Risk, Rationality and (Information) Resistance: De-rationalizing Elite-group Ignorance

Yong Xin Hui

Forthcoming in *Erkenntnis*

Abstract

There has been a movement aiming to teach agents about their privilege by making the information about their privilege as costless as possible. However, some argue that in risk-sensitive frameworks, such as Lara Buchak's (2013), it can be rational for privileged agents to shield themselves from learning about their privilege, even if the information is costless and relevant. This threatens the efficacy of these information-access efforts in alleviating the problem of elite-group ignorance.

In response, I show that even within the same framework, in this case David Kinney and Liam Kofi Bright's (2021), the rationality of this information avoidance rests on shaky ground in practice. In this framework, whether an agent should avoid information depends on the precise details of (1) how relevant they expect the information to be, (2) their priors about the value of various options, and (3) their risk attitudes. The model suggests that rationality of elite-group ignorance is a function of structural factors that are pervasive but nonetheless not insurmountable, thus offering a way out of pessimism about elite group education.

1 Introduction

In highly unequal societies with clear demarcations between those with higher privilege and power and those without, agents at the top of the social hierarchy (hereafter, "elite-group

agents") are often ignorant of the extent or even of the existence of their privilege, as well as of the struggles of those at the bottom.

This ignorance (which we shall call "elite-group ignorance") is not a benign one; it has dangerous societal consequences. Decision-makers, who are disproportionately from privileged backgrounds, might then lack important information about the problems that non-elite-group agents face, and might not understand and thereby be motivated to solve these issues. Even worse, even if we are careful about representation among policy-makers, mass voter ignorance can threaten the accountability of governing institutions and leave decision-makers open to political capture (Guerrero 2021), where decisions are made for the interests of the powerful instead of the rest of the population. Therefore, elite-group ignorance is inextricably tied to stable structures of inequity, such as white supremacy and patriarchy, and has great practical and moral import beyond ordinary ignorance because it "systemically [emerges] from our social practices and [is] importantly related to the persistence of [inequality]" (Martín 2021).

Perhaps especially saliently in recent years, there have been various efforts to increase access and decrease cost to elite-group agents to learn about their privilege. From DEI panels to Instagram infographic slides, the hope seems to be that by reducing the cost of learning, the elite-group agents will now be more likely to learn information that could alleviate their ignorance. However, empirical work suggests that DEI efforts that increase access to relevant information are often futile.[1]

In other words, what if these efforts have failed because the problem isn't cost, and even making information more readily accessible will do little to alleviate the problem of elite ignorance? What if it's because avoiding costless and relevant information about one's privilege could be rational and thus less likely to be swayed in response to criticisms of irrationality?

For example, David Kinney and Liam Kofi Bright (2021) argue that within some decision frameworks plausibly regarded as frameworks of normative rationality (they use L. M. Buchak (2013) as an exemplar[2]) , rational agents can permissibly avoid costless relevant information,

1. e.g. Duguid and Thomas-Hunt 2015; Wynn 2020, where debiasing efforts seem to have limited effectiveness in changing behaviour.

2. For the purposes of the upshots that Kinney and Bright are concerned with, it is significant that normative decision-theoretic models are predictable and stable insofar as the awareness that one is operating with the model, instead of motivating a revision of one's pattern of reasoning, reinforces one's belief that their model is rational. I remain agnostic on whether Buchak's framework, or for that matter any risk-sensitive expected utility framework, is normatively rational in any sense of normative. All I require from rationality is that a

and thus do not exhibit willful ignorance in a way that is rationally criticizable. Kinney and Bright worry that risk-averse frameworks can serve as a realistic and rigorous representation of how rational agents, including elite-group agents, make decisions in an instrumentally rational manner, *even* if these agents have morally-aligned values. If Buchak is right, the argument goes, and there are indeed agents that exhibit Buchakian behaviour, these agents may sometimes coherently and rationally choose to avoid costless and relevant information. If so, shoving low-cost or even costless information in the agents' faces may not prove sufficient for them to learn.[3] If this behaviour is rational, it explains why the agent may not, *ceteris paribus*, have instrumental reason to change their information-avoiding behavior, even if they are aware that they are shielding themselves from learning. This behavior can be thus "reinforced and encouraged,"[4] conferring stability to the above-mentioned harms. In other words, we are caught in an "epistemic trap" of harmful behaviour and beliefs that cannot be dislodged by just providing costless information alone or pointing out flaws in reasoning, with inquiry "pulling you more tightly into the trap"(Nguyen 2022). Thus they argue interventions focusing on "changing people's psychological and emotional orientation towards information about their own privilege" are limited, and therefore that hope of a "informed and benevolent elite" and mitigation of harm is vain.

I argue that these conclusions are overly pessimistic by showing that Buchakian agents only avoid information in highly circumscribed settings. In particular, deviations on any of a number of parameters of the decision problem will lead Buchakian agents from *avoiding* information to being willing to *pay* for it. Variations in 1) How decision-relevant the information is, 2) the agent's priors about the information they accept, or 3) the agent's risk attitudes can all lead Buchakian agents to seek out information. Like a graduate student doing a literature review right before their deadline, the agent prefers to seek out as much (currently) irrelevant information as possible just in case they ever encounter a situation where the certain view acquired by learning this information will come in handy. Therefore, even the most Buchakian of agents might not be rationally permitted to avoid learning about their own privilege, and I

rational agent is robustly resilient against intervention.

3. For this paper, I take 'learning' to be modeled as when an agent lets the information affect how they take further actions. This means that even if an agent encounters information, they may choose not to update on it and therefore 'not learn' on my account.

4. Kinney and Bright 2021, pp.20.

destabilize the pull of this epistemic trap.

If my findings hold water, the rationality of elite-group ignorance is then generally a function of structural factors that are **pervasive but nonetheless not insurmountable**. As persistent as rational information resistance currently could be, I argue that there are cracks in the fortress of elite-group ignorance that we can focus our destabilizing efforts on, such that our information arrows can penetrate their targets.

In §2 , I outline Buchak's model of rational risk-averse reasoning to show how a rational elite-group agent can choose to avoid relevant costless information about their own privilege. §3 shows how the agent's rational standing in their information avoidance is due to myriad factors, such as the perceived relevance of the information, the agent's priors, and their risk attitudes. I then explore the political upshots of my findings in §4 before concluding in §5.

2 Setting the Scene: Buchakian Rational Information Resistance

Kinney and Bright reject the assertion that motivated elite-group ignorance is necessarily irrational.[5] Elite agents, even with their privileged access to educational resources, might still rationally choose to reject costless and relevant information about their privilege. Kinney and Bright frame their argument using Buchak's risk-sensitive model of instrumental rationality via an example where it is rational to reject costless information.

2.1 Good's Theorem and *Harassment*

The value of information, according to Good (1967), is always non-negative. Those who do not conditionalize beliefs on costless and relevant information do so under pain of irrationality if they will always choose the act with the highest expected utility. Good's Theorem reflects the intuitive notion that learning more information should make the agent more sensitive to different factors, which should allow for more fine-tuned decision-making options.[6]

5. This in partial response to how Mills (2007) has been interpreted by some as presenting an argument that white ignorance, as an example of elite-group ignorance, is motivated and irrational.

