Demographic Statistical Evidence with a Humane Face

Somayeh Tohidi

A thesis submitted for the degree of
Doctor of Philosophy

Department of Philosophy, Logic and Scientific Method
London School of Economics
12/01/2025

Declaration

I hereby declare that the work presented in this thesis is my own. I have not plagiarized or made use of any external sources except where properly cited. This thesis is submitted in partial fulfillment of the requirements for the degree of PhD in Philosophy at the London School of Economics.

The work contained herein is my own original research, carried out under the supervision of Dr. Liam Bright and Prof. Richard Bradley. However, I would like to note that Chapter 1 of this thesis is co-authored by Dr. Lewis Ross, and I have fully acknowledged his contributions in the appropriate section.

The final word count of this thesis is: **32,257** words.

Somayeh Tohidi
Date: January 13, 2025

Abstract

This thesis shows that suspicion about the existence of bias in demographic statistical studies which align with social stereotypes and potentially support one side of a major political controversy can provide sufficient grounds for rationally suspending judgment about their content and refusing to use them in decision-making. Using the framework of bounded Bayesianism, it explores the perspectives of both statistically sophisticated agents and statistically novice yet socially-aware agents towards these studies. It concludes that both types of agents can have adequate reasons to rationally reject believing and using such studies. The thesis also examines the language of statistical reports about social groups and shows how, in different contexts, they can imply harmful essentialist claims about those groups.

To Omid,
for setting me free

Acknowledgments

To my mum,
for showing me how love can bend even the most stubborn social norms to its will.

To my dad,
for living as though love and integrity were worth more than anything money could buy.

To the team Liam and Richard,
for their unreserved support and their magic in resurrecting a long-dead self-confidence.

To Liam,
for never giving up on me and showing me how philosophy can be relevant to the social world.

To Richard,
for helping me find the most versatile ground for thinking by teaching me decision theory.

To Kevin,
for giving me a lifetime opportunity by hosting me at MIT.

To Lewis,
for his genuine support and for making me feel welcome in a new environment.

To Sally,
in whom I found a role model.

To Zainab and Ali,
for proving that hearts can stay close even when oceans apart.

To Bele, Dominic, and Katariina,
for being my anchors through stormy waters.

To Arefeh, Soodabeh, Mehdi, Mohammad, and Sajjad,
for bringing the warmth of family to a distant land.

Contents

Introduction

A woman from New York has witnessed a horrible crime committed by an immigrant against another woman during her daily walk from work. Outraged, she decides to weather her storm by posting a video on social media. In the video, she addresses "feminist women on the left" and expresses her anger toward them for not supporting anti-immigrant policies. She concludes her speech by asking, "How do you sleep at night? How do you vote for this, and *do you not feel like a literal walking statistic living in these places*?"

She seems to be saying that she is annoyed to be part of a population for which there is clear statistical data that feminist women on the left ignore. But what is that statistic? Does she mean statistics about public gender-based violence? Or statistics about the rate of such violence among immigrants? Maybe she has a specific statistic in mind about the comparative rate of crime against women in public among immigrants and non-immigrants. But does such a statistical study really exist? If it does, how reliable is it, and what does it exactly prove? Is it merely a correlation or a causation? If it is causation, how stable is it? And more importantly, how can we know the answer to these questions as laypeople with limited knowledge of the ever-expanding realm of statistical methods?

This work started with a concern over how statistics about demographic groups are increasingly used to justify policies and behaviors against minorities and oppressed groups in public discourse. A concern over the gap between how reliable and relevant these statistics actually are as pieces of evidence to justify those policies and behaviors, and how they are treated by us as science consumers (as opposed to scientists). This concern leads to three *normative* questions about using, believing, and reporting *the results of statistical studies about demographic groups*:

1. Should we *use* those results in our decision-making?
2. Should we *believe* those results?
3. How should we *report* those results?

There are numerous occasions when it is hard not to feel conflicted and torn between morality and rationality when facing the first two questions. These are occasions in which the statistics report a high prevalence or intensity of an undesirable characteristic among a demographic. For example, it may seem morally wrong, yet epistemically right, to use such statistics to *form beliefs about random members* of that demographic[22]. There is an ongoing debate on this conflict, which has led to theses such as the moral encroachment view [5] and the suspension of judgment view [53].

What is rarely discussed in the literature is the conflict we feel when using demographic statistics to make *decisions under uncertainty* about members of different demographics. This is because, if we make use of such evidence, the expected utility-maximizing norm often dictates choosing actions and behaviors that can align with oppressive social practices and thus become morally suspicious. For example, if a landlord uses statistics showing a significant difference in the rate of crime between immigrants and non-immigrants in a neighborhood, they may never sign a rental contract with an immigrant family in that neighborhood. Similarly, if an employer uses statistics showing that, in their field, the majority of women do not perform well in management positions, they may end up rejecting all female applications for a management position[1].

This conflict stems from the effect of statistics about demographics on our credences (rather than beliefs) about random members of those demographics. If we expose our credences to those statistics, then instrumental rationality can easily lead us to make decisions that go against our moral intuitions. What should we do in the face of this conflict? Should we shield our credences by ignoring those statistics? Are we able to do that at all, and even if we are, is it rational to do so? In mainstream Bayesianism, credences are updated automatically, and the agent has no agency to shield them. Moreover, even if we can, shielding our credences against a piece of evidence goes against Good [24]'s principle of total evidence, according to which ignoring a piece of relevant evidence before making a decision is instrumentally irrational.

If we start thinking outside the box of mainstream Bayesianism, we will be able to shield our credences without going against any rationality norm. This is because it can be argued that, as bounded Bayesians, we have control over when to update our credence function. Moreover, the principle of total evidence is not blatantly applicable to us. This is because Good's principle assumes no limitations for rational agents and hence no cost in evidence gathering and processing. But that is not the case for bounded Bayesians! Even if we assume that sometimes evidence just hits us, making evidence gathering cost-negligible, evidence processing is rarely cost-negligible for us as cognitively bounded agents.

Updating our credence based on a new proposition requires using conditional credences. For example, if we want to update our credence of X based on E, we need to make a judgment about the precise probabilistic dependence between X and E, i.e., $P(X|E)$, and that is not something that happens to us instantly or without cognitive labor. Conditional credences, unlike what mainstream Bayesianism depicts, are not always readily accessible. The space of possibilities over which we maintain a credence function is very coarse-grained. We do not have a readily accessible credence for every combination of propositions. When we face a proposition that is not already included in the algebra of our credence function, we need to add a level of fine-graining to our space of possibilities, and that is cognitively costly, as it involves multiple cognitive tasks. We need to *estimate* the precise probabilistic dependence between the new proposition and the old ones while performing mathematical calculations to ensure that the resulting credence function respects Kolmogorov's rules.

[1] Bovens [9] shows that the expected utility-maximizing norm does not necessarily demand such a decision. However, his argument focuses on statistical studies of the relative *degree of performance* in a trait between two demographics, rather than the relative *prevalence* of that trait, and is therefore not generalizable to all cases of statistical studies about demographic groups. For example, if we have detailed statistics on the distribution of scores among men and women on a math test and use them for recruiting decisions for a math-relevant position, then Bovens [9]'s argument is applicable. However, if we only have statistics on the *prevalence* of those scoring above a threshold among men and women and use them for decision-making, his argument is not applicable.

