members of elite groups are sometimes ignorant as to their own level of privilege can have serious social consequences (we take an elite group to be any group of individuals that enjoy an advantage over others with respect to some salient dimension of well-being). First, very often this ignorance is an impediment to political decision-making that could begin to alleviate harms caused by social hierarchies. Given that those wielding power tend disproportionately to come from socially powerful groups, ignorance on the part of the powerful as to their privileged status hinders their ability to ameliorate social inequalities. This is because, even if the will to solve social problems exists, those in a position to make policy are likely to be ignorant of pertinent facts that need to be taken into account, such as what actual disparities exist and how severe they are. Second, however, such a will to solve problems cannot be assumed, since for the very same reason those in a position to arrange the payment of reparations or secure restitutive justice by other means are more likely to be ignorant of the reasons why such reparations may be needed. Hence ignorance by members of elite groups as to their own privileged status can hinder attempts at restitutive justice (Mills, 2007).

Next there are arguments that this ignorance is both a social and personal harm to members of marginalized groups. For, third, it is plausibly a kind of intrinsic harm, an insult to one who is not in some socially privileged groups, to have to live in a society which systematically refuses to see the plain facts about one's lack of privilege. Under these conditions, non-elite agents feel invisible, and this is a harm unto itself (Ellison, 1952). And fourth, some philosophers have argued that the formation of racist beliefs on the part of whites and others at the top of existing racial hierarchies (which could include the false belief that one is not privileged in virtue of one's race) is a harm unto itself at an interpersonal level. This is distinct in at least some respects from the insult of living in an unjust society. See Basu (2019a, 2019b) for arguments that racist beliefs are themselves intrinsically harmful.

The variety of ignorance described here is thus a topic of interest for contemporary social and political philosophers studying the mechanisms through which hierarchies are maintained. Most notably, Charles Mills explicitly discusses the historical and contemporary phenomenon of "white ignorance" of racial inequality, and the role that this white ignorance plays in the maintaining of said inequality. Mills' work on white ignorance can be situated within a broader context of work on the political effects of elite-group ignorance. This broader literature includes work on propaganda that explains how people can be unaware of the true consequences of their normative commitments (e.g., Stanley 2015) or work on silencing which argues that testimony that might disrupt ignorance of an agent's privileged status is often sidelined (e.g., Dotson, 2011).

It might seem natural to think that if widespread ignorance of social inequality is a problem, then public information campaigns are the appropriate way to ameliorate things. And, indeed, many organizations nowadays are investing resources in diversity and inclusion trainings, with nearly 40% of Fortune 500 companies having hired a diversity and inclusion executive since 2015 (Tonneson, 2020). We take it to be an implicit assumption of these strategies that when people are presented with putatively accurate information about material inequalities between social groups, they will actually take it on board and update upon said information, rather than seek to diminish or ignore it. Indeed, the rationality of this kind of information processing is apparently validated by Good's well-known theorem showing that agents who maximize expected utility should always take on board free evidence (1967).

However, in what follows we will give some reason for pause about this anti-bigotry strategy. Using Buchak's (2010, 2013) model of rational risk averse reasoning, we argue that there is a well-motivated theory of rational decision-making that can explain a class of decisions in which agents seek not to learn information about their own privileged status. Of course, Buchak herself notes that her theory of risk averse rationality renders the verdict that it is rational for risk averse agents to avoid free information under some circumstances. Our aim is thus not merely to point this out, but rather to show that there are well-motivated models of situations that privileged agents in an unjustly hierarchical society may plausibly face wherein risk aversion would cause them to wish not to gather more information. The upshot of these efforts is an account of elite-group ignorance that represents agents who remain deliberately ignorant of their own privileges as following rules of rationality under conditions of risk aversion. Such agents could be expected to ignore, diminish, or refuse to update upon information even if it is freely available. In so far as members of elite groups behave like this, we shall argue, mere informational trainings are liable to be a costly and inefficient means of securing a better world.

We thus aim to complement and elaborate upon the central thrust of the social and political tradition described in the previous paragraphs. Like Mills, we will argue that some agents who are members of elite groups engage in motivated ignorance of their own privilege, that this ignorance has social consequences, and that these agents cultivate said ignorance at least partially because they benefit from inhabiting a superior social position. To this we add that seen from both the point of view of an amoral and narrow self-interest, and even sometimes from the point of view of a generally instrumentally rational agent, this motivated ignorance is rationally so maintained. Properly understanding this can help us understand what sort of incentives or policies need to be in place to break this ignorance down, as we discuss in Section 6 of this paper.

Before moving forward, we note that in discussions of possible counterexamples to Good's argument that rational agents should always seek costless information, one can often get bogged down in debates as to whether the information that is putatively ignored is genuinely costless. One of our primary reasons for relying on Buchak's formal framework for instrumental rationality is that it is uncontroversially demonstrable therein that rational agents may pay to avoid information. Thus, while one can argue against our proposal here by questioning the fittingness of Buchak's framework when modeling the dynamics of elite-group ignorance, our use of Buchak's framework ensures that, insofar as our examples are stated in the language of that framework, they represent genuine instances in which agents decline to receive costless information. To head off another misunderstanding of our claim, note that we are not claiming that Buchak's framework is anything like the One True Model of elite-group risk aversion. All that is necessary for our purposes is that sensible decision theories – that is to say, a theory wherein one could act as it would have you act given your beliefs and values without thereby being irrational in an informal sense – can give rise to failures of Good's theorem in non-recherchè scenarios that can plausibly model oppressive social dynamics. Another example of such a decision theory, though not one that we discuss here, is the Γ-maximin decision rule for imprecise probabilities developed by Gärdenfors and Sahlin (1983) and defended by Seidenfeld (2004); see Bradley and Steele (2016) for an illustration of how this decision rule admits failures of Good's argument. Since the class of appropriately sensible decision theories is too vaguely defined to prove general results about when elite-group ignorance can be considered rational, we instead opt to use a single clear instance of such a decision theory, namely Buchak's.

2. Why Elite-Group Ignorance Seems Irrational

When an agent deliberates between a set of actions, with the goal of choosing which action to perform, they rank actions with respect to the *choiceworthiness* of those actions. One action is more choiceworthy than another for an agent if that agent would prefer to perform the first action rather than the second.[2] We presuppose here that a *rational* agent is one who, when faced with a set of possible actions, chooses an action that is maximally choiceworthy.

How should agents rank the choiceworthiness of actions? An influential answer can be found in the axiomatization of expected value theory due to Savage (1972), which we present here in a highly condensed form, glossing over many controversial issues. For a more detailed articulation, see, among others, Resnik (1987), Bradley (2017), and Thoma (2019a). Let A be a set of actions that is assumed to be fixed and finite for the sake of mathematical tractability. Let $S = \{s_1, s_2, \dots, s_n\}$ be a set of states of the world that partitions the set of all possible worlds. Let $P(\cdot)$ be a probability distribution over S, i.e., a function that assigns a probability to each element of S.[3] Throughout this paper, we assume that all probabilities are subjective, i.e., that they represent an agent's degree of belief that the actual world is in a given state. Let $V(\cdot)$ be a function from the Cartesian product $A \times S$ into the real numbers. That is, $V(\cdot)$ assigns a real number to each pair composed of one action and one state of the world. This real number represents the value to the agent of performing a particular action in a particular state of the world. The expected value of an action $a \in A$ is given by the following equation:

$$EV(a) = \sum_{i=1}^{n} P\,(s_i)V(a, s_i)$$

The expected value of each action determines its position in an agent's choiceworthiness ranking. The most choiceworthy action is the action with greatest expected value, the second-most choiceworthy action is the action with second-greatest expected value, and so on. If two actions have the same expected value, then the agent regards them as equally choiceworthy.[4]

In philosophical applications of expected utility theory, one may wish to distinguish between action that is more generally instrumental rational and action that amounts to an efficient pursuit of self-interest. Throughout this paper, we adopt the view that both standard and

[2] We assume throughout this paper that an agent's ranking of actions according to the choiceworthiness of those actions forms a total order over the set of actions.

[3] More precisely, $P(\cdot)$ is a function such that: 1) its domain is an algebra on S, i.e., a set of subsets of S that is closed under union, intersection, and complement, 2) its range is the unit interval, and 3) it obeys Kolmogorov's axioms of non-negativity, unitarity, and countable additivity.

[4] One might object here that our representation of agents as *precise Bayesians* (i.e., as agents whose uncertainty about which element of some partition the actual world is in can be represented as a probability function with precise numerical values) is inherently unrealistic. Real-world agents, the objection might go, do not have mental states corresponding to such point-valued partial beliefs. In response, we hold that *all* social scientific models idealize their targets in some way or another, and that the representation of agents as precise Bayesians is just such an idealization.

risk-weighted expected value theory can be used to model either sort of behavior. Where an agent's value function solely represents the value to that agent of a given state of the world from a purely self-interested perspective, the agent's choosing the action with maximal expected value (or, in what follows, the action with maximal risk-weighted expected value) amounts to the efficient pursuit of self-interest. However, should the agent's value function represent the agent's conative attitudes in a way that goes beyond an agent's broad self-interest (e.g., the agent may value outcomes that have no direct bearing on their physical well-being), then the agent's behavior in accordance with either standard or risk-weighted expected value theory can be understood as broadly instrumentally rational, although perhaps not narrowly self-interested. We take our arguments in what follows to apply to both of these possible representational uses of an agent's value function.