6. Good's Theorem is a technical result; I offer the intuitive gloss because it is all that is required for this project. If agents rank the choiceworthiness of their actions according to their expected value, and they also

Why then do agents remain unaware of their privilege even with relevant and costless information? Kinney and Bright consider a choice situation involving an elite-group agent with a choice between a gamble and a sure result, as well as a decision to learn about their privileged position before they make that choice. I agree with the authors that such cases are common. Here's a schematic version of one that illustrates why:

> **Harassment:** John, an elite-group agent (he's a white man), is taking the train. He sees that a passenger, Stacey, is being called racial slurs by another passenger, Al. If he leaves the scene, there will be trouble; Stacey's harm results in a utility of -50 (John is sensitive to the social good and includes it in his choiceworthiness calculations). Staying and intervening will be a gamble; either he mitigates the harm (Utility 0) or Al turns on him too and the situation more than doubles (Utility -250). He faces a decision as presented in Fig. 1: should he stay or should he go?[7]

The expected utility of intervention depends on the likelihood of escalation, which is then a function of whether John is in a biased world. If John is in a biased world (i.e. B), Al is biased towards white people and thus respects him more; the probability of escalation is 0.1. Otherwise, Al would attack him regardless of his race (call this case 'in an unbiased world'), and the probability of escalation is 0.5. His current credence that Al's animosity is equal opportunity (i.e. that ¬B) is 0.8. Therefore, his current expected utility of intervening is $EU(intervene) = ((0.8 \times 0.5) + (0.2 \times 0.1)) \times -250 = 0.42 \times -250 = -105$, less than the flat -50 utility of leaving. Therefore $leave \succ intervene$; he prefers to leave. Because John currently has a 0.42 credence that he'll be attacked, he would leave if his goal is to maximize expected utility; his expected utility, or EU(base), is thus -50.

However, what if, as illustrated in Fig. 2, he could get access to information and thus know whether he was in a biased world (i.e. 'Whether B'), *and then* make his decision to intervene based on that? John currently sees a 20% chance that B (with a 0.1 chance of escalation), and 80% chance that ¬B (with equal odds of escalation). Now suppose John learned with certainty

rank the choiceworthiness of actions conditional upon a partitioned state according to their expected value conditioned upon the world being in each state, then Good's Theorem shows that the difference between the expected utility of the preferred action based on knowing which finely-grained partitions the agent is in (i.e. the expected utility of learning) and the EU of the preferred action without knowledge about learning about the partitioning (i.e. the expected utility if the agent did not learn) is non-negative.

7. Inspired by The Clash (1982).

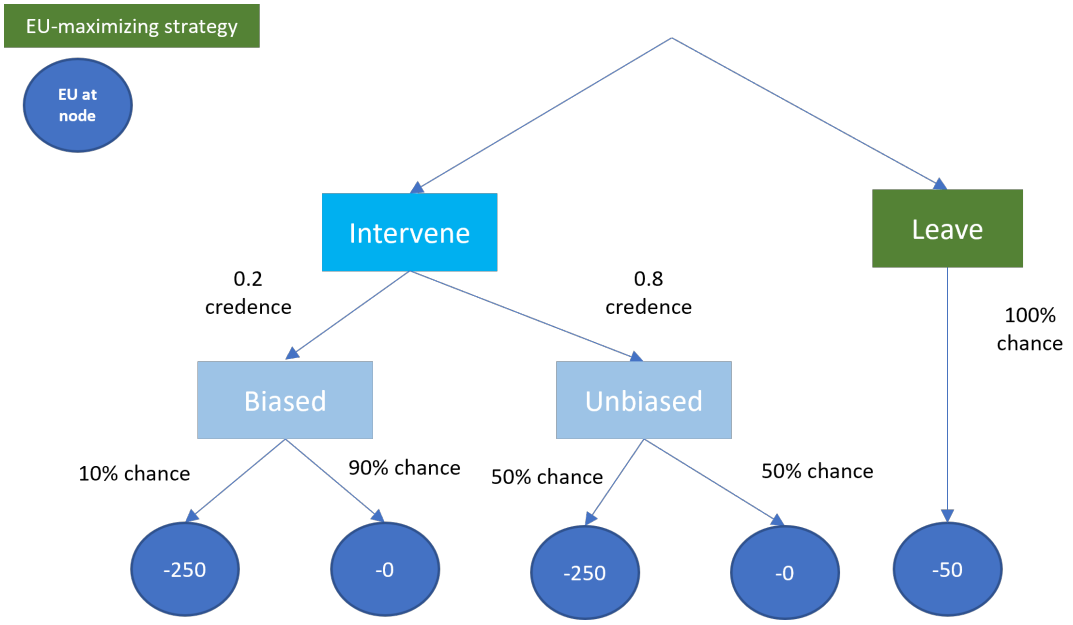

Figure 1: An illustration of John's current choices and the expected utility of each end-node. John prefers to leave at this point.

whether B. If B, the expected utility of intervening is $(0.1 \times (-250)) + (0.9 \times 0) = (-25)$, more than the expected utility of leaving. The choiceworthy move would be to intervene. If not B, he would stick with leaving; $EU(leave) = (-50) > (-125) = 0.5 \times (-250) = EU(intervene)$. The expected utility of learning is now $(0.8 \times (-50)) + (0.2 \times ((0.1 \times (-250) + 0.9 \times (-0))) = (-45)$.

So far, so Good: expected utility theory says that the value of information is $(-45) - (-50) = 5$, a positive number. This is an instance of Good's Theorem, which applies generally. So if we see that John chooses *not* to learn whether B, can we infer that he's irrational? Not so fast. Enter risk-weighted expected utility (REU).

2.2 Reconceptualizing EUT

Before going into the Buchakian framework, let us first reconceptualize the expected utility in terms of differential utility levels and probabilities[8]. Let's calculate the expected utility of John's learning again. The base utility level corresponds to the utility of the worst possible case, which arises if John learns that B, intervenes, but gets attacked. That's a utility of -250.

8. This is only an algebraic manipulation of the probability calculus; the total expected utility will be the same. See L. Buchak (2017) for a graphic explanation.