In fact, estimating the probabilistic dependence is quite difficult in the case of statistical evidence. If X is “a random immigrant in this neighborhood has committed a crime” and E is “a statistical study shows 40% of immigrants in this neighborhood have a criminal record”, then what is $P(X|E)$? Is it 0.4? Why? That number is just the result of a statistical study on ‘having a criminal record’ and not the objective chance of ‘having committed a crime’. Babic et al. [2] details a plausible procedure for finding that conditional probability. Reviewing that procedure makes it obvious that such a probabilistic judgment is not easy at all.

So, if we take evidence processing as updating our credence function, then it involves both *estimation* (for judging probabilistic dependence) and *mathematical calculation* (for respecting probabilism). It is very implausible, if not impossible, that a cognitive task involving estimation and calculation is unintentional and involuntary. We need intention to perform these kinds of cognitive tasks. So, we may decide not to do it, and not doing it, given the inapplicability of Good’s theorem to us as bounded agents, does not necessarily go against instrumental rationality. Does that mean we are done and can easily go ahead and ignore demographic statistics whenever they lead to morally suspicious decisions? Of course not! Updating credence improves accuracy of the credence function and can also improve our expected utility with respect to the decision under uncertainty we are about to make. That happens when updating changes our mind about the optimal option in that decision. So, we need to take into account this possibility when deciding whether to update.

So, it seems that the decision to update, which is in fact the decision to fine-grain our space of possibilities to make the conditional probability accessible, is itself a decision under uncertainty about whether that updating leads to a change in the decision we are about to make or not. That uncertainty itself can be translated into two uncertainties: (i) uncertainty about the *strength of the evidence* with respect to the decision that we are about to make later (degree of relevance to the uncertainty of that decision), and (ii) uncertainty about what we may learn later in our *line of enquiry* and before making the decision.

It can be proved that we are rationally obliged to update our credence of X on a piece of evidence E *only if* we are sufficiently confident that E is going to change our mind about the decision under uncertainty about X we were about to make, AND sufficiently confident that we will not learn any other piece of evidence that undermines the effect of E on that decision. The threshold for this ’sufficient confidence’ is a function of the cognitive cost (disutility) that we assign to fine-graining (updating) and the utilities involved in our decision (under uncertainty about X). The details of this proof have been presented in Chapter 0.

So, if for any reason we consider it *very unlikely* that E will change our mind about the decision, or that E will be easily undermined by other evidence in our line of enquiry, we are rationally obliged to disregard E and avoid wasting cognitive resources on updating based on it. In Chapter 1, I draw on Babic et al. [2]’s account of Bayesian updating on noisy statistics to argue that if E is a statistical report aligned with a social stereotype, and we are statistically sophisticated, that is, familiar with Bayesian statistical methods, then in many plausible decision-making situations, we will not be sufficiently confident that E will change our mind about the decision. Consequently, ignoring such statistics and refraining from updating is rational, resolving any potential conflict between rationality and morality.

But what if we are not statistically sophisticated? How should we treat these studies as laypeople? In Chapter 2, I argue that as socially aware agents, when we receive statistical studies *supporting one side of a political controversy*, we should consider the possibility of bias in those studies. Even if we can't directly spot methodological bias, our social awareness makes it likely that we will notice bias through the studies' connection to political bodies. This means we can't be sufficiently confident that trusting these studies, or letting them influence our decisions, won't later be undermined by new evidence. So, in many cases, it is rational to ignore such studies and not update based on them.

Being open to the possibility of bias in these studies has an interesting consequence for the rationality of maintaining categorical belief in them. In Chapter 2, I draw on Leitgeb [39]'s probabilistic account of belief to argue that fine-graining our space of possibilities to include the possibility of bias undermines the stability of our credence in their content. So, even if we treat the statistical study as expert testimony and start with a high prior credence in its truth, the fine-graining with respect to the possibility of bias makes our credence unstable, rendering belief in its content irrational.

But why does *believing* the content of a statistical study matter? Is there anything morally contemptible about it? There is clearly a strong moral intuition against believing that a random individual has a pernicious trait just because of statistics about their group - this is what the literature on moral encroachment addresses. But what makes believing the statistics itself morally wrong? At first glance, believing the statistics might seem harmless, except that it enables morally suspicious beliefs about specific individuals. However, there is more at stake here, because demographic statistics are probabilistic propositions about social groups.

In Chapter 3, I will discuss the propensity interpretation of probabilistic propositions and show why this interpretation is a plausible interpretation for probabilistic propositions about social groups. I will show why this interpretation is equivalent to essentialist claims about social groups, believing which can have harmful consequences. I also show why this interpretation can be considered as the conventional implicature of demographic statistical statements. Then I explain why, in decision-making conversations, these statements can imply unwarranted suggestions about the optimal intervention, which are aligned with oppressive social practices. I finish that chapter by introducing two strategies to block these harmful implicatures.

I wrote the first two chapters of this thesis as independent papers. Although both were grounded in similar intuitions, I initially failed to recognize their connection. It was only after simplifying the model that I realized they are based on exactly the same framework. As a result, their findings are generalizable to one another and can be seamlessly merged. So, in Chapter 4, I will present a refined version of the model that served as the foundation for both Chapter 1 and Chapter 2 and review some of their content. This more developed model helps show how the findings of these two chapters connect to each other.

Chapter 1

A Good Bayesian Has Faith in You

Abstract

A great deal has been written about the rational and moral status of demographic profiling—drawing inferences about people based on characteristics like race, gender, or age. Much of this work attempts to thread the needle between vindicating the idea that profiling is morally suspect, while remaining faithful to plausible rational principles about the formation of belief. There has been much less attention on the problem of profiling from the perspective of forming credences. The problem of profiling seems especially difficult in a credence-based framework, since it is widely accepted that credences should be formed based on all the available evidence. In this paper, we explain why and when it can be entirely rational for a social egalitarian to disregard demographic statistical evidence when forming beliefs about individuals and making decisions about them. Our argument is based on a principle we introduce for evidence contemplation for a bounded (non-ideal) Bayesian and Babic et al. [2]'s claim about the effect of updating upon noisy statistical evidence on one's credence. We then use Buchak [11]'s notion of faith to interpret the social egalitarian's refusal to engage with demographic statistical evidence in decision-making about individuals as having 'faith' in them.

1.1 Introduction

There is evidence suggesting that certain characteristics and conditions are more prevalent in some demographic groups than others. This includes characteristics and conditions often thought to be undesirable. For example, some demographic groups have comparatively high levels of criminality, alcoholism, or depression. This raises the question—how we should deploy this evidence when forming opinions about members of these groups?

A live debate concerns the permissibility of drawing profiling inferences ('profiling'). By profiling we mean inferring that some individual has an undesirable property just based on their membership in a socially salient demographic group, like gender or race. It is reasonably straightforward to envisage moral reasons to be uncomfortable with profiling as a social practice. Yet it is more difficult to see why these concerns would render it epistemically irrational to profile people, if we grant that there can be some accurate statistical information about different demographics. Nevertheless, various philosophers have baulked at Tamar Gendler's 'sad conclusion' that we can be rationally compelled to engage in profiling against members of certain demographic groups[1]. There are now several accounts aiming to explain why we are rationally entitled not to draw profiling inferences, with these views often diverging in what they say about the intersection between moral and epistemic norms.

Although these proposals contain much that strikes us as insightful, extant views tend to focus primarily on belief. This leaves open the possibility that profiling may be required when forming credences rather than full beliefs. Indeed, one might think that the inclusion of profiling evidence is straightforwardly required by widely endorsed Bayesian norms for conditionalizing credence on all available evidence.