A famous result from Good (1967), which can be read historically as a corollary of earlier work by Blackwell (1953), demonstrates that agents who rank the choiceworthiness of actions according to their expected value will never avoid free information. A more detailed explanation of Good's theorem is given in Appendix A. If one assumes that all rational agents have a choiceworthiness ranking over actions that tracks the expected value of those actions, then elite-group agents who are deliberately ignorant of their group-based privileges are acting irrationally. To illustrate, consider John, who is a white man. John is planning on taking a train journey, and learns that tickets are fifty dollars. John's roommate is away for the weekend, and would not mind if John borrowed his train pass. John knows that if his train pass is closely scrutinized, then he will be fined $250 for using another person's pass. To clarify, train passes in this scenario do not have pictures of the pass-holder's face on them; the issue is not whether John looks like his roommate. Rather, there is some other information on the train pass that will lead a train conductor to become suspicious that John is the passholder if the pass is closely scrutinized. For example, suppose that John is older than his roommate. If a train conductor looks closely, the conductor will read John's roommate's age on the back of the pass and become suspicious that John is in fact the pass-holder. This will lead the conductor to ask for John's driver's license, which will lead to the discovery that John has borrowed his roommate's pass.

	No Close Scrutiny	Close Scrutiny
Use Roommate’s Pass	$0	-$250
Buy a Ticket	-$50	-$50

Table 1: John’s Initial Decision Problem

The decision problem that John faces is modeled in Table 1. Suppose further that John can learn whether or not the process by which some people’s train passes are subject to close scrutiny is biased in favor of whites. That is, we introduce a partition B of the set of possible worlds where $B = \{Biased, Non\text{-}Biased\}$. John believes that if the actual world is in the set of biased worlds, then the probability that his pass will not be subject to scrutiny is $.9$, whereas if the actual world is in the set of non-biased worlds, then the probability that his pass will not be subject to close scrutiny is $.5$. John assigns probability $.8$ to the proposition that the process is not biased in favor of whites, and probability $.2$ to the proposition that the process is biased in favor of whites. However, unbeknownst to John, train security is, in fact, biased in his favor in virtue of his race. This encodes the assumption that John is in an epistemic situation of white ignorance, like that described by Mills. It is entailed by these assumptions that, prior to John’s learning whether train security is biased in favor of whites, he must assign probability $.42$ to his pass being scrutinized. Under these conditions, Good’s theorem entails that John will accept information about whether or not train security is biased in favor of whites. In fact, John should pay up to $5 to learn whether or not train security is biased in this way, even though he already has a fairly high degree of belief that it is not biased. Details of the calculation by which this value is obtained are given in Appendix B. Thus, by the lights of standard expected value maximization, if John chooses to maintain his ignorance when information that would alleviate it is freely or cheaply available, then he does so under pain of irrationality.

3. Risk Aversion and Decision Theory

Risk aversion is a type of attitude that an agent can have towards a set of actions, which affects how that agent ranks those actions with respect to their choiceworthiness. To illustrate, suppose that an agent can perform one of two actions: they can gamble on the outcome of a fair coin toss, such that they are given zero dollars if the coin comes up heads, and two dollars if the coin comes up tails, or they can request one dollar and receive one dollar, regardless of the outcome of the coin toss. This decision problem is represented in Table 2. A risk-neutral agent will regard these two actions as equally choice-worthy, since both actions have the same expected payoff.[5] However, a risk averse agent will regard requesting one dollar as a more choice-worthy action than gambling. This is because the risk averse agent would rather receive one dollar as a matter of certainty than face the possibility of receiving no money, even though the possibility of receiving no money is accompanied by the equally likely possibility of receiving two dollars.

	Heads	Tails
Request $1	$1	$1
Gamble	$0	$2

Table 2: A basic decision problem.

There is substantial empirical evidence that many real-world agents are risk averse in this way; see Samuelson (1952), Allias (1979), and Kahneman, Knetsch and Thaler (1991) for especially well-known examples of this kind of behavior. However, the standard theory of how agents rank actions according to their choiceworthiness (viz., Savage's expected value theory) cannot account for agents who rank actions in a risk averse manner. Buchak's model

[5] More precisely, a risk-neutral agent with an increasing, linear value function over money will regard the two actions as equally choiceworthy.

provides a normative decision theory on which it is permissible for agents to exhibit risk aversion in their choiceworthiness ranking over actions.

On Buchak's model, risk averse agents form a choiceworthiness ranking over actions according to the *risk-weighted expected value* (REV) of each action. That is, actions are regarded as more choiceworthy to the extent that they have greater REV. The REV of an action is calculated as follows. First, for a given action a, action-state pairs are ordered $(a, s_1) \leq (a, s_2) \leq \cdots \leq (a, s_n)$, where the ordering is designed so that $V(a, s_1) \leq V(a, s_2) \leq \cdots \leq V(a, s_n)$. That is, for a given action a, action-state pairs are ordered from the state that has least value to an agent that performs a to the state that has most value to an agent that performs a. Once the states of the world are ordered in this way, the REV for the action a is defined by the equation:

$$REV(a) = V(a, s_1) + \sum_{i=2}^{n} R\left(\sum_{j=i}^{n} P\left(s_j\right)\right)\big(V(a, s_i) - V(a, s_{i-1})\big)$$

A less precise, but possibly more illuminating way of writing this equation is as follows:

$$\begin{gathered} REV(a) = V(a, s_1) + R\big(P(s_2) + P(s_3) + \cdots + P(s_n)\big)\big(V(a, s_2) - V(a, s_1)\big) \\ +R\big(P(s_3) + P(s_4) + \cdots + P(s_n)\big)\big(V(a, s_3) - V(a, s_2)\big) \\ + \cdots + R\big(P(s_n)\big)\big(V(a, s_n) - V(a, s_{n-1})\big) \end{gathered}$$

The *risk-weighting function* $R(\cdot)$ is a function from the interval $[0,1]$ into the interval $[0,1]$ such that $R(0) = 0$, $R(1) = 1$, and $R(\cdot)$ is non-decreasing. If, in addition to these requirements, $R(\cdot)$ is convex, then an agent whose choiceworthiness ranking over actions is determined by the risk-weighted expected value of those actions is said to be risk averse. Although it is tangential to our arguments here, agents with concave risk-weighting functions will exhibit a preference for risk in the way that they rank the choiceworthiness of actions.

To illustrate how this works, consider the action of gambling in the decision problem modeled in Table 2. The risk-weighted expected value of gambling is $\$0 + R(.5)(\$2)$. For any convex function $R(\cdot)$ satisfying the other constraints listed above, it is the case that $\$0 + R(.5)(\$2) < \$1$. For instance, if we define $R(\cdot)$ such that for all $x \in [0,1]$, $R(x) = x^2$, then $\$0 + R(.5)(\$2) = \$.5$. This means that an agent who determines their choiceworthiness ranking of actions according to the risk-weighted expected value of those actions, and who calculates risk-weighted expected value using a convex risk-weighting function, will prefer to request one dollar than to gamble in the decision problem modelled in Table 2. Such an agent, for reasons discussed above, is best thought of as having a choiceworthiness ranking over actions that demonstrates risk aversion. In what follows, we will show how representing risk aversion in this way allows us to model agents who deliberately avoid information about the extent of their own group-based privilege as behaving rationally.[6]

4. Risk Aversion and Rational Ignorance

We are now in a position to demonstrate how Buchak's decision theory for risk averse agents allows us to model a rational agent who avoids costless information as to their own privileged status, of the sort considered in the introduction. Importantly, we wish to stress at this stage that what we present here is a *merely possible* explanation of ignorance among elite-group members with respect to their own privilege. An argument that Buchak-style risk aversion *actually* explains the social phenomenon of elite-group ignorance would require more empirical evidence than currently exists. At the conclusion of this section, we discuss in more detail the sort of empirical evidence that would lend further support to the claim that our model actually explains elite-group ignorance, as well as the evidence that already provides support for the actual explanatory power of our model.