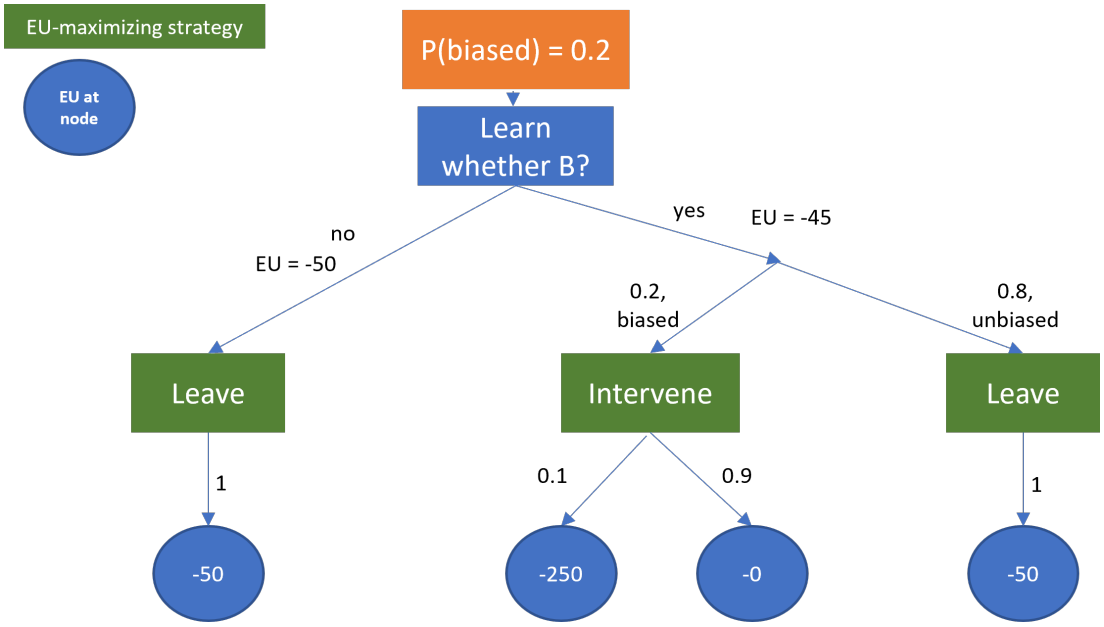

Figure 2: An update on Fig. 1, given that John can now choose to learn whether B.

The next worst possible utility level corresponds to the scenario when not B and he leaves, with a differential utility from the worst utility of $-50-(-250)=200$ and a probability of $0.8+0.2\times0.9=0.98$ that the utility is at least -50 (i.e. either where John leaves or he intervenes successfully). The best possible utility level corresponds to when B is true, he intervenes, and mitigates the harm; the utility differential is $0-(-50)=50$ and the probability of that is $0.2\times0.9=0.18$. So the expected utility using a "wedding cake"-like formula is

$$\begin{aligned} EU(learning) &= U_1+((P_2+P_3)\times(U_2-U_1))+(P_3\times(U_3-U_2)) \\ &= -250+(0.98\times200)+(0.18\times50)=-45 \end{aligned}$$

where the utilities are in ascending order: U_3 is the best case scenario utility with a corresponding probability of P_3.

In Buchak's model, the agent also factors in risk into their choiceworthiness calculations, where agents instead maximize their REU, incorporating both their expected utility and a risk function $R(x)$. She models risk-sensitive agents by scaling up or down the expected utility of an outcome, such that riskier utility-states are less or more valuable to the agent respectively.

Going back to the "wedding cake" formulation: assuming John was risk-averse, we add a risk function to each probability such that we get our "compressing" effect. Let's take $R(x) = x^2$, following Kinney and Bright.[9] Now the formula is:

$$\begin{aligned} REU(learning) &= U_1 + (R(P_2 + P_3) \times (U_2 - U_1)) + (R(P_3) \times (U_3 - U_2)) \\ &= -250 + (0.98^2 \times 200) + (0.18^2 \times 50) = -56.3 \end{aligned}$$

which $< REU(base)$

$$\begin{aligned} &= Max(REU(leave), REU(intervene)) \\ = Max((-50), ((-250) + (0.8 \times 0.5) + ((0.2 \times 0.9)^2 \times 200) + ((0.2 \times 0.9)^2 \times 50))) \\ &= Max((-50), (-181.1)) = (-50) \end{aligned}$$

The preferred strategy[10] is now to avoid learning whether B, despite its costlessness, because $(-56) < (-50)$.[11] If John was an elite-group agent with a consistent and precise reasoning process, and subscribed to the Buchakian decision framework with a risk function of $R(x) = x^2$, it would be coherently instrumentally rationally permissible given what he believes to avoid learning whether B. He is thus not compelled to amend his decision model as it serves his purposes as he sees them. Generalizing, John could rationally be permitted to avoid information, including information about his own privilege, that enables him to make riskier decisions.

Kinney and Bright contend that agents like John "whose behavior constitutes or perpetuates white ignorance [may possibly be] modelled as agents who are risk averse and rational." Their reasoning process is consistent and stable even upon "vigorous intellectual effort" (Nguyen 2022), and because they are elite-groupa agents, their credences are a "[systemic overestimation of] their likelihood of facing negative consequences for certain risky behaviors" (Kinney and Bright 2021). However, given that John's reasoning process may be robust against criticisms (either externally or through internal reflection) of irrationality given that he is a Buchakian

9. This risk function would represent risk-aversion.

10. I use the word "strategy" instead of "act" here, following Briggs (2015); a strategy is a series of acts.

11. Note that not all risk-averse and rank-preserving decision frameworks will result in permissibly rational rejection of costless but relevant information. If, for example, $R(x) = x^{1.1}$, the REU(learning) would be about -46.8, which is still more than the REU(base) which is $Max(-112.7, -50) = -50$.

agent, Kinney and Bright worry that he will continue to have these mistaken credences.[12] The Buchakian model is thus a contender for an accurate model of instrumentally rational elite-group ignorance.[13]

To summarise, Buchakian risk-averse agents, who are privileged such that learning about their privilege would lead to learning that the world was biased, understand that it could be coherent for them to take the riskier choice. Given their current priors, they might rationally "protect" their future self from taking that gamble via information avoidance. Therefore, according to Kinney and Bright, it would be coherent in this rational framework for John to avoid learning whether B, even if it is costless and relevant. This is worrying if true; elite-group ignorance is harmful in itself and tends to beget more ignorance.

3 Instability and Uncertainty – How Stable is John's Rationality?

In this section, I show that the outcome of the Buchakian model relies on how strongly the agent can anticipate the decision consequences when choosing whether or not to seek or accept information about whether B (call this factor "Expected Relevance"). In addition, the effects of Relevance on the value of the information are also dependent on two other factors: The agent's **priors** when they're confronted with the information, as well as their **risk attitudes**[14].

3.1 The Relevance of Expected Relevance

Recall how John was given the chance to learn whether B *only as* he is about to decide whether to intervene – he knows that whatever information he learns will be immediately relevant to his situation. He worries that the information is misleading[15], and his risk-aversion kicks in.

12. Note that this not require John to be aware or intentional *at all* about his learning. What matters is that if John were presented with his decisions like in the figures above, he would not find them inconsistent; they would be in line with his risk-function.

13. Especially given that empirical research suggests that agents "who have not previously engaged in risky behavior (especially criminal behavior) tend to overestimate the probability of negative consequences."(Kinney and Bright 2021, pp.15)

14. I'll be referring to the two factors as "priors" and "risk attitudes" respectively.

15. I'm only focusing on 'misleading' in the Buchakian sense of the word. Note that the information can be completely accurate, but the agent sees the consequences of learning the information as leading them to choose dangerous actions.