[1] Gendler [22].

Our paper explains why not integrating profiles into the formation of credences is rationally permissible. We also identify situations in which, plausibly, it is rationally impermissible to integrate that information. We begin by surveying two promising accounts of why profiling is impermissible: the moral encroachment view and the suspension of judgement view. While acknowledging the strengths of each, we raise problems for both accounts. We then introduce a framework in formal epistemology for evidence contemplation for a non-ideal Bayesian and we will show why it can accommodate the suspension of judgement view and resolves the problems it faces.

1.2 Moral Encroachment

Among the most prominent anti-profiling accounts appeals to the idea that epistemic rationality is 'stake-sensitive', roughly: sensitive to the costs of error. This view is generally referred to as a form of moral encroachment, drawing on the older tradition of pragmatic encroachment in epistemology on which the pragmatic 'encroaches on' the intellectual[2]. It is also closely related to the phenomenon of inductive risk familiar to philosophers of science [54][31][17][3]. In short, moral encroachment theories take the epistemic permissibility of making a judgment to be partly determined by the moral risks of accepting that judgement[4]. The greater the moral risks involved in accepting a given proposition, the more demanding the standards that must be met to rationally believe that proposition[5].

Renee Jorgensen, an influential defender of this view, writes:

> [A] crucial element of epistemic responsibility is to accurately judge when the evidence we have is sufficient for acceptance, given what we risk in the event of a mistake, and when we must keep inquiry open and continue to seek better evidence. [7, p. 2422]

Integrating demographic evidence in your opinions about particular individuals then using it to justify an action imposes a risk of harm on person being profiled. In some individual cases the harm may simply be embarrassment (e.g. a doctor making an assumption about a patient), but in others it may license much more serious harm (e.g. discriminatory policing). Moreover, even seemingly trivial cases, multiplied over many like instances, can contribute to an oppressive environment for members of certain groups. As Frye [20] evocatively puts the idea: while each individual wire is no serious impediment, the entire network of wires constitutes a birdcage which seriously impedes a life.

The moral risks inherent in profiling, on moral encroachment views, raise the bar

[2] See Kim [38] for a general introduction.

[3] For a contemporary review of all such arguments in philosophy of science see Ward [63].

[4] This is to be distinguished from a different view, explored in Basu [4], on which certain beliefs can be morally wrong *even if* epistemically rational.

[5] For further discussion see, among others, Pace [49], Moss [46], Gardiner [21], Fritz and Jackson [19].

for when it is epistemically rational to form profiling beliefs. However, moral encroachment views generally do not claim that statistical generalisations are necessarily impotent as support for rational inferences about members of a demographic group, even when these inferences concern undesirable characteristics. Indeed, as Jorgensen herself is careful to say (ibid.), the rejection of profiling: "[I]sn't because generalizations are intrinsically incapable of grounding moral conclusions." Rather, the rejection of profiling evidence is explained by a situational recognition of the risks that inhere in profiling people in different scenarios[6]. This means that the injunction not to profile is only contingent on the cost/benefit calculus aligning in a particular way in a particular case. The moral encroachment view does not, for example, identify any flaw in 'private profiling', where no particular action is at stake. Whether this is a limitation of the moral encroachment view is not entirely uncontroversial, but it does certainly seem like there is something unsavoury about profiling others even as a pure spectator.

Regardless of what one thinks of private profiling, we contend that there is a fundamental problem with the moral encroachment view: namely, it only generates a prohibition on profiling in relatively trivial cases. The typical examples used to motivate the moral encroachment view are those wherein the reasons *favouring* profiling are rather weak, e.g. the availability of a convenient heuristic when trying to work out: whether someone is a waiter, how much someone is likely to tip, or whether someone has stolen some cutlery[7]. Against these scant convenience-based rewards, proponents of moral encroachment enjoin us to weigh the risks of causing considerable offence and upholding destructive patterns of social and institutional racism. However, the risks and rewards in a given case may not always organise themselves to raise the epistemic standards for engaging in profiling. For example, we can devise cases where the costs of not profiling are more serious. The following case is fanciful and extreme, but brings out a general tactic for generating cases where profiling is permitted and (arguably) epistemically obligatory, by the lights of the moral encroachment view:

> *Heist.* You are head of security for an extremely effective charity. A wealthy patron is coming in to donate a jewel to be auctioned—for charitable purposes—later that evening at the Fancy Club. You are guarding the jewel. Stepping out for a cigarette, you return to see the jewel gone. Someone shouts, "One of the waiters snatched it, go after him! Down the corridor!" Sprinting down the corridor, you come to a fork. On your right, you see a black face behind an elevator door as it closes shut, with the arrow pointing UP. Down the left-hand corridor, you see a white face behind a second elevator door as it closes shut, with the arrow pointing DOWN. You know there are exits on each floor. Knowing that black people in your society are much more often waiters than wealthy patrons, which way should you go?

[6] For example, Bolinger [7, p. 2426] suggests that a profiling proposition might be rational to believe when one is a mere spectator of a scenario but irrational to believe when one is an active participant.

[7] See Gendler [22] and Bolinger [7] on being a waiter, Basu [4] on tipping, Ross [53] on cutlery.

Here, the risks of *not* profiling are much higher than the risks of profiling—you risk losing thousands of dollars for charitable causes (that will save scores of lives, we can stipulate), weighed against the harm caused to someone who is incorrectly profiled as a waiter and for criminal activity. So, the moral encroachment view cannot say about this case that the standards for profiling are increased[8]. Indeed, if anything, the view would seem to suggest that the epistemic standards for profiling ought to be lower than usual—given the possible downstream harms of failing to retrieve the jewel.

This case is stylised, but you can in fact substitute any more realistic and humdrum case where there is a risk of losing a certain amount of money or a risk of suffering personal harm which outweighs the disutility of the other party being incorrectly profiled. The disutility of being the subject of a profile presumably has some (expected) value, and this can be outweighed by countervailing (expected) values. This illustrates general concerns with the moral encroachment response. Firstly, the moral encroachment view characterises epistemic responsibility as a matter of weighing up the risks and rewards of profiling in each case, while it may seem preferable to adopt a *policy* of refusing to profile across the board, viewing it as a generally undesirable cognitive practice rather than simply a generic instance of weighing expected utilities. Secondly, the moral encroachment view can have the result that we are *rationally compelled* (and, indeed, *morally compelled*) to engage in profiling where the risks of not doing so are high enough, when we might prefer to say that an agent who does have a general policy against profiling is commendable rather than open to criticism.

Of course, profiling is problematic in large part not because any individual instance of profiling is terribly harmful, but because, when iterated, it creates broader oppressive patterns. For example, Andreas Mogensen draws an analogy with Derek Parfit's famous 'harmless torturer' case, where turning a single dial on a torture machine makes no discernible difference to the victim, but where repeated turns of the dial will leave them in excruciating pain. Mogensen writes:

> Imagine a very large group of torturers, each of whom turns the dial just once, resulting eventually in terrible suffering for the victim. We're inclined to say that the torturers have acted wrongly. But since each through her action seems to make no difference to the level of discomfort suffered by the victim, it seems we must appeal to the cumulative effect of these actions in order to explain why each torturer has acted impermissibly. [45][p. 466](emphasis added)

Perhaps the cumulative effect of profiling is enough to render each instance of profiling impermissibly harmful by the lights of the moral encroachment view? There are reasons to resist this diagnosis. Firstly, it is simply unclear whether the eventual badness of a cumulative pattern really extends backwards to condemn individual contributors to that pattern. For example, suppose someone buys a single-use coffee cup.