To illustrate how Buchak-style risk aversion *could* explain elite-group ignorance, let us reconsider the case of John, who is deciding whether to borrow his roommate's train pass. Now suppose that instead of choosing the action with maximal expected value, John ranks the choiceworthiness of actions according to their risk-weighted expected value. Suppose further that John's risk-weighting function $R(\cdot)$ is such that for all $x \in [0,1]$, $R(x) = x^2$. Note that this change in risk attitudes is *all* that changes about John's epistemic and conative states; otherwise, he is an identical agent as compared to the previous case. Under these conditions, as demonstrated by calculations given in Appendix C, John would actually pay up $6.30 to *avoid* being given information regarding bias in train security procedures. In this case, John prefers not to learn whether or not the actual world is one in which the train pass scrutiny system is biased in favor of whites. This is because John fears being in a world in which the train pass scrutiny system *is* biased in favor of whites, and yet his train pass is nevertheless subject to scrutiny. In such a world, the information that the system is biased in favor of whites will lead John to take his roommate's pass, but he will still end up paying a penalty. John can avoid this outcome by simply not learning that the world is biased in his favor, an option that he regards as choiceworthy by his risk averse lights. The idea that risk averse agents can sometimes rationally decline free information is defended in detail by Buchak (2010), Ahmed and Salow (2017), and Campbell-Moore and Salow (2020). For other examples of adjustments to the standard expected-value framework that allow for violations of Good's theorem, see Dorst (2020) and Das (forthcoming).

To illustrate this point, consider the decision tree in Figure 1. Each path represents a possible course of action that John might take. If John chooses at Γ to learn whether the ticket-

[6] Note that Thoma (2019b) raises a compelling argument against the general applicability of Buchak's framework to sequences of decisions. Thus, we reiterate that we do not wish to commit here to the claim that Buchak's theory is *the* true theory of risk aversion. Instead, we use Buchak's theory as a proof of concept, showing that decision theories that allow for risk aversion can also allow for the avoidance of free information.

scrutiny system is biased in his favor in virtue of his race, then his path to one of three possible outcomes will be determined by what he ends up learning. The worst-case scenario for John is that he moves from *Γ* to *Δ* to E and then ends up in a world where his ticket is closely scrutinized, causing him to lose $250. One way for John to avoid this worst-case scenario is to choose, at *Γ*, not to learn whether or not the ticket-scrutiny system is biased in his favor. Given the level of risk aversion represented by John's risk-weighting function, this is the action that he should take if he wishes to maximize his risk-weighted expected value.

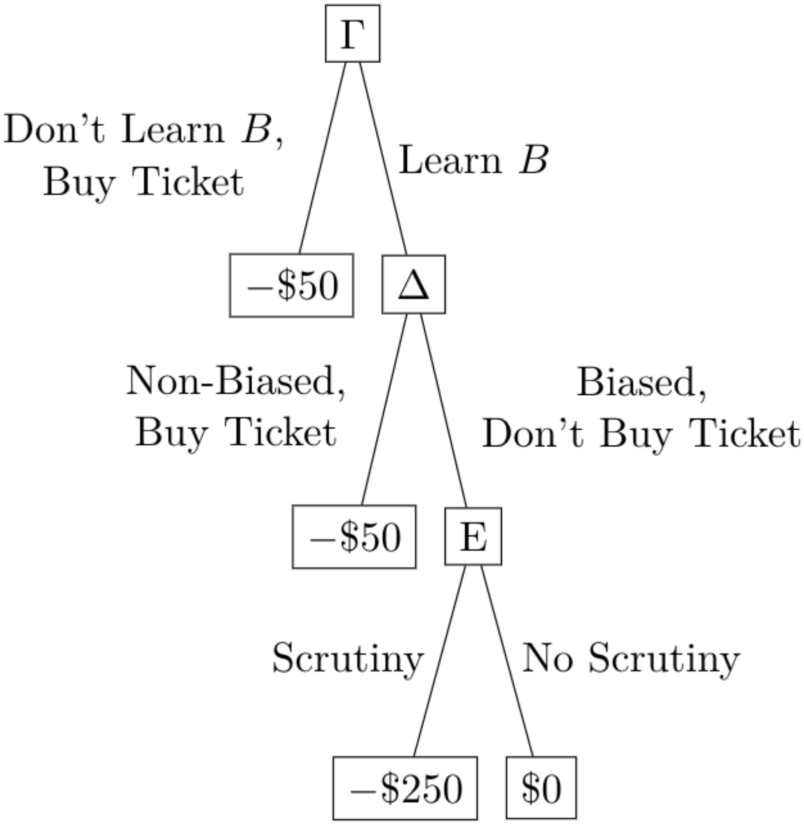

Figure 1: Tree Diagram of John's Decision

As Buchak (2010, p. 99-100) argues, these kinds of cases can be understood informally as cases in which an agent declines to perform an experiment, or otherwise learn about the world, in order to avoid receiving accurate but misleading information. In the case above, John could query whether the ticket-checking process is biased in his favor in virtue of his perceived race. However, doing so opens up the possibility of receiving accurate but misleading information. Specifically, if it turns out that, in general, the ticket-checking process *is* biased in favor of whites, but it is also the case that John's ticket *will* be checked, then, prior to John's boarding the train, the information that white people's tickets tend not to be checked carefully is misleading. As Buchak notes, said information is both instrumentally and epistemically misleading. It is instrumentally misleading in that it will lead John to pursue the suboptimal course of action of borrowing his roommate's train pass. It is also epistemically misleading in that, while it does convey to John to true state of racial bias in the ticket-checking process (namely, that it exists), it also leads John to the false belief that his ticket will not be subject to close scrutiny (when, in fact, it will). From his risk-averse perspective, John would rather avoid the risk of receiving accurate but misleading information than take this risk on in order to possibly take a free train ride. Thus, he chooses not to seek information about his race-based privilege in the ticket-checking process, and take the safer route of buying his own tickets. By the lights of Buchak's risk-weighted expected utility theory, such behavior is entirely rational.

To underscore the role that John's status as a white person plays, within the context of this model, in creating a state of affairs such that he ought to remain ignorant of his social dominance, consider what would happen if the same scenario were faced by a person of color. Call this person Kelly. Even if Kelly has the same value function over money and risk function over probabilities as John, there are conditions such that Kelly would choose to buy a train ticket whether or not the process of scrutinizing train passes is biased in favor of white people. This is because, whether or not the scrutiny process is biased in favor of whites, the probability that a person of color's train pass will be subject to close scrutiny is such that, for a person of color such as Kelly, borrowing another person's train pass has lower expected value than the cost of a train ticket. Thus, for Kelly, there is no risk of being led down a path towards a $250 penalty by learning whether or not the process is biased in favor of whites. So Kelly should, by the lights of risk-weighted expected value theory, accept free evidence about the racial biases that influence the process whereby some train passes are subject to close scrutiny.[7] A comparison of John and Kelly's attitudes towards information about the existence of structural racism establishes that, within the context of this model, John's status as a white person plays at least some role in explaining his rational ignorance of the existence of said racism. What we are saying here is thus in accord with Mills' discussion of white ignorance in which "white racial domination [...] plays a crucial causal role," provided that we accept a counterfactual account of causal explanation (2007, p. 20).

Let us use a more realistic example to further illustrate the point. Given that police can be biased in the extent to which they suspect different people of committing a crime, or what neighborhoods they consider important to patrol, a stop-and-search (or stop-and-frisk) policy induces different risks of actually getting caught for people in different demographic groups. A rational member of an elite group, e.g., a rational white middle class American teenager, might therefore be such that if they looked carefully into the data on police behavior then they would be induced to take more risks, e.g., they might habitually carry around small amounts of illegal drugs, since it is very unlikely that they will get searched and caught. If they are risk averse, with their eye on eventually going to a good college and

[7] Thus, there is a sense in which Kelly possesses what Du Bois (1903) called "double consciousness." That is, she must understand in detail how she is viewed by both white and Black people.

entering into polite society absent a criminal record, it might therefore be rational for them to avoid information about police bias in searches. Remaining ignorant in this way does not require them to bear any active ill will towards Black people; indeed, it does not require them to especially consider Black people at all. No particular attitude to Black people is required or necessarily relevant, even though the result of their rational ignorance will be a self-serving ignorance of the racist conditions Black people live with. Rather, the teenager's remaining ignorant of racial inequalities in policing only requires that the teenager be rationally self-interested and risk averse. In the same way, whether John explicitly considers the possibility of a racial bias in the ticket-inspection process is ultimately not essential to his decision to avoid information about said biases; it is enough that he is self-interested and risk averse. This likewise could fit with Mills' discussion of white ignorance that is "operative even if the cognizer in question is not racist," where it is understood that the sort of racism in question here is that of interpersonal animosity (2007, p. 21).

Indeed, according to our model, an agent can be rationally ignorant of their own privileged status *even when the agent adopts attitudes that are explicitly antithetical to unfair social structures*. For example, an actively anti-racist white person can, on our model, still rationally remain ignorant of their own white privilege. To see this, suppose that an agent actively desires an end to racial inequalities, so that they are best represented as having a valuation function over act-state pairs that does not reflect their personal self-interest in performing a certain action in a certain state, but rather the social good of performing an action in a state, where the social good consists in eliminating material racial inequalities. Such an agent has aligned their sense of what is personally good with a thoroughly anti-racist worldview. Nevertheless, our previous arguments establish that such an agent could still avoid information about their own privilege, due to risk aversion — so long as they do not believe the struggle against racial injustice will be affected by their purchase of a train ticket. This is in keeping with our comment in Section 2 that our conclusions here apply to both broadly instrumentally rational agents (such as the explicitly anti-racist agent described above) and more narrowly self-interested agents.