However, if an agent cannot predict that they will definitely be confronted with said decision consequences, it will not immediately be obvious to them whether they should avoid learning. If learning is costless to begin with, they may be at least ambivalent about receiving the new information. This is because they are not yet able to calculate the risk-weighted utility given that they do not have the outcomes in mind.[16]

How could we model this? Perhaps we could say that John simply does not have the *Harassment* scenario as a consequence in his mind, and therefore that it does not factor into his calculation of the risk-weighted expected utility of learning whether there is bias. A perhaps more realistic option if John is indeed a really risk-sensitive agent is to model John as being uncertain about whether he will end up witnessing harassment with an allyship pamphlet in hand. If he's, for example, presented with an infographic about active bystanding when scrolling in bed at 2am with nary a heckler to be seen, he may not immediately believe that he will use this information to intervene in the future. Given that Buchak's model does not distinguish not anticipating a certain outcome with simply attributing 0% credence to it, the former option of not having the intervention scenario consequence can be represented as an extreme case of having < 1 credence that he will encounter *Harassment* (i.e. that his credence that he will encounter the scenario is 0).[17]

Now recall that when John's credence that B was 20% ('Possibility I'), the REU of the preferred choice (i.e. not learning whether B) in *Harassment* is -50. When the credence in B was 100% ('Possibility II'), the REU of the preferred choice (i.e. intervention) in *Harassment* is -47.5; at 0% ('Possibility III') it is also to reject learning and leave at -50.

However, John isn't always going to be 100% sure that he'd encounter *Harassment*. If, for example, he assigns a credence of 0.5 that he'd witness *Harassment*, we can add in Stage S, where John doesn't know whether he'll be encountering *Harassment* and is making a decision about whether to learn whether B, and possibilities I', II' and III' such that these stages are *after* John has made his decision to learn whether B (i.e. that he's living in a biased world) or

16. My argument structure mirrors the Jeffrey (1956) classic response to the problem of inductive risk.

17. While the actual phenomenology of attributing 0% credence to an outcome and neglecting to think about it is obviously different, they will result in the same decisions being made if the agent assumes that all the outcomes they can envision have a total probability of 1. To clarify, for modeling purposes the agent's credence that learning whether B will be relevant is interchangeable with how relevant they think the information will be.

avoid the information, but *before* he knows for certain that he'll be encountering *Harassment*.[18]

The utility of not learning at Stage S is simply $(-50) + (0.5)^2 \times 50 = (-37.5)$. In this case, since $(-36.575) > (-37.5)$, Learning is preferred to not learning at S in some cases where John does not have sufficiently high credence that he will witness Harassment. When it is less clear whether B will come in handy, John goes from refusing to learn whether B to being willing to pay for it.

Therein lies the crux of Expected Relevance: whether it is rational for privileged agents to avoid such information depends on how sure they are that they'll use the information; in fact, when the consequences of learning whether B are not yet obvious, they may instead choose to seek information.

Generalizing over varying credences that the information is relevant, John's attitude towards the information takes a parabolic curve (see 3). Intuitively: at 0 credence, he is absolutely sure that the information will never be useful, and thus is indifferent to learning it. The benefit from deciding over finer partitions then entices him to learn whether B from credence 0 to 0.61, before the fear of being tempted into the worst-case scenario takes over and he becomes averse to the information.

3.2 Ignorance Begets Ignorance

What if we instead tweak John's starting priors that B? For example, what if the numerous protests and activism efforts around John have caused more than an inkling of doubt that he was indeed privileged, and he had a credence of, say, 0.5 that B instead? How would that affect his attitude towards learning given varying credences about the relevance of the information?

Let's model this by modifying John's priors at stage S (Fig. 4). Not only is John's response to information dependent on Expected Relevance, but the interactions between the value of information and Expected Relevance are also a function of his prior in B. Specifically, the higher his initial credence that B, the higher his Expected Relevance has to be for him to rationally avoid the information.

In fact, while the relationship between the value of information and Expected Relevance remains parabolic at this risk attitude, depending on the agent's prior at stage S that the world

18. For calculations at each stage, as well as a detailed diagram, see 20.

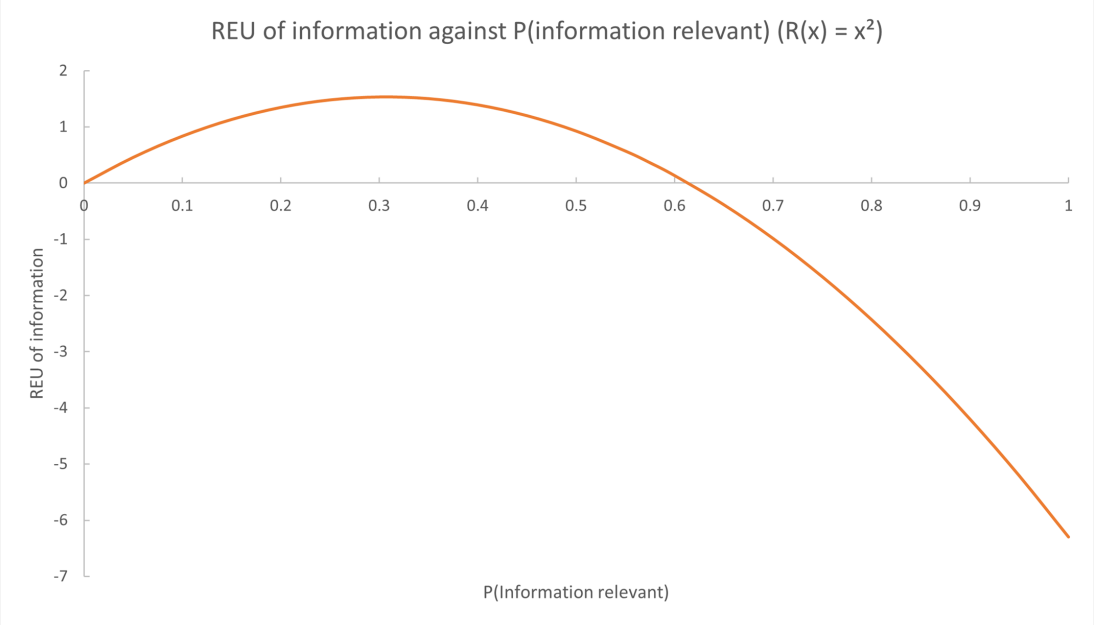

Figure 3: How REU(information) changes for John against his credence that the information will be relevant. Figures 3-6 have the same axes; they do not include the data point where P(information relevant) = 1.

is biased, they switch from seeking to avoiding information (at the x-intercepts) at different credences that the information will be relevant. For example, if John had had a prior of 0.1 that B, *ceteris paribus*[19], the set of situations in which he'd rationally be information-avoidant are a superset of the set of situations in which he'd rationally avoid the information with a prior of 0.5. Perhaps we could say that his higher credence has granted him some "immunity" against rational information avoidance.