[8] Indeed, depending on the relative details of the view in question, it might be that the standards for profiling are *lowered* below those that would usually govern belief-formation.

They are making a miniscule contribution to a global system in which huge quantities of rubbish needlessly finds its way into landfills and waterways, harming future generations. But it is not obvious that the existence of this broader pattern makes the individual act substantially morally *worse* than it would have been had there not been this pattern, rather than it just being deeply regrettable that individually negligible actions can aggregate in such a way. But a more basic worry is the following: if we are careful *only* to profile when the rewards are high enough, then the cumulative harm may not be particularly great for certain individuals who are being profiled. For, as the rewards of profiling increase, so does the rarity of the cases. There will be some point—on the cost/benefit analysis—where profiling is the optimal option, even once the risk of contributing to an oppressive pattern is considered. The moral encroachment view, depending on how the risk and reward details are filled in, could license a rather permissive attitude towards profiling.

1.3 Suspension of Judgement

A second view, due to Ross [53], rejects profiling in a different fashion. Ross argues that even if we accept a traditional 'purist' theory of epistemic rationality, we still ought to *suspend judgement* instead of relying on group-level generalisations about the prevalence of certain characteristics—viz. intelligence, virtue and vice—when drawing inferences about individuals. Suspending judgement is compatible with supposing—as might a traditional view about epistemic rationality—that profiling beliefs *would be* epistemically rational if formed. This is because, as several authors have suggested, suspended judgement can be justified by non-epistemic factors, even in cases where forming a belief rather than suspending would be epistemically justified. For example, one might be justified in suspending judgement because of the pragmatic fact that you will soon receive utterly decisive evidence (even if your current evidence is already fairly strong)[9]; because the norms of close interpersonal relationships might involve giving certain people the benefit of the doubt (even if some evidence suggests that they have erred in some way)[10]; or because moral requirements stemming from certain roles (e.g. qua juror or interview panelist) might require you resist forming certain judgements until you have utterly decisive evidence[11].

On Ross's account, we ought to institute a *policy* of suspending judgement in profiling cases. This is partly because, as moral encroachment theorists concur, we should have a concern for the risks involved in acting on such beliefs. But he also argues that there are non-contingent reasons to eschew profiling, by drawing from the social egalitarian tradition in political philosophy. Social (or relational) egalitarians argue that over and above distributive equality, it is important that societies eschew hierarchical structures and modes of relating that impedes *seeing and treating one's fellow citizens as equals.* Moreover, it seems important to the spirit of this programme that

[9] Raz [51], Schroeder [57], and McGrath [44] discuss such cases.

[10] For example, see Keller [37] or Stroud [60] for relevant discussion.

[11] Ross [53][p.819] examines such a case.

social equality is not a hostage to fortune, dependent on some contingent investigation happening to validate certain empirical claims about equality[12]. Ross argues that the ideals of social equality are naturally interpreted as ranging over our cognitive lives, viz. the attitudes and beliefs we have concerning others. A primary claim he makes is that there is an intrinsic disvalue in cognitive attitudes that violate the requirements for social egalitarian relations. Regarding profiling beliefs, he writes:

> The beliefs that result from demographic profiling undermine the attitudinal requirements of social equality. By harbouring antecedent beliefs about the esteem-relevant characteristics of our fellows before an individual has had the opportunity to personally distinguish themselves in one way or the other, we are not providing those whom we encounter with a level playing field. [53][p. 816]

The suspension of judgement view has two main advantages. Firstly, it is not downstream of a controversial theory of epistemic rationality; it is compatible with both traditional and encroached theories. This is because one can maintain the idea that beliefs *once formed* should only be assessed against epistemic norms, remaining agnostic on whether the relevant epistemic norms are encroached by non-intellectual factors or not. Rather, according to Ross and the growing literature from which he draws, *suspension of judgement is a cognitive act* with its own norms, norms that are separable from those of belief. And secondly, by drawing on the moral importance of social equality, the suspension-based response to profiling offers a *non-contingent* reason to reject profiling—rather than have it that the reasons against profiling be entirely contingent in the manner of the moral encroachment view.

We have some concerns with this view. Ross suggests that suspending judgement is a way to ensure that one's cognitive states are in harmony with what he calls the 'attitudinal requirements of social equality'. In his terms, this means not having as one's default intellectual view that certain people are less morally or intellectually good than others. However, Ross's position only secures a rather limited argument against profiling. His position is that: prior to having individualized evidence about the individual in question, we should suspend judgement about members of certain demographic groups with respect to intellectual and moral virtue/vice. But suspending judgement, in Ross's sense, seems compatible with having a *credence*, but not judging outright, that a member of a particular group is less morally or intellectually good than a member of another group. At very least, there seems to be a serious tension between the argument that our cognitive states should conform to a social egalitarian standard while allowing that it is epistemically legitimate to have mental states such as the following:

Based on Race	"Given the ethnic backgrounds of A and B, I am much more confident that A is a criminal than B despite knowing neither of them."

[12] Phillips [50]

The worry is that demographic evidence poses a much broader challenge to social-egalitarian ideals than Ross acknowledges. Merely suspending judgement about individuals does not deal with socially inegalitarian levels of confidence. And levels of confidence seem like just the sort of thing that we typically rely upon, in lieu of an outright judgement, to justify and motivate action. For example, if you have statistical evidence about a lottery proposition like 'I will lose this ten-million ticket lottery', one view in epistemology is that this simply rationalizes a high credence that you will not win the lottery rather than an outright belief. Yet, nevertheless, relying on this credence is exactly what you should do when making practical plans—like deciding *not* to make an offer for an expensive mansion, given your paltry academic salary. So, if we are to judge our response to the profiling problem against a social-egalitarian standard, then it is not clear that Ross's suspending judgement view constitutes a fully satisfying response. This worry also extends to the moral encroachment view, insofar as it only discusses the moral encroachment of belief. Precisely this objection has been pressed by Jackson and Fritz and Jackson [19], who argue that accepting a morally encroached theory of belief might require either accepting a puzzling asymmetry between the rational status of full belief and credence or accepting a morally encroached theory of credence, something which has yet to be defended and comes with its own independent costs.

The second issue with Ross's view concerns whether profiling evidence can ever be relevant. Ross argues that we should suspend judgement "*until we have individualized evidence*" about the person we are considering. This is sensible if you think it implausible to hold that we can never form beliefs about the moral or intellectual characteristics of others. Waiting until we have individualized evidence is supposed to be way of fulfilling the requirement that our cognitive *default* is to treat members of different groups as social equals, while acknowledging that sometimes we really do acquire evidence that entitles us to depart from this egalitarian default. However, the question arises: can we integrate demographic evidence into our judgement *after* we have received some individualized evidence?

There are two possibilities:

WAIT: After receiving some individualized evidence, the statistical evidence then becomes relevant.

EXCLUDE: The statistical evidence ought to be excluded, even after receiving some individualized evidence.

If we take the **WAIT** option, i.e. incorporate profile evidence into our decision-making after receiving any individualized evidence, then we might think that there is not really a 'level playing field' between different demographics with respect to the attitudes we are rationally licensed to form. An example will make this clear.