One might conclude from this that such an agent would be behaving both morally and rationally, but would nevertheless maintain ignorance of their own privileged status. According to this line of argument, the agent is moral because of their anti-racist valuation function, and rational because they maximize risk averse expected value. Nevertheless, they remain ignorant of their own white privilege. This suggests a stronger thesis than the one that we articulate in the introduction; per this stronger view, ignorance of one's elite-group privileges can be both rationally and morally maintained. Such a conclusion would also be a stronger departure from Mills. That is, whereas Mills views white ignorance as both immoral and irrational, one could take the line that white ignorance can be both morally and rationally maintained, because it is possible for an actively anti-racist but risk averse white agent to be rationally ignorant of their own racial advantages.

However, this is not the conclusion we wish to draw. Rather, we begin from the premise that morality will typically require agents to become aware of their own privilege. This is because of the usefulness of such awareness in effectively dismantling unfair social structures. Thus, if such an agent avoids said information due to risk aversion, then in order to act morally they must work to become less risk averse, or else change their valuation function over action-state pairs so that they do not avoid information about their relative privilege due to risk aversion. However, we recognize that this reading of the implications of our argument rests on particular moral and empirical premises (namely, that it is immoral to remain ignorant of one's privilege, especially when there are no costs to alleviating said ignorance, due to the possible downstream effects of said ignorance in perpetuating unfair inequalities) that may not be universally shared and which we cannot defend here in full. Moreover, we note that these cases will typically arise under a narrow decision model in which agents never instrumentally value information about their privilege, and in which the same agents, though valuing fair social arrangements, do not explicitly *dis*value holding false beliefs about their own privilege.[8]

This suggests a particular positioning of our argument within the three types of accounts of white ignorance highlighted in recent work by Annette Martìn (forthcoming). Martìn identifies three species of accounts of white ignorance: the willful ignorance account, the cognitivist account, and the structuralist account. According to the willful ignorance account, white agents deliberately ignore information about their own privileged position in social hierarchies. On the cognitivist account, white ignorance is due to faulty reasoning. On the structuralist account, white ignorance "systematically arises as part of some social structural process(es) that systematically gives rise to racial injustice" (Martìn forthcoming, p. 12). Though Martìn argues in favor of a structuralist account, for our part, we believe that all three accounts can explain at least some instances of white ignorance. However, our particular use of risk weighted rationality to explain elite ignorance falls somewhere between the willful ignorance and structuralist accounts. By showing that, under risk aversion, agents will avoid information about their own privilege while still maintaining rationality, we allow for rational agents to nevertheless exhibit willful ignorance of their privilege. At the same time, we are entirely sympathetic to the idea that the kinds of decision scenarios that are faced by elite agents and which encourage ignorance of their privileged status are produced by exactly the sorts of social structural processes that Martìn has in mind. What our argument here pushes back against is the idea that a cognitivist account of white ignorance can be a panacea for explaining all cases; indeed, our examples show that under a plausible model of rationality, it is *not* the case that the maintenance of white ignorance is attributable to faulty reasoning.

To summarize, Buchak's model of instrumental rationality for risk averse agents allows us to represent agents like John as behaving rationally when they deliberately avoid information that could be probative as to their own level of privilege. This entails that agents

[8] We are grateful to an anonymous reviewer for pointing this out.

whose behavior constitutes or perpetuates white ignorance can be modelled as agents who are risk averse and rational, in accordance with Buchak's decision theory. In keeping with our proviso at the beginning of this section, we stress here that we have shown only that white ignorance can be explained as the result of Buchak-style risk aversion, and not that it is in fact so explained. For this move from a *how-possibly* explanation to a *how-actually* explanation, more empirical evidence is needed.

More specifically, we identify two main categories of experiment and data collection that we believe would bolster the case for the actual explanatory power of the model presented here. First, there is a need to empirically test the extent to which actual agents are willing to decline free information in order to avoid receiving news that would license riskier decision-making. This is primarily a research program for experimental cognitive psychology and experimental philosophy. It is worth noting that there are already some promising results showing a correlation between uncertainty aversion and closed-mindedness; see for instance Jost et al. (2003) and Thórisdóttir and Jost (2011).

Second, we note that our model can only be representationally accurate if it is in fact the case that elite agents systematically *overestimate* their likelihood of facing negative consequences for certain risky behaviors. While the data clearly indicate that elite-group ignorance exists with respect to issues such as policing, it is possible that elite agents who are ignorant of their own privilege generally assume that the risks of negative interactions with police are lower for all people than they are in reality. Upon learning of a bias in their favor, said agents only need to revise upward their degree of belief that a member of a non-elite group is more likely to suffer negative consequences when taking risky actions. Such agents, no matter their level of risk aversion, will not avoid being told whether or not there is a bias in their favor, such that our explanation of elite-group ignorance will not apply. Thus, further validation for our model requires a concrete investigation of the extent to which those belonging to elite groups overestimate their own risk of negative consequences in social situations where different salient groups tend to experience highly unequal outcomes.

Regarding this second avenue for empirical research, there is already some promising evidence from empirical criminology and sociology that is suggestive of a general phenomenon in which people who have not previously engaged in risky behavior (especially criminal behavior) tend to overestimate the probability of negative consequences. Jensen (1969) finds evidence of this pattern of overestimation, as does Tittle (1980), who calls it "the shell of illusion" (p. 69). Subsequent findings that are in keeping with the shell of illusion phenomenon include Paternoster et al. (1985), Pogarsky et al. (2004, p. 349), Matsueda et al. (2006, p. 107), and Schulz (2014, p. 226). If we grant in addition the plausible assumption that members of disadvantaged groups are more likely to be placed in situations in which they are more likely to be required to engage in risky action, including criminal behavior (see Wolff and De-Shalit 2007, Chapter 3 for case studies that speak in favor of this premise), then it stands to reason that members of elite groups will be subject to the shell of illusion in their attitudes towards the risks of criminal behavior. That said, while the evidence in favor of the shell of illusion is certainly consonant with our explanatory model, stronger evidential support would be provided by studies showing that effect of the shell of illusion is more pronounced among members of elite groups. To our knowledge, no such study has been attempted, hence this is an intriguing avenue for future work.

5. Risk Attitudes and Privilege

If the possible explanation of ignorance by privileged agents that we have presented here is to be robustly representationally adequate with respect to the actual social world, then it must be the case that members of socially dominant groups typically rank the choiceworthiness of possible actions in a risk averse way. One may doubt that this condition holds. Indeed, one could argue that privileged agents typically have a *risk seeking* attitude towards possible choices, such that a member of a dominant social group may prefer a gamble with a risk-neutral expected value of $1 but an upside of $2 to receiving $1 as a matter of certainty. A real-life phenomenon that would seem to bolster this argument is the putative presence of risk seeking behavior among finance industry professionals, especially investment bankers and traders. For instance, Shefrin (2010) argues that in the lead-up to the 2008 financial crisis, executives at UBS developed an increased tolerance for risk with respect to the makeup of the firm's investment portfolio, "to the point where they became risk neutral, if not risk seeking" (p. 7). These retrospective, observational findings cohere with experimental results from Haigh and List (2005), Abdellaoui et al. (2013), and Woo and Kang (2016) showing that financial professionals exhibit more risk seeking behavior than other agents in cases where gains are unlikely and losses are probable. In most industrialized countries, if any group could be described as "socially dominant", it would be professionals with decision-making power at large financial firms. Thus, the results described above would seem to throw a wrench in our theory that ignorance of certain facts on the part of members of social dominant groups can be explained by the prevalence of risk aversion among members of those same groups.

In response, we first note that financial professionals are found to be risk-seeking in cases where losses are likely and gains are unlikely. These agents are willing to take on additional risk in order to "chase" the remote possibility of avoiding loss. Thus, these agents are perhaps better described, in the language of Bernartzi and Thaler (1995) as exhibiting "myopic loss aversion" rather than exhibiting genuine risk-seeking behavior. That is, traders and bankers regard likely losses as much worse than other agents would, such that they are willing to risk more to give themselves the possibility of making some gain. One can note that in our example above, if John learns that the ticket scrutiny process is biased in favor of white people, his subsequent decision to borrow his roommate's train pass, risking a fine, is not one in which he faces a likely loss and an unlikely gain. After all, the probability that he will lose $250 is only ten percent; the gain of a free train ride is the more likely outcome. Thus, our example is not a case in which, by the lights of empirical studies on risk-seeking behavior by agents, John is likely to be risk-seeking. In fact, the decision to learn that the scrutiny process is biased in favor of whites is closer to cases where Abdellaoui et al. (2013) have found financial professionals to be risk averse; namely, those cases in which agents face a likely gain and an unlikely loss. Indeed, the central claim of our paper is that that risk aversion can explain white ignorance in cases where knowledge of one's privilege would potentially license taking actions that have a high probability of a modest gain and a low probability of significant loss. This coheres with the predictions of Prospect Theory, as formulated by Kahneman and Tversky (1979). All this leads to the conclusion that our models do not detrimentally deviate from real-world conditions by including privileged agents who are risk averse.