In this case, John's prior ignorance has made him more likely to further avoid the information. Ignorance begets further ignorance here, and John stays in his Nguyenian inquiry trap as his priors "function to re-assert" the original belief system(Nguyen 2022). This could explain elite group ignorance's resilience and stability to intervention, corroborating both canonical and recent works in standpoint theory, and show how the persistence of this ignorance can arise from structural inequities such as faulty educational practices.[20]

19. This toy example also assumes that he'll only use information about B for the *Harassment* scenario.
20. For a survey, see Sullivan and Tuana (2007); for recent modelling work of standpoint epistemology, see Wu (2022).

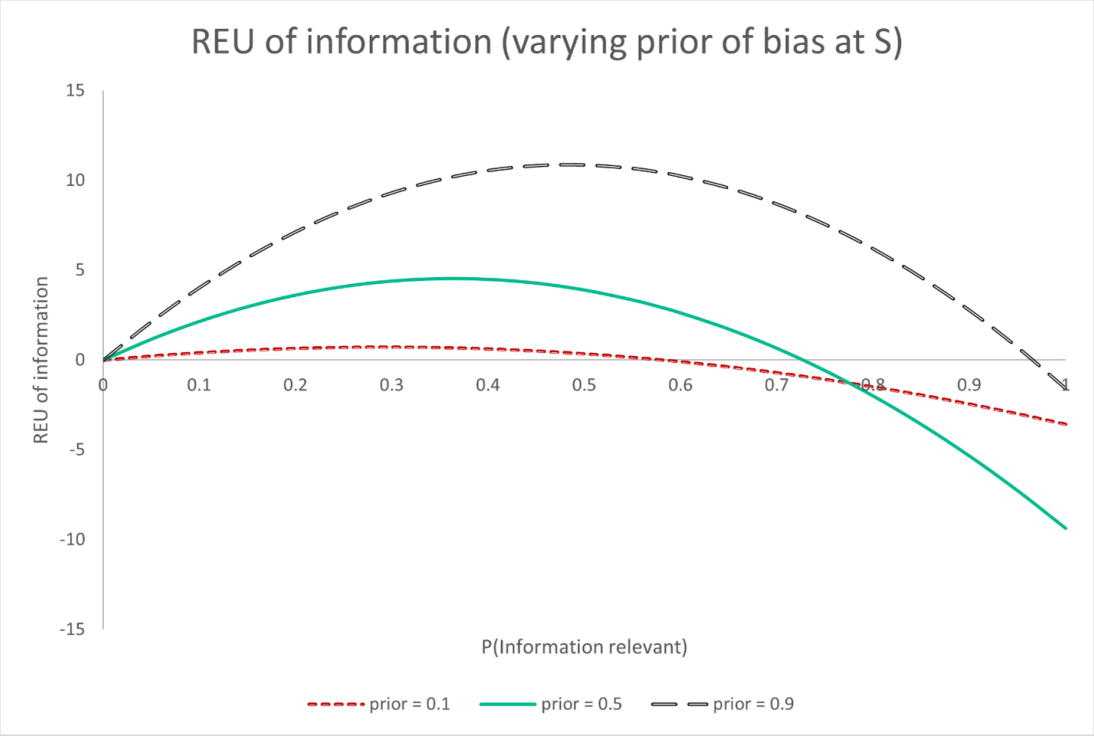

Figure 4: The axes are the same as in Fig. 3, but each line represents a different prior at stage S. Risk function: $R(x) = x^2$.

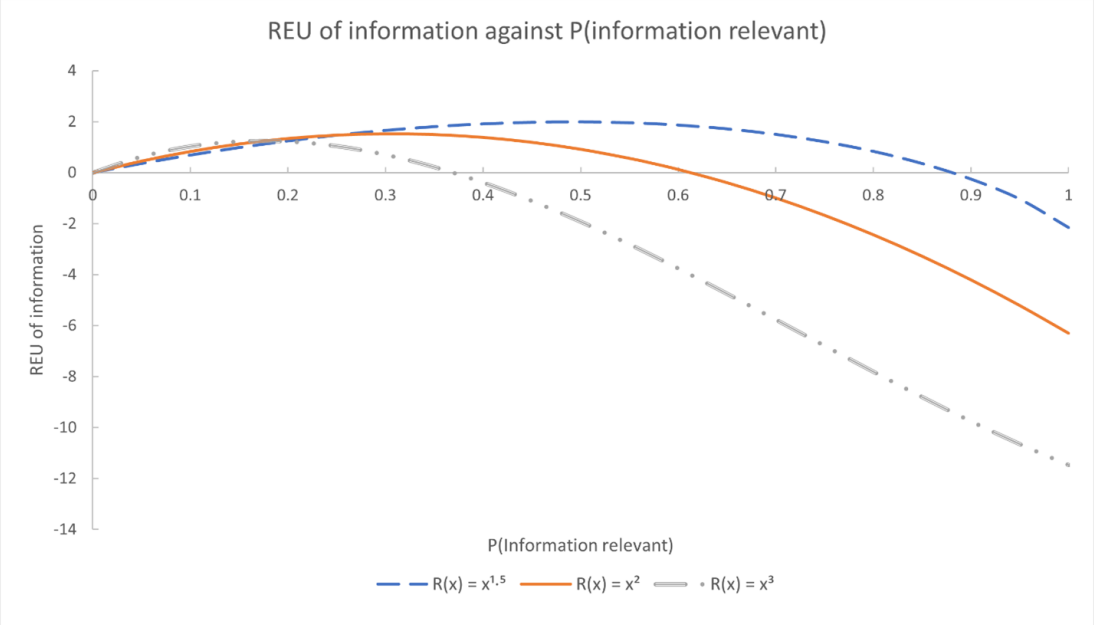

Figure 5: The curve for $R(x) = x^3$ is the most risk-averse curve of the three.

3.3 Risk Attitudes and Ignorance

What if John was more or less risk-averse, or perhaps even risk-seeking? From Fig. 5, we see that among risk-averse agents, the more risk-averse an agent is, the lower the x-intercept, i.e. the earlier it will be that the agent becomes information-avoidant as relevance increases. Ceteris paribus, then, a more risk-averse agent would rationally avoid information in at least as many cases as a less risk-averse one. Besides his priors, Expected Relevance's effect on the REU of learning is thus also modulated by the agent's risk attitudes.

This is perhaps unsurprising: the more risk-averse an agent is, the more they scale up the worst-case scenario in their utility calculations, and therefore the lower their threshold of relevance before they get scared off learning. In addition, elite-group agents are likely to be risk-averse in gambles like Harassment, where the gamble is between a high probability of modest gain and low probability of significant loss (Kinney and Bright 2021); this therefore explains why elite-group agents robustly avoid information partly due to their risk-aversion.