Suppose you see someone drink orange juice in a context where the custom is to drink alcohol, like the evening party of a wedding in many European countries. This is individualized evidence (*they* are drinking the juice). This individualized evidence might (only very slightly) support the proposition that 'the person is or was an addict' (because people in recovery will prefer non-alcoholic drinks in contexts where most would prefer alcoholic drinks). Of course, this individualized evidence by itself hardly supports the proposition that someone is or was an alcoholic, because other possibilities are more likely (e.g. they are driving; they are pregnant; they are health conscious; they don't want to embarrass themselves on the dance floor). However, suppose there existed an accurate statistical individualized about members of their demographic which stated that they have a considerable tendency to alcoholism. The **WAIT** option would have it that, since we have some individualized evidence, we can *now* draw on the statistical generalization when forming judgements about this person. It hardly seems to conform to the 'level playing field' that Ross advocates for, if seeing someone drink some orange juice entitles us to be confident they are an alcoholic if they belong to group Y but not to group X. Rather, the upshot would be that some people are given unequal treatment in certain situations, in light of background statistical individualized about the group of which they are a member.

However, if we instead take the **EXCLUDE** option, to say that demographic evidence must be excluded even *after* receiving individualized evidence, then Ross's argument ends up being much more radical than it first appears. For, the idea would not only be to suspend judgement in some scenarios but rather to permanently exclude a certain type of evidence. This represents a much more radical view whereby we consciously reject evidence that we accept to have relevance for the question at hand. Our worry is that as it stands this is an unmotivated violation of the principle of total evidence. According to this principle one's credences should take into account all of the information one has available. The principle is supported by powerful conceptual and mathematical arguments (Carnap [13], Blackwell [6], Good [23], Hosiasson [33]). If one simply violates it without explanation as to what about one's epistemic circumstances license this, on the surface it is difficult to maintain that one is rational in so doing.

In the rest of this paper, we will show how both of the worries raised against Ross [53] can be resolved in the framework of bounded Bayesianism. We will show how a rational agent can rationally resist changing their level of confidence (credence) after receiving demographic evidence and why the **EXCLUDE** option can be consistent with the principle of total evidence. Here is a summary of the overall argument:

1. Updating (changing) credence after receiving evidence is a cognitive *act* for a bounded agent and its rationality should be assessed by the norms of instrumental rationality.

2. If the cost of updating credence upon a piece of evidence is non-negligible compared to utilities involved in the decision that an agent is about to make based on that credence, then updating upon the evidence is rational *only if* they deem it sufficiently likely that updating on that evidence will change their mind about the decision they face. The threshold of being *sufficiently likely* is determined by the disutility that they assign to doing the cognitive act of updating and the utilities that they assign to the outcomes of the decision that they are about to make based on their credence.

3. Demographic evidence about an individual is only useful when accompanied by a *statistical study about their demographics*. In a stereotype-driven society, some statistical studies about demographics are *noisy*. Bayesian updating upon a noisy statistical study is cognitively taxing and only causes minimal shift in the credence [2]. On the other hand, the disutility of updating on demographic evidence is high for a social egalitarian given the intrinsic disvalue of maintaining different mental attitudes towards random members of different demographics for them. So, a bounded Bayesian with social egalitarian tendencies typically do not deem it *sufficiently likely* that updating on stereotypical statistical demographic evidence will change their mind about the decision they face.

4. Therefore, it is typically instrumentally rational for a bounded Bayesian with social egalitarian tendencies to refuse to update their credence based on stereotypical statistical demographic evidence.

1.4 Bayesian Updating as a Cognitive Act for a Bounded Agent

We have seen that suspension of judgment is considered a cognitive act and hence is not subject to epistemic norms. So, if an agent decides to suspend judgment about an individual even after receiving demographic statistical evidence, they are not violating any epistemic norm. The question is whether there is a counterpart in formal epistemology for this move, i.e. suspension of judgement about a proposition even after receiving relevant evidence. In this section, I will argue that the counterpart for a non-ideal Bayesian is *refusing to update their credence for that proposition after learning relevant evidence*. I show why updating credence upon a piece of evidence is a cognitive act for a non-ideal (bounded) Bayesian and hence not subject to epistemic norms of rationality. I will then introduce a necessary condition for *practical rationality* of this cognitive act.

Bayesians treat credences as the outputs of a probability function (P) which is a function that gets its input from a space[13] of possible propositions and produces outputs between 0 and 1 and respects certain rules, called Kolmogorov's laws[14]. The classic approach is to assume that a Bayesian has immediate access to a prior credence towards every possible proposition X (let's call that $P^0(X)$) before starting an enquiry and when they learn a proposition like Y during the enquiry two things happen instantly and without any (or with a negligible) cognitive cost for the agent:

- Their posterior credence towards Y becomes 1: $P^1(Y) = 1$
- Their posterior credence towards other propositions (like X) is calculated by

[13] That space is called Borel space and should have certain characteristics, the details of which is unnecessary for the purpose of this paper.

[14] Kolmogorov's laws are (i) the probability of an event cannot be negative, (ii) the probability of the reference set is 1, and (iii) the probability of the disjunction of mutually exclusive vents is the sum of their individual probabilities.

conditionalization on Y, which has the following formula: $P^1(X) = P^0(X|Y) = \frac{P^0(X\&Y)}{P^0(Y)}$

It is not hard to admit that this picture is too simplistic for a non-ideal agent with limited memory, computational capacity and attention scope. A bounded agent does not have immediate and cost-negligible access to their prior credence towards every possible proposition before starting an enquiry. That is because their attention scope is limited and cannot accommodate every salient relevant proposition[15]. Neither can they immediately and without any cognitive cost conditionalize and update their credence function on any proposition that they learn during their enquiry. That is mainly because maintaining a prior judgment about how pairs of propositions are probabilistically correlated is cognitively taxing. One may be able to instantly judge how likely it is that they miss the bus tomorrow ($P^0(X)$) and how likely it is that tomorrow is rainy ($P^0(Y)$), but it may take them a while and some effort to answer the question 'how likely it is that they miss the bus if the weather is rainy' ($P^0(X|Y)$). That is because that question invites consulting their background knowledge of the causal structure of the world and/or their record of bus-catching throughout the year and this consultation is time-consuming and cognitively taxing[16].

So, updating one's credence function after a learning episode does not happen instantly or automatically. That is because it needs estimating the degree of probabilistic dependence between the evidence proposition and the original proposition, and such an estimation needs consulting one's background knowledge. In fact, the cost is more than just consulting one's background knowledge. We are bounded agents and we cannot maintain a credence function over every logically possible combination of propositions. So, it is natural to assume that we have a very coarse-grained space of possibilities and, as new evidence becomes salient to us, we include those new evidence propositions in the algebra of our credence function by fine-graining that algebra. We need to do that in order to calculate our conditional credence and be able to update our credence. Such fine-graining not only needs consulting our background knowledge to estimate the degree of probabilistic dependence between the new proposition and old ones, but also needs observing some mathematical rules, known as probabilism, to make sure the credence function over the new more-fine-grained algebra is still a probability function. These cognitive tasks, i.e., estimation and mathematical calculation, need at least some level of volition and intention. Therefore, updating is counted as an *intentional action* for us as bounded agents and is subject to the norms of practical rationality rather than epistemic rationality.

The important question is whether this cognitive act can be practically *irrational* given that Good [23] proved that cost-free conditionalizition always maximizes the *expected utility of a decision* and it is also proved that conditionalization always improves *epistemic utility* (accuracy) of the credence function [48], [26]. The answer is

[15] Moreover, which propositions become salient for an agent depends on their background knowledge which changes during an enquiry.