These studies typically model risk aversion using Prospect Theory and standard expected utility theory, rather than Buchak-style risk weighted decision theory. As such, there is some room for disagreement regarding their probative value with respect to our arguments here. We take these studies as providing evidence of risk averse behavior among privileged agents, and note that while the authors may model said behavior using other methods, risk averse behavior *can* be modelled using Buchak's approach. Thus, we take these findings to be at least consonant with our argument about that elite-group ignorance may be attributable to risk aversion on the part of elite agents. If, on the other hand, the same studies had found that privileged agents are robustly risk-seeking, then it would be harder for us to make the case for the possible explanatory value of our model.

Secondly, putatively risk seeking behavior by financial professionals and other privileged agents may not be correctly described as risk seeking in some cases. After the immediate onset of the 2008 crisis many of the largest institutions were provided emergency liquidity by the United States Federal Reserve in order to keep these institutions from declaring bankruptcy (Bernanke 2012; Graeber 2011, pp. 15-6). This socialization of losses but not of gains suggests that an individual trader or banker's value function over outcomes might not vary linearly with gains and losses in money, but instead heavily discount the negative value of possible losses. This would suggest that putatively risk seeking financial professionals, and other privileged agents who may face less severe penalties than other agents in similar circumstances, might not be correctly modelled as having a risk seeking attitude towards possible actions.

6. Political Consequences

What is to be done about elite-group ignorance? Plenty. But here we will advance just one more specific thesis about the kinds of political interventions that can be used to combat white ignorance. Specifically, we argue that, in light of our arguments above, interventions aimed at changing people's psychological and emotional orientation towards information about their own privilege may be limited in their effectiveness. This is because, even if elite agents do not incur any emotional costs from processing such information, the conjunction of their risk attitudes and the nature of the decision problems that they face may be such that there are strong incentives to avoid such information.

To make this argument, we begin by noting that in response to all that we have said so far, one could argue that our approach to modeling the rational origins of elite-group ignorance fails to be representationally accurate on the grounds that the cost of receiving information about one's privilege, rather than one's attitude towards risk, provides a better explanation of elite-group ignorance of privilege. To illustrate, learning that one has been the beneficiary of structural inequalities in virtue of one's race or gender may be deeply unpleasant, because it brings about feelings of guilt or shame (for a philosophical discussion of this phenomenon, see Cherry 2020). Further, evidence of one's racial privilege may be highly disturbing to perceive. The video, taken in 2020 by Darnella Frazier, of George Floyd being killed by Minneapolis police officer Derek Chauvin is, for many, clear evidence of racial inequalities in the treatment of individuals by the police, and yet many who view the video are horrified simply in virtue of witnessing someone being violently killed. Scenes of poverty and desperation can also be inherently upsetting to many people, independently of whether they also constitute evidence of unfair material inequalities. Once one acknowledges these real costs to consuming information about one's privilege as a member of a systematically advantaged group, ignorance of said privileged status by members of that group can be modelled as a rational decision without appeal to risk aversion.

We respond to this line of objection by agreeing that, in many cases, agents do indeed turn down information because of the psychological or emotional costs associated with processing that information. However, our model still shows that if risk attitudes are taken to be an irreducible input to an agent's rational decision-making processes, then there are cases such that *even if* an agent manages to condition themselves so as not to incur an emotional or psychological cost to processing evidence of social inequality, that agent will still avoid information about their relative privilege within said social inequalities. Thus, if one wishes to argue that the emotional cost of receiving information about one's social privilege fully explains how ignorance on the part of members of elite groups as to their own privileged status is rationally maintained in all possible cases, then one must deny that risk attitudes can ever play an irreducible role in rational decision-making. We take this to be a significant argumentative cost, such that it is preferable to maintain our position that risk attitudes can explain, in some instances, ignorance among elites as to their own relative privilege. Moreover, recent work in experimental psychology (see Landy et al. 2018, Ivuoma et al. 2020) provides some empirical support for the claim that false beliefs about demographics and the relative privilege of various socially salient groups can be highly entrenched, even when agents are prompted to be receptive to information about their own levels of privilege.

This response allows us to clarify some of the implications for our argument with respect to strategies for alleviating elite-group ignorance. In particular, it shows that psychological interventions aimed at removing the unpleasant feelings that elite agents may feel when encountering evidence of their privilege may not be sufficient to eliminate ignorance of racial inequality. Similarly, simply encouraging agents to "push through" or learn to "sit with" the discomfort of confronting the realities of their privilege, as is advocated by consultants such as Robin DiAngelo (2018), may also not be sufficient to alleviate ignorance of said material privilege. Rather, as long as agents remain risk averse in some situations (and there may be good reasons for agents to be generally risk averse), our model is such that rational incentives to avoid evidence of one's privilege can exist. This suggests that strategies for alleviating elite group ignorance which aim primarily to intervene on the psychologies of privileged people (e.g., diversity initiatives at large corporations, or direct attempts to get people to acknowledge their privilege) may be limited in their ability to achieve their stated aims. When guiding action, one needs to take into account the potential risk aversion agents will display even after one's interventions.

To explain this more fully, there is a *prima facie* plausible, if somewhat optimistic, line of argument that runs roughly as follows: if ordinary members of elite groups simply understood the extent of the inequality that existed in society, then they would put in place measures, at both a personal and political level, to help eliminate these inequalities. Thus, the line of argument concludes, anti-inequality efforts should be focused on educating people, especially those at the top of social hierarchies, about the extent of inequality in society. Such a line of argument may often be behind, for instance, attempts to carry out climate surveys in academic departments, which would make it apparent to local elite agents just how others are experiencing departmental life. They may likewise motivate seminars or consultancies designed to let those involved in business hiring decisions know more about biases or historical obstacles to minority-group participation. Our work here shows that even if it is true that members of elite groups would do more to combat inequalities if they knew about them, it does not follow that anti-inequality efforts should be focused on educating or otherwise intervening on the psychological states of members of elite groups. This is because, according to our model, members of elite groups acting in their own rational self-interest may actively avoid information about their own privileged position in social hierarchies, such that these education efforts would be more costly than perhaps initially realized. This is all consistent with acknowledging that if one could just press a button to bring about greater knowledge on the part of hiring committees or one's academic colleagues of the difficulties faced by some groups, then of course one should press the button.

However, risk averse elite agents can be faced with the possibility that if they are more informed in this way then they might take riskier actions. For instance, they might be rationally required to give members of non-elite groups closer consideration when hiring (as argued in Bovens 2016), and thereby incur additional costs in terms of the time and effort required to find an optimal candidate. Thus, there may be rational incentives to avoid this kind of information, such that good faith efforts to educate hiring committees may be met by deliberate information-avoidance on the part of some agents. For instance, this could take the form of simply nodding and smiling one's way through educational programs but never actually thinking about what one is being told. Alternatively, these programs can be designed or selected by members of various elite groups, and may exhibit and thus perpetuate ignorance about the stark material inequalities that give rise to the very need for them. And after all, if this avoidant behavior has been engaged in, since it is validated by a plausible theory of rationality that agents themselves may (consciously or intuitively) endorse, then the agent will look back and it will seem to them that they behaved as they ought. There will thus not be a track record of decisions clearly felt to be irrational that might provide impetus for change. This conclusion echoes empirical work by Wynn (2019) on the ineffectiveness of diversity initiatives at large corporations. In light of these arguments, we may be better off aiming to eliminate social inequalities through direct interventions, rather than relying on information campaigns aimed at the presently powerful.

Note that nothing about this argument requires that agents consciously and explicitly calculate their rational self-interest in avoiding information and acting accordingly. Rather, if the models of risk averse reasoning used in this paper can in fact accurately predict behavior, then the fact that they allow for information avoidance is politically significant. An agent who has picked up the habit of checking out when presented with potentially uncomfortable information about racism or misogyny, or who just reliably fails to actually integrate what they are being told into their broader worldview, could well be instantiating the sort of rational information avoidance we are concerned with here. What we wish to stress is not the individual malfeasance of agents finding ways to ignore the information they have available, but that in so far as this behavior is rational from the point of view of risk averse agents, they will tend to find their subsequent behavior optimal from their own perspective. As such, the behavior will be in a certain sense reinforced and encouraged, and this can be expected in rational risk averse agents even if (hypothetically) they were devoid of irrational biases or raw emotional resistance to information about racism.