That being said, we have so far been focused on risk-averse agents, while Buchak's theory purports to allow for risk-neutral and seeking agents too.[21] Consider Fig. 6: The agent's

21. Many have taken issue with the consequences of risk-seeking behaviour in Buchak's framework, e.g. Zollman

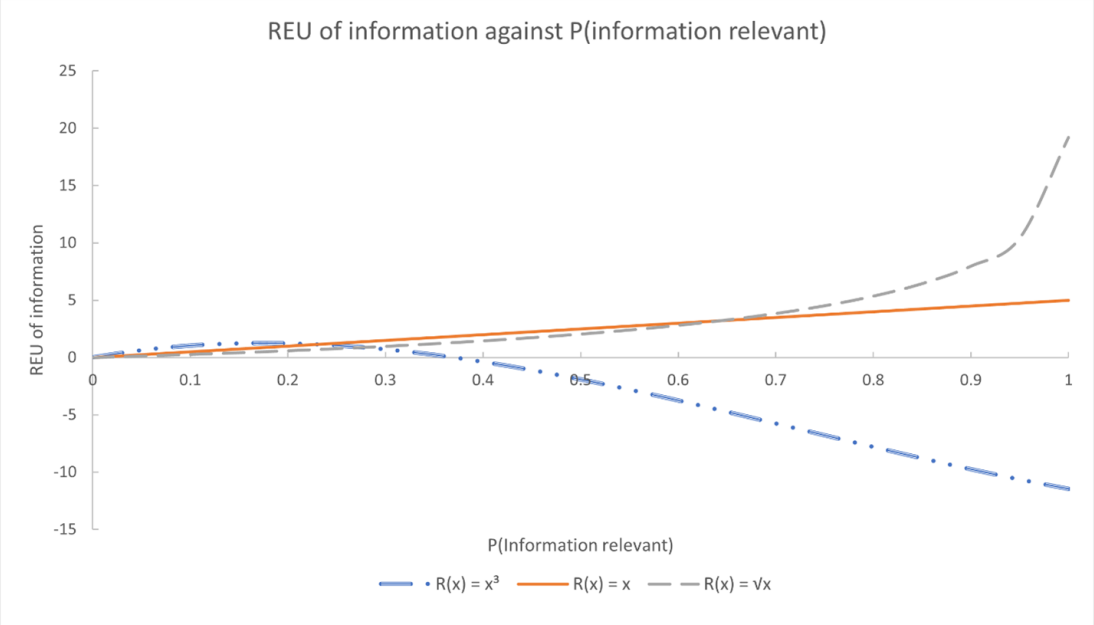

Figure 6: Depending on their risk function, an agent's attitude towards info evolves differently across their credence of information relevance. As above, $R(x) = x^3$ is when the agent is risk-averse; $R(x) = x$ is an example of risk-neutrality and $R(x) = \sqrt{x}$ of risk-seekingness.

attitudes towards information differs wildly if they are risk-seeking or risk-neutral compared to being risk-averse; for the risk-seeking agent, their risk attitudes no longer compete with the benefits of learning.

This raises questions for Buchak's framework as an explanation of attitudes towards information. Particularly, do risk-seeking agents take to learning like ducks to water, as the graph seems to suggest? This finding warrants further investigation beyond this paper's scope.

4 Resistance and Revolution: Political Upshots

Kinney and Bright dismiss the hope of "an informed and benevolent elite." In asserting the "[vanity]" of hoping that the elite could become "informed and benevolent," they presuppose that elite-group ignorance persists as a necessary consequence of an elite class. My findings in §3 are friendly to their arguments that the rationality of elite-group ignorance is a result of structural factors, especially the existence of elite groups in the first place. However, I push back on the vanity of interventions on the elite. Though elite-group ignorance and the elite

(2020).

class actively reinforce each other, I retain hope that, if we use the same framework, we can still intervene in the meantime.

Consider Martín's analogy to a castle under siege, which she uses to describe the "multiple kinds of mechanisms for active ignorance"(Martín 2021) both at the individual level and upstream of the individual. According to the analogy, on an individual level, targeting John's rationality is like one-on-one combat against him. However, on another level, there are "coordinated manoeuvres" like mass archery barrages that provide a "significant layer of protection" that make it hard for John to lose the fight, not to mention a moat or even the geography that can protect his eliteness from the commoners outside. Even if John could only defend himself from woke ally conversion because of multiple structural factors such as educational practices and the simple fact that he is an elite, his ignorance remains difficult to intervene on.

However, if the findings in §3 hold water, we can and should still chip at the defenses while the revolution brews. With these medium-term interventions, I believe we may mitigate some harms in the short run (e.g. in *Harassment*). Furthermore, given that elite-group ignorance is an active player in maintaining the social hierarchies we see(Martín 2021), if we undercut elite-group ignorance enough, we might even weaken the social structures that depend on it. When we finally storm the castle, it wouldn't hurt if the tower defenses have been disabled!

Thus, while I agree that John could rationally avoid information, the stability of that rational refusal is highly dependent on John's starting credences about what information he's going to receive and how useful it will be, as well as his risk attitudes. Even if we can't intervene on John's behaviour through criticizing his irrationality (because we're assuming that Buchkian agents are rational), these findings suggest other ways of destabilizing his rationality. If Kinney and Bright's employment of Buchak's framework holds water here, then it seems we have cause to intervene on the factors that result in John's rationality[22]. Perhaps then, as Martín hopes, we could replace practices that make elite-group ignorance easy to sustain with those that make it difficult to maintain.

22. Recall that I rely on a very thin notion of rationality – all I require from John's rationality is that it leads to the 'sticky' resilience against intervention.

4.1 While the Guards Sleep: Destabilizing Interventions

According to the modeling above, agents are rationally incentivized to avoid information when Expected Relevance exceeds a threshold. If this is true, then contra in-your-face efforts to convince elites to intentionally "sit with" the information of their privilege (e.g. DiAngelo 2018), we could consider treating learning about their privilege as a self-effacing end that "cannot be acquired through direct pursuit."(Nguyen 2022; see also Parfit 1984; Nguyen 2020). Like blending white beans into broccoli cheddar soup instead of forcing a bean-hater to chug a can of beans, we could incentivize John to learn about his privilege through other autotelic activities – i.e. activities that are done for their own sake – that distract him from the consequences of learning.

For example, he could choose to learn about his white privilege by playing a game that implicitly teaches the differential incarceration rates in the U.S. and how they impact communities. If John is so invested in the investigation itself that he is distracted from the worst-case scenario (i.e. the information turns him into a daredevil), his temporal decision-making horizon could be intentionally restricted such that he achieves the self-effacing goal of learning the information, but circumvents the clutches of his own instrumental rationality.[23]

Intervening when the ignorance is already resilient is unsurprisingly difficult; instead, the results suggest that early intervention will also pay dividends in narrowing the possibilities for rational information avoidance. Especially given that formalized education adds its own incentive structure beyond the consequences of acting with one's privilege in mind, it will also serve as an intervention on Expected Relevance. The findings above then corroborate calls for interventions within the early education system.