[16] A subjective Bayesian may object by saying that we do not need to consult objective facts when forming priors, as priors do not need to be based on objective facts, but most theorists think there are at least some constraints on permissible conditional priors [65],[56].

yes, if we think of ourselves as decision-makers, who are about to make a decision under uncertainty (henceforth referred to as the 'macro decision'), and we assume that (i) the *mental cost of updating* is non-negligible compared to the *utility of making a better macro-decision* and (ii) the *epistemic utility of having a more accurate credence function* is negligible compared to the utility of making a better macro-decision. It is important to note that this mental cost involves both the *procedural* cost and the *non-procedural* cost. By procedural cost, I mean the cognitive and hedonic tax that a bounded agent has to pay in order to update their credence function, and by non-procedural cost, I mean the tax they have to pay by upholding the resulting updated credence, which is a mental attitude that may go against their *moral values* or personal preferences.

In the following, I argue that whether updating the credence function with respect to a piece of evidence is rational for a bounded agent depends on two factors: (i) how confident the agent is that the evidence is going to change their mind about the macro-decision and (ii) how much they disvalue paying the mental cost of updating. Suppose an agent faces the following macro-decision. They are deciding between two actions A_1 and A_2 and there are two possible states: X and $\neg X$. Suppose their prior credence for X is below 0.5 and they are about to choose A_2 (hence the little arrow[17] besides A_2):

	X	$\neg X$
A_1	10	0
→ A_2	0	10

Macro-decision

Also suppose before they get to choose A_2, they learn proposition E. Unfortunately, they do not have cost-free access to their credence towards X conditional on E, ($P^0(X|E)$). So, they have to decide whether to pay the cost and update their credence for X by conditionalizing on E before making the macro-decision. They know that if $P^0(X|E)$ is above 0.5, they will choose A_1 over A_2 in the macro decision, but since they do not have access to $P^0(X|E)$, they do not know what their posterior credence after updating with respect to E will be and whether it will be above 0.5. Therefore, they do not know which option they will choose in the macro-decision after updating upon E and what their expected utility with respect to that decision will be.

So, the decision-relevant proposition that they are uncertain about is whether the conditional credence $P^0(X|E)$ is above 0.5. To put it informally, they are uncertain whether whether updating on E would change their choice in the macro-decision. If it would and they update their credence then they will choose A_1 in the macro-decision

[17] Henceforth in the decision tables, I use the little arrow to show the initial inclination of the agent i.e. the decision they would make based on their prior credence.

but they will undergo the cognitive labour of updating. But if updating on E would not change their choice in the macro-decision and they update then they keep choosing A_2 in the macro-decision and they also undergo the cognitive labour of updating. If they do not update, they keep choosing A_2 and they undergo no cognitive labour.

	E is decision-changing	**E is not decision-changing**
Update	Choosing A_1 and paying the cost of updating	Choosing A_2 and paying the cost of updating
Don't update	Choosing A_2 and paying no cost of updating	Choosing A_2 and paying no cost of updating

The utility of each outcome is determined by finding the sum of the expected value of choosing the relevant option under the relevant credence function in the macro-decision and the cost of updating. If we denote the credence function after updating with P^1 and the credence function before updating with P^0 then the utility of choosing A_1 after updating is $P^1(X) * 10 + P^1(\neg X) * 0$, the utility of choosing A_2 after updating is $P^1(X) * 0 + P^1(\neg X) * 10$, and the utility of choosing A_2 before updating is $P^0(X) * 0 + P^0(\neg X) * 10$. If we denote the mental cost of updating by c, then the following table represents the utilities of each outcome:

	$\boldsymbol{D}$: E is decision-changing	$\neg\boldsymbol{D}$: E is not decision-changing
$\boldsymbol{U}$ **:** Update on E	$P^1(X) * 10 + P^1(\neg X) * 0 - c$	$P^1(X) * 0 + P^1(\neg X) * 10 - c$
$\neg\boldsymbol{U}$ **:** Don't update on E	$P^0(X) * 0 + P^0(\neg X) * 10$	$P^0(X) * 0 + P^0(\neg X) * 10$

Table 1.1: Micro-decision of updating on the evidence E

Using this table, we can calculate the expected utility of updating and not updating as cognitive acts. It can be proved that[18]:

$$EU(U) > EU(\neg U) \textit{ only if } P^0(D) > \frac{c}{20}$$

This means that updating on a piece of evidence is instrumentally rational for a bounded agent *only if* her prior credence towards that piece of evidence being decision-changing is beyond a threshold determined by the cost of the cognitive act of updating and the utilities involved in the macro decision. This principle can be easily generalized for a macro-decision with the following utilities:

	X	$\neg X$
A_1	H_1	L_2
→ A_2	L_1	H_2

Generalized macro-decision[19]

If a bounded agent is about to make the above macro-decision, and their initial

[18] Please see the appendix for the proof.

credence towards X is low enough[20] to make them inclined to choose A_2 and they receive a piece of evidence E relevant to X, then they should update their credence on E *only if* their prior confidence that updating on E will change their mind about the macro decision (proposition D) is above a threshold:

$$EU(U) > EU(\neg U) \; only \; if P^0(D) > \frac{c}{(H_1 - L_1) + (H_2 - L_2)} \tag{1.1}$$

The next question is whether judging the above inequality i.e. whether one's prior credence towards D (E being decision-changing) is above that threshold is less costly than actually updating on E. I believe there are many contexts that such judgment can be made with negligible cognitive cost. Those include contexts in which (i) we have reasons to think that it is extremely unlikely that updating on E will change our mind about the macro-decision (the LHS of the inequality is too low) and (ii) the mental cost of updating is too high.

But what kind of reason can make one think that it is extremely unlikely that updating on a piece of evidence can change their mind about the decision that they are about to make? There are two factors that affect whether updating on a piece of evidence can change one's mind about the decision they are about to make:

1. How far the threshold credence for changing their mind about the macro-decision is from their prior credence.
2. How strong the evidence is generally considered to be.

If the threshold credence is very far from the prior credence and the evidence is weak, then the agent can justifiably believe that it is extremely unlikely (even impossible) that updating on that piece of evidence will change their mind about the decision they face.

Suppose you are new to a neighborhood and you are deciding on how to behave to new people that you meet. You are undecided between two modes of behaviour: being very nice and being neutral. You think being very nice to untrustworthy people increases the chance of being exploited by them and being neutral to them does not affect that chance. On the other hand, being very nice to trustworthy people increases the chance of making new friends and being neutral to them does not affect that chance. So, if you meet a new person like Andrew in the neighbourhood, you face the following decision under uncertainty:

	T : Andrew is trustworthy	**$\neg T$: Andrew is not trustworthy**
B : Be very nice to Andrew	Increasing the chance of making a new friend	Increasing the chance of being exploited by him
$\neg B$: Don't be very nice to Andrew	Status quo	Status quo

Also suppose that your anxious personality is such that your fear of being exploited by untrustworthy people is much greater than the joy you experience when making

[20] This means $P(X) < \dfrac{H_2 - L_2}{(H_1 - L_1) + (H_2 - L_2)}$

new friends:

		$\boldsymbol{T}$	$\neg\boldsymbol{T}$
	$\boldsymbol{B}$	$+10$	-100
→	$\neg\boldsymbol{B}$	0	0

Deciding on how to treat new people for an overly anxious person

It is easy to show that in this scenario, you need to be at least 90.9% sure that Andrew is trustworthy in order to treat him very nicely. You are an optimist and have a prior credence of 70% that any random person is trustworthy. But of course, you are still not convinced to treat him very nicely and you need at least 20.9% confidence boost in order to be convinced. Andrew is a white man and you know in this new neighborhood the rate of crime among white men is only 15%. Should you update your credence about Andrew being trustworthy (T) upon this piece of evidence?