7. Conclusion

We conclude by indicating several avenues for future work that we believe would be fruitful, beyond the program for empirical research outlined at the conclusion of Section 4. To begin, the account of white ignorance developed by Mills is motivated by a desire to provide a needed dose of realism to the social epistemological frameworks developed by, among others, Kitcher (1994), Kornblith (1994), and Goldman (1999). That is, Mills is aiming to develop a social epistemology in which learning the truth is understood as the foremost epistemic goal of both individual and collective agents, but which also takes seriously the roadblocks that social hierarchies place in the way of achieving that goal. At the same time, Buchak's risk-weighted expected value theory is part of an instrumental rationality tradition in which formal theories of rationality aim to provide normative bounds for rational behavior, while also taking seriously real aspects of human behavior such as risk aversion. What both of these projects have in common is their desire to provide rigorous *and realistic* theoretical frameworks for understanding some set of social phenomena. Here, we have shown that combining insights from these two research programs can be fruitful. Since Mills and Buchak are not the only two philosophers whose research programs aim at both rigor and realism, we believe that it is worth exploring other ways in which research programs of this sort can be shown to shed light on real-world phenomena.

Another possible line of future research concerns the epistemological consequences of social transitions. Our model of rational agents who avoid information about their group-based privileges requires that these agents be initially uncertain about the extent of the structural advantages shared by certain social groups. In the example used throughout this paper, John must assign some non-extreme probability to the proposition that the pass-checking system on trains is biased in favor of whites, in order to generate the result that John will seek to avoid information that would settle the question of whether such bias exists. In a context in which racial and other social hierarchies are explicitly codified (e.g. the Jim Crow South or apartheid-era South Africa), agents are likely to be certain about the existence of various advantages and disadvantages possessed by people who share their socially salient characteristics. Under these conditions of explicit social hierarchy, our analysis of elite-group ignorance will not work. However, in a society in which some social hierarchies are not explicitly codified, and where one is more likely to get mixed messages about how bad things are for those at the bottom and how good things are for those at the top, agents may well be uncertain about whether some process is biased in a way that helps or harms certain groups. Thus, our model is an instance of a kind of social epistemology for a transitional world in which it is possible for a reasonable person to be unsure as to the existence of real social inequalities (perhaps on the condition that the standard of reasonableness is fairly low). There may be other instances in which social epistemologists can yield new insights by studying the unique dynamics of these sorts of transitions, and of other social transitions more broadly.

Finally, we wish to make clear where we take ourselves to sit within the broader landscape of social epistemology and standpoint epistemology. First, we note that this paper is meant to be part of an inter-disciplinary (or multi-method) approach to studying social reality. The paper begins by acknowledging the significant role that social hierarchies play in the existence of ignorance on the part of some agents, especially members of dominant social groups. This acknowledgment is rooted in historical or social reality, as recorded and studied by various modes of inquiry. We then show that this historical reality is not consonant with the standard framework for modelling the rationality of evidence-seeking. However, we go on to show that more nuanced frameworks can accommodate the historical reality of willful ignorance by members of privileged groups. This sort of approach is entirely in keeping with some accounts of standpoint epistemology. As Harding puts it, "marginalized lives provide the scientific problems, the research agendas, for standpoint theories, [...] [t]hinking from marginal lives leads one to question the adequacy of the conceptual frameworks that the natural and social sciences have designed to explain (for themselves) themselves and the world around them" (1992, p. 451). In our case, we begin with the historical and contemporary reality of elite-group ignorance, especially relevant to our focus has been white ignorance, before investigating the relationship between the fact of elite group ignorance and the theory of rational decision and risk aversion. More generally, reflecting on and refining the conceptual apparatus used to inform and formulate the claims of historicized inquiry is an important part of the collective endeavor of research.

This discussion provides an additional opportunity to clarify our aim in this paper. We do not hope to explain exactly what occurs in every instance in which an individual agent ignores the dominant status of a social group to which that agent belongs. Rather, our aim is to demonstrate that it is possible for an agent to be ignorant in this way without necessarily being irrational. Without meaning to ourselves endorse behaviors that are rational, we take "irrational" to be a thick normative term - it has both an evaluative and descriptive element. We believe that the best way to refute such evaluatively laden claims about the rationality of behavior is to show that such behavior is in fact permissible within an independently plausible normative theory of instrumental rationality. To perform its task such a theory must be sufficiently connected to everyday experience that it can plausibly describe important aspects of real decisions, but retain enough distance from actual events that it retains genuinely critical potential - i.e., the ability to normatively appraise events by comparing them to some independently plausible ideal. We use Buchak's mathematical account of risk averse rationality as such a normative theory because it allows us to precisely delineate those cases in which the sort of ignorance Mills describes is rationally permitted in just this way. On these grounds, we believe that our methodology here is fit to purpose. And if it has thus guided us to important truths about elite group ignorance, then we have good reason to change our ways. We ought to adopt direct redistributive and restructuring policies, rather than waste our time chasing the vain hope of an informed and benevolent elite.[9]

[9] We are grateful to Kevin Dorst, Tena Thau, Zina B. Ward, Xin Hui Yong, several anonymous reviewers, the LSE PhD Student work in progress seminar, and the audience of the Santa Fe Institute's Science Club for comments on earlier versions of this paper.

References

Abdellaoui, Mohammed, Han Bleichrodt, and Hilda Kammoun. “Do financial professionals behave according to prospect theory? An experimental study.” *Theory and Decision* 74, no. 3 (2013): 411–429.

Ahmed, Arif, and Bernhard Salow. “Don’t Look Now.” *The British Journal for the Philosophy of Science* 70, no. 2 (2017): 327–350.

Allais, Maurice. “The foundations of a positive theory of choice involving risk and a criticism of the postulates and axioms of the American School (1952).” In *Expected utility hypotheses and the Allais paradox*, 27–145. Springer, 1979.

Basu, Rima. "Radical moral encroachment: The moral stakes of racist beliefs." *Philosophical Issues* 29, no. 1 (2019a): 9-23.

Basu, Rima. "The wrongs of racist beliefs." *Philosophical Studies* 176, no. 9 (2019b): 2497-2515.

Benartzi, Shlomo, and Richard H Thaler. “Myopic loss aversion and the equity premium puzzle.” *The quarterly journal of Economics* 110, no. 1 (1995): 73–92.

Bernanke, Ben S. *Some Reflections on the Crisis and the Policy Response: a speech at the Russell Sage Foundation and The Century Foundation Conference on “Rethinking Finance,” New York, New York, April 13, 2012*. Technical report. 2012.

Blackwell, David. “Equivalent comparisons of experiments.” *The annals of mathematical statis-tics*, 1953, 265–272.

Bovens, Luc. “Selection Under Uncertainty: Affirmative Action at Shortlisting Stage.” *Mind* 125, no. 498 (2016): 421–437.

Bradley, Seamus, and Katie Steele. “Can free evidence be bad? Value of information for the imprecise probabilist.” *Philosophy of Science* 83, no. 1 (2016): 1–28.

Buchak, Lara. “Instrumental Rationality, Epistemic Rationality, and Evidence-Gathering.” *Philosophical Perspectives* 24, no. 1 (2010): 85–120.

Buchak, Lara. *Risk and rationality*. Oxford University Press, 2013.

Camp, Nicholas P., Rob Voigt, Dan Jurafsky, and Jennifer L. Eberhardt. "The thin blue waveform: Racial disparities in officer prosody undermine institutional trust in the police." *Journal of Personality and Social Psychology* (2021)

Campbell-Moore, Catrin, and Bernhard Salow. “Avoiding Risk and Avoiding Evidence.” *Australasian Journal of Philosophy* 0, no. 0 (2020): 1–21.

Cherry, Myisha. Solidarity Care: How to Take Care of Each Other in Times of Struggle. *Public Philosophy Journal* 3, no. 1 (2020): 1-12.

Chetty, Raj, Nathaniel Hendren, Maggie R Jones, and Sonya R Porter. *Race and economic opportunity in the United States: An intergenerational perspective*. Technical report. National Bureau of Economic Research, 2018.

Das, Nilanjan. “The Value of Biased Information.” *The British Journal for the Philosophy of Science*, forthcoming.

DiAngelo, Robin. *White fragility: Why it’s so hard for white people to talk about racism*. Beacon Press, 2018.

Dorst, Kevin. “Evidence: A guide for the uncertain.” *Philosophy and Phenomenological Research* 100, no. 3 (2020): 586–632.

Dotson, Kristie. “Tracking epistemic violence, tracking practices of silencing.” *Hypatia* 26, no.2 (2011): 236–257.

Du Bois, William Edward Burghardt. *The souls of black folk*. A.C. McClurg & Co., 1903.

Ellison, Ralph. *Invisible man*. Random House, 1952.

Gärdenfors, Peter, and Nils-Eric Sahlin. “Decision making with unreliable probabilities.” *British Journal of Mathematical and Statistical Psychology* 36, no. 2 (1983): 240–251.

Goldman, Alvin I. *Knowledge in a social world*. Oxford University Press, 1999.