While all risk attitudes may be permissible, some may be less convenient than others. Therefore, a possible intervention could be to cultivate flexibility in terms of risk attitudes through being open-minded. Perhaps through education or gameplay, John could learn to foster an "openness to surprise"(Lugones 1987) that will allow him to perspective shift. Perspective taking and "active open-mindedness" have been at least somewhat effective in diminishing outgroup stereotypes(Lilienfeld, Ammirati, and Landfield 2009); perhaps there remain ways

23. For more on intentional framing, see Thoma 2018.

to get past the tower defenses and allow for effective elite-group privilege education, albeit sneakily.

4.2 Checking our Tools: Testing the Model

Interventions are high-stakes and require care. Given how speculative the model and therefore the findings are, it would be responsible to only pursue the suggestions above after the model has been sufficiently tested. How much are the results in §3 reflected in real life? While current studies cannot yet empirically determine how the interventions would actually affect information avoidance, I propose a few directions the research could take. In addition, independent of the applications, my findings illustrate *just how unstable* the framework could be.

The effectiveness of interventions on Expected Relevance require empirical research, such as investigations on whether elite-group agents could ever be distracted from the consequences of learning. What would "distracting" the agents look like, and how much would it cost? Telling people at a DEI meeting that the information they're learning is irrelevant to them doesn't seem like it'd create a great learning environment, for instance, but could we measure how varying whether the module is conducted via a game or just a lecture affects the audience's ability to recognize how their privilege affects the stakes they've been given, and perhaps even if they are more receptive to future learning opportunities (which my findings suggest would be a result of having successfully learned about their privilege and thus updated their priors).

Why do my current results still matter? Models are tools of inquiry as much as they are tools of prediction; I hope that my findings provide motivation for investment of research resources to empirically ground or dispute the model. Even if empirical results suggest wariness about the effectiveness of the interventions above, they will provide impetus to revise or search for models that better capture these important phenomena. If it turns out that decreasing Expected Relevance does not improve learning outcomes, then perhaps not only should we reject the results in §3, but also the Buchakian model altogether, as a fitting model of elite-group ignorance.

5 Conclusion

We started by outlining how Kinney and Bright employ a Buchakian risk-averse decision model as a proof of concept that not all elite-group ignorance is necessarily irrational. They argue that in these cases, because Good's Theorem fails to apply, risk-averse elite agents plausibly rationally refuse costless relevant information about their privilege; even decreasing costs of education will not result in the elimination of elite-group ignorance. This has worrying consequences about the efficacy of current efforts to educate the elite and trust that they will make better decisions with access to more information.

In response, this paper has investigated the different points at which an agent could be rationally required to learn. We have seen that Kinney and Bright's model results are not robust to variations in parameters, including the agent's Expected Relevance of the information. The degree to which the agent is sensitive to the relevance of the information depends too on their priors and their risk attitudes. Therefore, while the rationality of elite-group ignorance sits upon powerful structural factors, I suggest that there are still cracks where we can intervene to destabilize the rational information resistance, and instead make it rational to learn information about one's privilege.

The effectiveness of the suggestions above require substantiation with further empirical work before actually being implemented. But not all hope is yet lost for intervention, even with how entrenched elite-group ignorance is as a tool and product of oppression. Redistribution and restructuring can remain our end goals; that does not, however, preclude us from taking steps now to prime ourselves for a successful revolution.[24]

24. I owe a large debt of gratitude to (in alphabetical order) Liam Kofi Bright, Kevin Dorst, Carolina Flores, Seth Goldwasser, Shelby Hanna, David Kinney, Taylor Koles, Annette Martín, Sven Neth, Felipe Pereira, Joseph Schiavone, Gabriel Vasquez-Peterson, Mark Wilson and Elise Woodard, and several anonymous reviewers, and the audiences at PFEW, PSA 2022, APA Central 2022, and the National University of Singapore (where the paper was presented under the name "Accidentally I learnt: On Relevance and Information Resistance") for their extensive feedback and discussions on previous versions of this paper.

Calculations for Section 3

Recall that:

$$REU(Learning) = U_1 + R(P_2 + P_3) \times (U_2 - U_1) + R(P_3) \times (U_3 - U_2)$$

$$U_1 = (-250); U_2 = (-50); U_3 = 0; R(x) = x^2$$

Stage/Possibility	P_1	$P_2 + P_3$ $(= 1 - P_1)$	P_3	REU(Learning)
S	$= 0.5 \times (0.2 \times 0.1)$ $= 0.01$	0.99	$=$ no *Harassment* + *Harassment*, biased, safe intervention $= 0.5 + (0.5 \times (0.2 \times 0.9))$ $= 0.59$	$= (-250) + (0.99^2 \times 200) + (0.59^2 \times 50)$ $= (-36.575)$
I' = III'	0 (John will always leave)	1	$=$ no *Harassment* $=0.5$	$= (-250) + (1^2 \times 200) + (0.5^2 \times 50)$ $= (-37.5)$
II'	$= 0.5 \times 0.1$ $= 0.05$	0.95	$=$ no *Harassment* + *Harassment*, biased, safe intervention $= 0.5 + (0.5 \times 0.9)$ $= 0.95$	$= (-250) + (0.95^2 \times 200) + (0.95^2 \times 50)$ $= (-24.375)$