The answer is no. That is because that piece of evidence will not generate the sufficient confidence boost that you would need for changing your mind about how to treat Andrew. Even if you assume committing/not committing a crime to be a perfect indicator of being untrustworthy/trustworthy[21] and the statistical study about the rate of crime in that neighborhood is perfect in terms of the size of the sample and the absence of noise, your confidence towards Andrew being trustworthy after updating upon that piece of demographic evidence will be at most around 85%, i.e. the *maximum confidence boost* that it can afford is 15%. However, you needed at least 20.9% confidence boost to treat him very nicely[22]. So, your confidence towards this piece of evidence being decision-changing is *zero* and the inequality (1) does not hold for it. So, instrumental rationality dictates that, as a bounded agent, you should ignore this piece of evidence altogether and not update your credence based on it.

The next question is what if we have a lower threshold for changing our mind about the decision we are about to make? In the above example, what if your personality is only moderately anxious? If you attribute -40 utiles to the outcome of increasing the chance of being exploited by Andrew, you only need to be 80% sure that Andrew is trustworthy to treat him very nicely.

[21] This is a natural assumption if we think of 'crime' as a felony (major crime) OR repeating misdemeanour. If someone commits a major crime or a repeating misdemeanour then they are untrustworthy and if someone does not commit a major crime or a repeating misdemeanour, then they are trustworthy.

[22] I would like to emphasize that this is the *maximum* credence that you can have after updating on that piece of statistical evidence i.e. the credence that you would have if the (i) statistical evidence perfectly reflects the *chance* of having a criminal record and (ii) not having a criminal record is the perfect indicator of being trustworthy. That is not the case for most cases of demographic statistical evidence i.e. (i) they use a sample that may not represent the population with negligible error, resulting in a *non-representative sample*, (ii) they use an indicator that does not track our sought-after trait accurately, resulting in a *noisy sample.* In the next section I will discuss, how a Bayesian should update their credence upon a possibly noisy non-representative piece of demographic statistical evidence and how it affects their micro-decision of updating.

	$\boldsymbol{T}$	$\neg\boldsymbol{T}$
$\boldsymbol{B}$	$+10$	-40
$\rightarrow$ $\neg\boldsymbol{B}$	0	0

Deciding on how to treat new people for a moderately anxious person

You have a prior credence of 70% that he is trustworthy. You also know that he is a member of a demographic group (white men) 85% of which does not have a criminal record. Does updating your credence on that fact boost your credence above 80%? How likely is it that it happens? Is it likely enough to justify the instrumental rationality of doing the cognitive act of updating for you as a bounded agent?

Obviously, there is no general yes or no answer to these questions and the answer depends on the details of the statistical evidence about that demographic group. In the next section, I will argue that in many decision-making contexts, if the statistical evidence is about a demographic group which has been subject to stereotypes in a society and you are a social egalitarian then it is instrumentally rational for you to ignore that demographic evidence.

1.5 Bayesian Updating on Demographic Evidence in a Stereotype-driven Society

Demographic evidence is only useful when accompanied by relevant statistical data. If you're considering whether someone has a certain characteristic, such as trustworthiness, and you learn that they belong to a particular demographic group, but you have no idea how prevalent that characteristic is within the group, that information is useless. The same applies when your understanding of the prevalence of that characteristic is based on word of mouth or unfounded stereotypes. What is needed is a statistical study on the prevalence of that trait within the demographic group in question.

However, there are two key problems that we face here: first, statistical studies are usually based on small samples, which may not accurately represent the large populations of demographic groups. For example, in the case of a statistical study on the prevalence of criminal records among white men in Andrew's neighborhood, one may question whether the sample truly represents the broader population of that demographic group with negligible error.

Second, we are often interested in characteristics for which no direct statistical studies exist. In such cases, we must rely on a study of a different trait within that demographic, hoping it tracks the prevalence of the trait we are interested in, with negligible error. However, there may be serious doubts about the accuracy of this tracking and whether it is affected by a systematic error. If we refer to the systematic error in that tracking as 'noise', we might be uncertain about the level of noise in a given piece of statistical evidence. For example, in Andrew's case, there is no statistical study on the prevalence of 'trustworthiness' or even '*actually* having committed a crime' within his demographic. Instead, we have a study on the prevalence of 'having a *criminal record*' within that group. We may question whether there is a systematic error in 'having a criminal record' tracking the 'actually having committed a crime', and be uncertain about the level of that noise.

In the following, I will first review Babic et al. [2]'s model of Bayesian updating on noisy statistical evidence. Then, I will argue why, in a stereotype-driven society, some demographic statistical evidence can hardly be considered decision-changing, and why a bounded Bayesian with social egalitarian tendencies is often rationally obliged to ignore such evidence and not update their credence based on it.

1.5.1 Bayesian Updating on Noisy Statistical Evidence: Cognitively Taxing and Minimally Effective

How should a Bayesian update their credence after receiving an imperfect piece of statistical evidence? As mentioned before, statistical evidence can be imperfect in two ways: 1- the sample may be small and not represent the population perfectly (non-representative sample), 2- there may be systematic errors in data collection from the sample, leading to systematic misclassification (noisy sample). Following Babic et al. [2], I will first explain how a Bayesian should update their credence based on statistical evidence derived from non-noisy data collected from a *sample* rather than the entire population. Then, I will discuss how they should account for the possibility of *noise* in their updating process, and how this consideration, while possible, is computationally taxing.

Let's start by an example: the aforementioned case of Andrew's demographics. Andrew is a random member of a well-defined demographics i.e. white men in the neighbourhood X. We know that 85% of white men in this neighbourhood do not have criminal record and we wonder how likely it is that a random member of this demographic group (like Andrew) is trustworthy. Let's first focus on how certain should we be towards Andrew having a criminal record after learning this statistics. The naive and wrong approach is to appeal to *principal principle* and demand that we should put our posterior credence equal to the rate of reported in the statistical study i.e. that we should 85% certain that a random person like Andrew has a criminal record after learning that data. Why is it wrong? Because principal principle requires the Bayesian to put their credence for an event equal to the *chance* of that event. But we do not know whether the number reported in a statistical study as the prevalence rate of a trait in a population is equal to the *chance* of having that trait in that population. The statistical studies are based on collecting data from a *sample*. The reported prevalence rate will be equal to the chance only if the sample includes the whole population (or is a perfect representation of that population) and there is no error (noise) in the data collection, i.e. if the statistical evidence is perfect. That is rarely the case. Most of statistical study is imperfect. So, we are not obliged to put our credence equal to the reported rate. What should we do then?