Good, Irving John. “On the principle of total evidence.” *The British Journal for the Philosophy of Science* 17, no. 4 (1967): 319–321.

Graeber, David. "Debt: The first five thousand years." *New York: Melville House* (2011).

Haigh, Michael S, and John A List. “Do professional traders exhibit myopic loss aversion? An experimental analysis.” *The Journal of Finance* 60, no. 1 (2005): 523–534.

Harding, Sandra. “Rethinking standpoint epistemology: What is” strong objectivity?”” *The Centennial Review* 36, no. 3 (1992): 437–470.

Jensen, Gary F. "“Crime doesn't pay”: Correlates of a shared misunderstanding." *Social Problems* 17, no. 2 (1969): 189-201.

Jost, John T, Jack Glaser, Arie W Kruglanski, and Frank J Sulloway. “Political conservatism as motivated social cognition.” *Psychological bulletin* 129, no. 3 (2003): 339.

Kahneman, Daniel, Jack L Knetsch, and Richard H Thaler. “Anomalies: The endowment effect, loss aversion, and status quo bias.” *Journal of Economic perspectives* 5, no. 1 (1991): 193–206.

Kahneman, Daniel, and Amos Tversky. “Prospect Theory: An Analysis of Decision under Risk.” *Econometrica* 47, no. 2 (March 1979): 263–291.

Kitcher, Philip. “Contrasting conceptions of social epistemology.” In *Socializing epistemology: The social dimensions of knowledge*, 111–134. Rowman / Littlefield: Lanham, MD, 1994.

Kolmogorov, Andrey Nikolaevich. *Foundations of probability*. Ergebnisse Der Mathematik, 1933.

Kornblith, Hilary. “Contrasting conceptions of social epistemology.” In *Socializing epistemology: The social dimensions of knowledge*, 93–110. Rowman / Littlefield: Lanham, MD, 1994.

Kraus, Michael W, Ivuoma N Onyeador, Natalie M Daumeyer, Julian M Rucker, and Jennifer A Richeson. “The misperception of racial economic inequality.” *Perspectives on Psychological Science* 14, no. 6 (2019): 899–921.

Landy, David, Brian Guay, and Tyler Marghetis. “Bias and ignorance in demographic perception.” *Psychonomic bulletin & review* 25, no. 5 (2018): 1606–1618.

Lum, Kristian, and William Isaac. “To Predict and Serve?” *Significance* 13, no. 5 (2016): 14–19.

Martín, Annette. “What is White Ignorance?” *The Philosophical Quarterly*, forthcoming.

Matsueda, Ross L., Derek A. Kreager, and David Huizinga. "Deterring delinquents: A rational choice model of theft and violence." *American sociological review* 71, no. 1 (2006): 95-122.

Mills, Charles. “White ignorance.” In *Race and Epistemologies of Ignorance*, edited by Shannon Sullivan and Nancy Tuana, 26–31. State University of New York Press Albany, 2007.

Mills, Charles. *Black Rights/White Wrongs: The Critique of Racial Liberalism*. Oxford University Press. 2017.

O’Connor, Cailin. “Modeling Minimal Conditions for Inequity.” *Unpublished Manuscript*, 2017. http://philsci-archive.pitt.edu/13473/.

Onyeador, Ivuoma N, Natalie M Daumeyer, Julian M Rucker, Ajua Duker, Michael W Kraus, and Jennifer A Richeson. “Disrupting beliefs in racial progress: Reminders of persistent racism alter perceptions of past, but not current, racial economic equality.” *Personality and Social Psychology Bulletin*, 2020, 0146167220942625.

Paternoster, Raymond, Linda E. Saltzman, Gordon P. Waldo, and Theodore G. Chiricos. "Assessments of risk and behavioral experience: An exploratory study of change." *Criminology* 23, no. 3 (1985): 417-436.

Pew Research Center. *On Views of Race and Inequality, Blacks and Whites Are Worlds Apart.* Technical report. 2016.

Pogarsky, Greg, Alex R. Piquero, and Ray Paternoster. "Modeling change in perceptions about sanction threats: The neglected linkage in deterrence theory." *Journal of Quantitative criminology 20,* no. 4 (2004): 343-369.

Resnik, Michael D. *Choices: An introduction to decision theory*. U of Minnesota Press, 1987.

Ross, Cody T. “A Multi-Level Bayesian Analysis of Racial Bias in Police Shootings at the County-Level in the United States, 2011–2014.” *PLoS one* 10, no. 11 (2015): e0141854.

Samuelson, Paul. “Risk and Uncertainty: A Fallacy of Large Numbers, Reprint.” *Scientia*, 1952.

Savage, Leonard J. *The foundations of statistics*. Courier Corporation, 1972.

Schulz, Sonja. "Individual differences in the deterrence process: Which individuals learn (most) from their offending experiences?." *Journal of Quantitative Criminology* 30, no. 2 (2014): 215-236.

Seidenfeld, Teddy. “A contrast between two decision rules for use with (convex) sets of proba-bilities: Γ-maximin versus E-admissibility.” *Synthese* 140, nos. 1/2 (2004): 69–88.

Shefrin, Hersh. “How psychological pitfalls generated the global financial crisis.” *Voices of Wisdom: Understanding the Global Financial Crisis, Laurence B. Siegel, ed., Research Foundation of CFA Institute*, 2010, 10–04.

Stanley, Jason. *How propaganda works*. Princeton University Press, 2015.

Tittle, Charles R. *Sanctions and social deviance: The question of deterrence.* Praeger, 1980.

Thoma, Johanna. "Decision Theory." In *The Open Handbook of Formal Epistemology*. Edited by Richard Pettigrew & Jonathan Weisberg, 57-106, PhilPapers Foundation, 2019a.

Thoma, Johanna. "Risk aversion and the long run." *Ethics* 129, no. 2 (2019b): 230–253.

Thórisdóttir, Hulda, and John T Jost. "Motivated closed-mindedness mediates the effect of threat on political conservatism." *Political Psychology* 32, no. 5 (2011): 785–811.

Tonneson, Stephanie. "Has corporate America reached a diversity tipping point?" ZoomInfo. June 23, 2020. https://zoominfo.medium.com/has-corporate-america-reached-a-diversity-tipping-point-fabe8ff6f07c.

Wolff, Jonathan, and Avner De-Shalit. *Disadvantage*. Oxford university press on demand, 2007.

Woo, Jung Seok, and Hyung-Goo Kang. "Risk attitudes of investment bankers: are they risk-lovers? Experiment and survey on investment bankers." In *Conference of the Korean Financial Association, Bangkok*, 705–773. 2016.

Wynn, Alison T. "Pathways toward Change: Ideologies and Gender Equality in a Silicon Valley Technology Company." *Gender & Society*, 2019, 1–25.

Appendix A.

We produce here not a reconstruction of Good's theorem itself, but rather a detailed explanation of how the value of information is calculated in a standard expected-value framework, and Good's key result that value of information is always positive for free information. The decision to seek free information is modelled as follows. Assume that an agent that deliberates between a set of actions A and has a value function $V(\cdot)$ over the Cartesian product $A \times S$, where S is a set of states of the world that partitions the set of possible worlds. Now, we introduce a second partition X over the set of possible worlds, and define a joint probability distribution $P(\cdot)$ over an algebra on the Cartesian product $X \times S$. Under the assumption that both S and X are of finite cardinality, this allows us to calculate the conditional probability that the actual world is in state $s_i \in S$, given that it is in state $x_k \in X$, according to the following "ratio formula":

$$P(s_i|x_k) = \frac{P(s_i, x_k)}{P(s_i)}$$

The ratio formula allows us to calculate the conditional expected value of performing a given action $a \in A$, conditional on the actual world being in state x_k, according to the following equation:

$$CEV(a|x_k) = \sum_{i=1}^{n} P\,(s_i|x_k)V(a, s_i)$$

Going forward, we assume that agents who generally rank the choiceworthiness of actions according to their expected value also rank the choiceworthiness of actions, conditional on the world being in some state x_k, according to their conditional expected value.

The action that maximizes expected value, conditional on the world being in some state x_k, is denoted mathematically as $argmax_{a \in A} CEV(a|x_k)$. By contrast, the action that maximizes unconditional expected value is denoted as $argmax_{a \in A} EV(a)$. Thus, if an agent that maximizes expected value learns that the actual world is in x_k, then they will choose to perform the action $argmax_{a \in A} CEV(a|x_k) \in A$. If the agent does not learn which element of X the actual world is in, then the agent will choose to perform the action $argmax_{a \in A} EV(a) \in A$. This notation allows us to calculate, in a general way, the expected value of choosing to learn which element of X the actual world is in. Recall that the agent's value function $V(\cdot)$ is defined over the product space $A \times S$. Recall further that if the agent learns that the actual world is in x_k, then they will perform $argmax_{a \in A} CEV(a|x_k)$. Thus, if the the actual world is in x_k and s_i and an agent learns that the actual world is in x_k, then that agent will perform $argmax_{a \in A} CEV(a|x_k)$ and receive the payoff $V(argmax_{a \in A} CEV(a|x_k), s_i)$. However, if the actual world is in x_k and s_i and the agent chooses not learn the which element of X the actual world is in, then the agent will perform $argmax_{a \in A} EV(a)$, and receive the payoff $V(argmax_{a \in A} EV(a), s_i)$. Thus, the decision problem of choosing to learn or not learn which element of X the actual world is in, supposing that X has m elements, is represented in Table 3.