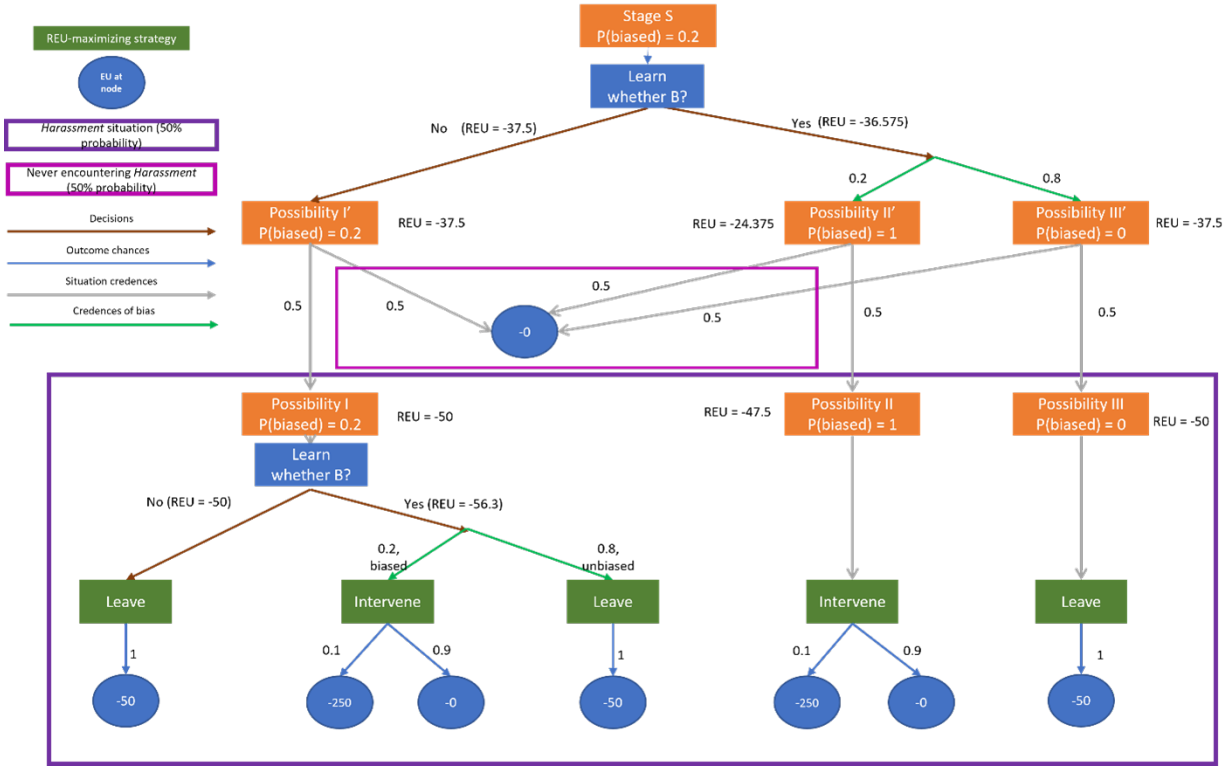

Figure 7: We've added S, I', II' and III' and the uncertainty of whether John will witness *Harassment* to John's decision process.

References

Briggs, R.A. 2015. "Costs of abandoning the Sure-Thing Principle" [in en]. *Canadian Journal of Philosophy* 45, nos. 5-6 (December): 827–840. ISSN: 0045-5091, 1911-0820, accessed June 8, 2021. https://doi.org/10.1080/00455091.2015.1122387.

Buchak, Lara. 2017. "Precis of Risk and Rationality" [in en]. *Philosophical Studies* 174, no. 9 (September): 2363–2368. ISSN: 0031-8116, 1573-0883, accessed June 8, 2021. https://doi.org/10.1007/s11098-017-0904-7. http://link.springer.com/10.1007/s11098-017-0904-7.

Buchak, Lara Marie. 2013. *Risk and rationality* [in en]. OCLC: ocn841520426. Oxford: Oxford University Press. ISBN: 978-0-19-967216-5.

DiAngelo, Dr Robin. 2018. *White Fragility: Why It's So Hard for White People to Talk About Racism* [in en]. Google-Books-ID: abZdDwAAQBAJ. Beacon Press, June. ISBN: 978-0-8070-4741-5.

Duguid, Michelle M., and Melissa C. Thomas-Hunt. 2015. "Condoning stereotyping? How awareness of stereotyping prevalence impacts expression of stereotypes" [in English]. Num Pages: 343-359 Place: Washington, US Publisher: American Psychological Association (US), *Journal of Applied Psychology* 100, no. 2 (March): 343–359. ISSN: 0021-9010, accessed August 22, 2022. https://doi.org/10.1037/a0037908. https://www.proquest.com/docview/1610752947/abstract/E725D75097CA4B5BPQ/1.

Good, Irving John. 1967. "On the Principle of Total Evidence." *The British Journal for the Philosophy of Science* 17.4:319–321. https://doi.org/https://doi.org/10.1093/bjps/17.4.319.

Guerrero, Alexander. 2021. "The Epistemic Pathologies of Elections and the Epistemic Promise of Lottocracy." In *Political Epistemology,* edited by Elizabeth Edenberg and Michael Hannon, 156–179. Oxford University Press. ISBN: 978-0-19-289333-8.

Jeffrey, Richard C. 1956. "Valuation and Acceptance of Scientific Hypotheses" [in en]. *Philosophy of Science* 23, no. 3 (July): 237–246.

Kinney, David, and Liam Kofi Bright. 2021. "Risk Aversion and Elite-Group Ignorance." *Philosophy and Phenomenological Research.*

Lilienfeld, Scott O., Rachel Ammirati, and Kristin Landfield. 2009. "Giving Debiasing Away: Can Psychological Research on Correcting Cognitive Errors Promote Human Welfare?" [In en]. Publisher: SAGE Publications Inc, *Perspectives on Psychological Science* 4, no. 4 (July): 390–398. ISSN: 1745-6916, accessed August 24, 2022. https://doi.org/10.1111/j.1745-6924.2009.01144.x. https://doi.org/10.1111/j.1745-6924.2009.01144.x.

Lugones, María. 1987. "Playfulness, "World"-Travelling, and Loving Perception." Place: Oxford, UK Publisher: Blackwell Publishing Ltd, *Hypatia* 2 (2): 3–19. ISSN: 0887-5367. https://doi.org/10.1111/j.1527-2001.1987.tb01062.x.

Martín, Annette. 2021. "What is White Ignorance?" [In en]. *The Philosophical Quarterly* 71, no. 4 (September). ISSN: 0031-8094, 1467-9213, accessed July 22, 2022. https://doi.org/10.1093/pq/pqaa073.

Mills, Charles W. 2007. "White Ignorance" [in en]. In *Race and epistemologies of ignorance,* edited by Shannon Sullivan and Nancy Tuana, 13–38. SUNY series, philosophy and race. OCLC: ocm70676503. Albany: State University of New York Press.

Nguyen, C. Thi. 2020. "Echo Chambers and Epistemic Bubbles" [in en]. Publisher: Cambridge University Press, *Episteme* 17, no. 2 (June): 141–161. ISSN: 1742-3600, 1750-0117, accessed May 6, 2022. https://doi.org/10.1017/epi.2018.32. https://www.cambridge.org/core/journals/episteme/article/abs/echo-chambers-and-epistemic-bubbles/5D4AC3A808C538E17C50A7C09EC706F0.

———. 2022. "Playfulness Versus Epistemic Traps." In *Social Virtue Epistemology,* edited by Mark Alfano, Colin Klein, and Jeroen de Ridder. Routledge.

Parfit, Derek. 1984. *Reasons and Persons.* Oxford University Press. ISBN: 0-19-824908-X.

Sullivan, Shannon, and Nancy Tuana, eds. 2007. *Race and epistemologies of ignorance* [in en]. SUNY series, philosophy and race. OCLC: ocm70676503. Albany: State University of New York Press.

The Clash. 1982. *Should I Stay Or Should I go.*

Thoma, Johanna. 2018. "Risk Aversion and the Long Run." *Ethics* 129:230–253.

Wu, Jingyi. 2022. "Epistemic Advantage on the Margin" [in en]. *Philosophy and Phenomenological Research,* 31.

Wynn, Alison T. 2020. "Pathways toward Change: Ideologies and Gender Equality in a Silicon Valley Technology Company." Publisher: SAGE Publications Inc, *Gender & Society* 34, no. 1 (February): 106–130. ISSN: 0891-2432, accessed August 19, 2022. https://doi.org/10.1177/0891243219876271. https://doi.org/10.1177/0891243219876271.

Zollman, Kevin J S. 2020. "On the normative status of mixed strategies" [in en] (August): 43.