	$x_1 \cap s_1$	$x_2 \cap s_1$	. . .	$x_m \cap s_n$
Learn X	V(argmax$_{a \in A}$$CEV$ $(a\|x_1)$, s_1)	V(argmax$_{a \in A}$$CEV$$(a\|x_2)$, s_1)	. . .	V(argmax$_{a \in A}$$CEV$$(a|x_m)$, s_n)
Don't Learn X	V (argmax$_{a \in A}$$EV$ (a), s_1)	V(argmax$_{a \in A}$$EV$ (a), s_1)	. . .	V(argmax$_{a \in A}$$CEV$ (a), s_n)

Table 3: Decision table showing the value an agent receives by learning or not learning which element of X the actual world is in, when the actual world is in each possible union of elements of X and S. As an example of an elided column of this table, if the actual world is in the set $x_3 \cap s_4$, then the value of learning which element of X the actual world is in is $V(argmax_{a \in A} CEV(a|x_3), s_4)$, and the value of not learning this information is $V(argmax_{a \in A} EV(a), s_4)$.

We can use this table to calculate the expected value of learning which element of X the actual world is in, by generalizing to a case in which the probability function is defined over the product space $X \times S$. This yields the following:

$$EV(LearnX) = \sum_{k=1}^{m} \sum_{i=1}^{n} P\,(x_k, s_i) V(argmax_{a \in A} CEV(a|x_k), s_i)$$

$$EV(Don't\,LearnX) = \sum_{k=1}^{m} \sum_{i=1}^{n} P(x_k, s_i) V(argmax_{a \in A} EV(a), s_i)$$

$$= \sum_{i=1}^{n} P(s_i) V(argmax_{a \in A} EV(a), s_i)$$

The *value of information* about which element of X the actual world is in, which we denote as $VOI(X)$, is defined as follows:

$$VOI(X) = EV(LearnX) - EV(Don't\,LearnX)$$

That is, $VOI(X)$ is the difference between the expected value of learning which element of X the actual world is in and the expected value of acting without learning this information. Good (1967) shows that $VOI(X)$ cannot be negative. Thus, an agent who ranks the choiceworthiness of actions according to expected value theory will never regard learning which element of X the actual world is in as more choiceworthy than not learning this information. If this agent is rational, then they will always accept free evidence about which element of X includes the actual world.

Appendix B.

Prior to learning about the existence of racial bias in train security, John's expected value for using his roommate's pass is $.52(\$0) + .42(-\$250) = -\$105$. Thus, in the absence of information, he will buy a ticket, incurring a cost of \$50. The conditional expected value of each action when the actual world is biased or not biased is as follows:

$$CEV(\textit{Use Roomate's Pass}|\textit{Biased}) = .9(\$0) + .1(-\$250) = -\$25$$

$$CEV(\textit{Buy a Ticket}|\textit{Biased}) = -\$50$$

$$CEV(\textit{Use Roomate's Pass}|\textit{Non-Biased}) = .5(\$0) + .5(-\$250) = -\$125$$

$$CEV(\textit{Buy a Ticket}|\textit{Non-Biased}) = -\$50$$

Thus, if John learns that the actual world is biased in favor of whites, then he will use his roommate's train pass. If he learns that the world is not biased in favor of whites, then he will buy a ticket. In the mathematical notation introduced above, $argmax_{a \in A} CEV(a|\textit{Biased}) = \textit{Use Roommate's Pass}$ and $argmax_{a \in A} CEV(a|\textit{Non-Biased}) = \textit{Buy Ticket}$.

	Biased, No Scrutiny	Not Biased, No Scrutiny	Biased, No Scrutiny	Not Biased, Scrutiny
Learn *B*	\$0	–\$50	–\$250	–\$50
Don't Learn *B*	–\$50	–\$50	–\$50	–\$50

Table 4: Decision table showing the value to John of learning or not learning whether the train pass scrutiny process is biased in favor of whites.

All of this yields the decision table shown in Table 4 for John's decision to learn whether or not the world is biased in favor of whites with respect to whether train passes will be inspected with close scrutiny. John initially believes that the actual world is equally likely to be biased or non-biased. This allows us to calculate the expected value of learning whether or not the train pass scrutiny process is biased in favor of whites as follows:

$$EV(\mathit{LearnB}) = P(\mathit{Biased, No\ Scrutiny})(\$0) + P(\mathit{Not\ Biased, No\ Scrutiny})(-\$50)$$
$$+P(\mathit{Biased, Scrutiny})(-\$250) + P(\mathit{Not\ Biased, Scrutiny})(-\$50)$$

$$EV(\mathit{LearnB}) = P(\mathit{No\ Scrutiny}|\mathit{Biased})P(\mathit{Biased})(\$0)$$
$$+P(\mathit{No\ Scrutiny}|\mathit{Not\ Biased})P(\mathit{Not\ Biased})(-\$50)$$
$$+P(\mathit{Scrutiny}|\mathit{Biased})P(\mathit{Biased})(-\$250)$$
$$+P(\mathit{Scrutiny}|\mathit{Not\ Biased})P(\mathit{Not\ Biased})(-\$50)$$

$$EV(\mathit{LearnB}) = .9(.2)(\$0) + .5(.8)(-\$50) + .1(.2)(-\$250) + .5(.8)(-\$50) = -\$45$$

Given that the expected value of not learning whether the train pass scrutiny process is biased in favor of whites is $-\$50$, John should pay up to \$5 to learn this information. Good's result shows that choosing to learn which element of a given partition the actual world is in will always have non-negative value by the lights of Savage's decision theory. Thus, if the only rational choiceworthiness ranking over actions tracks the expected value of those actions, then members of elite social groups who deliberately ignore evidence of their own group-based privileges act irrationality in this case.

Appendix C.

Prior to learning about the existence of racial bias in train security, the risk-weighted expected value for John of borrowing his roommate's train pass is $-\$250 +.52^2(\$250) = -\$182.4$. Thus, in the absence of information, he will buy a ticket, incurring a cost of \$50. The risk-weighted conditional expected value (RCEV) of each the two actions that John deliberates between, given his learning that the ticket scrutiny process is or is not biased in favor of whites, is given by each of the following equations:

$$RCEV(\mathit{Use\ Roomate's\ Pass}|\mathit{Biased}) = -\$250 +.9^2(\$0 - -\$250) = -\$47.5$$

$$RCEV(\mathit{Buy\ a\ Ticket}|\mathit{Biased}) = -\$50$$

$$RCEV(\mathit{Use\ Roomate's\ Pass}|\mathit{Non\text{-}Biased}) = -\$250 +.5^2(\$0 - -\$250) = -\$187.5$$

$$RCEV(\mathit{Buy\ a\ Ticket}|\mathit{Non\text{-}Biased}) = -\$50$$

This implies that the following equations hold:

$$\mathit{argmax}_{a \in A} RCEV(a|\mathit{Biased}) = \mathit{Use\ Roommate's\ Pass}$$

$$argmax_{a \in A} RCEV(a|Non\text{-}Biased) = Buy\ Ticket$$

Thus, under risk aversion, John's decision problem about whether to investigate bias in train pass inspection is again represented by Table 4. However, John's risk-weighted expected value of learning whether or not there is bias in train pass inspection differs from his risk-neutral expected value of learning this same information. This is demonstrated by the following calculation:

$$\begin{gathered} REV(LearnB) = -\$250 + R(P(No\ Scrutiny|Biased)P(Biased) \\ +P(No\ Scrutiny|Not\ Biased)P(Not\ Biased) \\ +P(Scrutiny|Not\ Biased)P(Not\ Biased))(-\$50 - -\$250) \\ +R(P(No\ Scrutiny|Biased)P(Biased) \\ +P(No\ Scrutiny|Not\ Biased)P(Not\ Biased))(-\$50 - -\$50) \\ +R\big(P(No\ Scrutiny|Biased)P(Biased)\big)(0 - -\$50) \end{gathered}$$

$$\begin{gathered} REV(LearnB) = -\$250 + (.9(.2) + .5(.8) + .5(.8))^2(-\$50 - -\$250) \\ +(.9(.2) + .5(.8))^2(-\$50 - -\$50) + (.9(.2))^2(0 - -\$50) = -\$56.30 \end{gathered}$$

Since the risk-weighted expected value of not learning B is $-\$50$, under risk aversion, John would pay up to \$6.30 to avoid information about bias in the train security process.