Statistical evidence about the proportion of a population having a characteristic Q can be treated as sampling from a *Bernouli process*[23]. As Bayesians, we maintain a prior towards the next outcome of a Bernouli process with outcomes Q and $\neg Q$ and this prior is usually taken to be the *mean of a beta distribution*[24] over all different possible chances of the outcome Q in that Bernouli process:

$$P^0(Q) = \frac{a_\theta}{a_\theta + b_\theta} \tag{1.2}$$

In this formula, a_θ and b_θ can be considered as pseudo observations of Q and non-Qs. The posterior after updating upon a sample with size n and x number of Qs is determined by the following formula:

$$P^1(Q) = \frac{a_\theta + x}{a_\theta + b_\theta + n} \tag{1.3}$$

This is called generalized Laplacian succession rule according to Huttegger [35]. Looking at the equations (2) and (3), it is not hard to see that the distance between $P^1(Q)$ and $P^0(Q)$ is an increasing function of $\frac{n}{a_\theta + b_\theta}$:

$$P^1(Q) - P^0(Q) \ = \ f(\frac{n}{a_\theta + b_\theta}) \tag{1.4}$$

That is to say, the greater the size of sample compared to the subtotal of pseudo-observations, the greater effect the statistical evidence has on our credence. For example, in the case of how to treat Andrew (a random person in the neighborhood), if your prior credence towards him having (actually) committed a crime[25] (C) is 30%, then we can assume 0.3 is the mean of a prior beta distribution over all possible chances with the following parameters:

- Psuedo-observation of having committed a crime: $a_\theta = 3$
- Psuedo-observation of not having committed a crime: $b_\theta = 7$
- Mean of the distribution: $P^0(C) = \frac{a_\theta}{a_\theta + b_\theta} = \frac{3}{10}$

[23] A Bernoulli process consists of a sequence of binary random variables that can have either a finite or infinite length.

[24] Beta distribution is conjugate to Bernouli process which implies our posterior will also be a beta distribution. A beta distribution is usually identified with two variables a_θ and b_θ.

[25] It may be argued that 'having (actually) committed a crime' is different from 'being untrustworthy' and that it should be assumed that your prior credence towards him being untrustworthy is more than him having committed a crime. As mentioned before, if we interpret 'crime' as a felony (major crime) OR repeating misdemeanour, then it is natural to assume that for most people, their prior credence towards one being trustworthy is roughly equal to their prior credence towards them having committed a crime. That is because committing a major crime (like roberry) or a repeating misdeamonour (like assault) is enough for losing trust in a person and lack of them is enough for trusting them. However, we assume 'having (actually) committed a crime' is different from 'having a criminal record'. In the following section, we will say more about this assumption and its role in our modelling.

The statistical study that we have is about the rate of having a criminal record among white men and that rate is arguably different from the rate of having actually committed a crime, depending on whether there is a bias in the judicial system against white men. Let's first assume that there is no such bias and the difference between the rate of having (actually) committed a crime and having a criminal record is negligible in that demographic group. If the statistical study has a sample size of $n = 100$, with $x = 85$ observations of not having a criminal record, then following equation (3), our posterior credence towards Andrew having committed a crime after updating on this evidence is:

$$P^1(C) = \frac{a_\theta + x}{a_\theta + b_\theta + n} = \frac{3 + 15}{10 + 100} = 0.164$$

However, if the sample size is $n = 20$ with $x = 3$ observations of having a criminal record, then following the same formula, our posterior will be $P^1(C) = \frac{3+3}{10+20} = 0.2$ which is closer to our prior credence (0.3). So, if the sample size is 100, updating your credence based on the statistical evidence lowers your credence towards Andrew having committed a crime ($P^1(C)$) and therefore being untrustworthy ($P^1(\neg T)$ below 0.2 which implies your credence towards him being trustworthy ($P^1(T)$) will be above 0.8 and you will decide to treat him very nicely rather than being neutral to him. However, if the sample size is 20, updating your credence based on the statistical evidence only lowers your credence towards Andrew having committed a crime ($P^1(C)$) and therefore being untrustworthy ($P^1(\neg T)$ to 0.2 which means your credence towards him being trustworthy ($P^1(T)$) will not go above 0.8 and you will decide to treat him neutrally.

What is important to note is that in order to update our credence upon a statistical study about the prevalence a trait in a population, which is based on a sample rather than the whole population, we need to maintain a prior distribution over different possible prevalences of that trait and decide on the shape of that distribution. That distribution is usually assumed to be a beta distribution. So, we need to decide on the value of variables a_θ and b_θ. That decision not only determines our prior about the next observation from the population having that trait (the mean of the distribution), but also the weight that we consider to our prior (pseudo-observations) compared to actual observations in the sample. Making such a decision is cognitively taxing on its own, but one may argue that its cost is negligible compared to the utilities involved in the macro-decision which we are about to make based on our credence. However, if we add the possibility of noise to the picture, it is hard to accept that the cost is negligible.

How should we update our credence if we consider the *possibility* of noise in the sample and why is it so costly? Let's first see what noise in a statistical study is and when its possibility is salient to us. Noise is usually defined as a *systematic misclassification* in the sample. When looking for the prevalence of trait Q in a population, systematic misclassification happens when there is an underlying cause for misclassifying Qs in the sample as non-Qs and/or vice versa. So, when we ask ourselves whether the statistical study is noisy, we are in fact asking ourselves whether there is an underlying cause for misclassification in the data collection. We are *not* looking for accidental clerical errors. We are looking for an underlying reason that can cause, even a small percentage of Qs (or non-Qs) in the sample getting misclassified in the other group.

In the example of the statistical study about crime rate among different demographic groups in a neighbourhood, the study claims that it reports the percentage of the members of a demographic group that committed crime. Let's take this demographic group to be white men. We need to know how the study categorized people in its sample into two groups of 'innocent' and 'guilty of crime'. Let's suppose it looked at their official *criminal record* for this categorization i.e. if a person had a criminal record, they were recorded as 'guilty of a crime' and if they did not have any criminal record, they have been put in the category of 'innocent'. Given this methodology, is the possibility of noise in this study salient to us? The answer depends on whether we can think of an underlying reason that causes white men who are innocent as guilty of a crime or vice versa? It is hard to find such a reason. So, the possibility of noise in that statistical study is probably not salient to us. What if we replace the demographics of 'white men' with 'black men' or 'immigrants'? Can we think of an underlying reason that causes innocent black men misclassified as guilty of a crime by looking at their criminal record? Yes! The bias against black men in the judicial system can cause such misclassification. So, the possibility of noise in the statistical study about crime rate among black men (or immigrants) is salient to us. It is important to note that the fact that it is salient to us does not imply that our estimation of it is high.

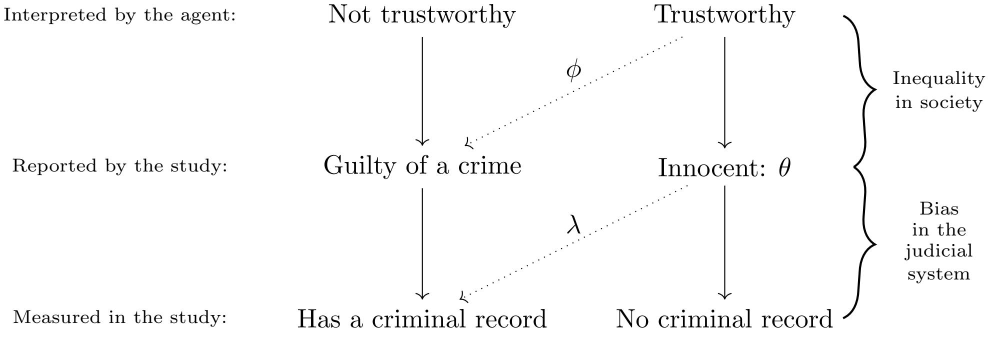
This misclassification is in fact the percentage of innocent people in the sample who were misclassified in the study as guilty of a crime because of having a criminal record. So, it is in fact the percentage of innocent people in the sample who have a criminal record (λ in the diagram). λ can be salient to us even though our estimation of it is rather low. We may estimate that only 10% of black innocent people in the sample have a criminal record (because of the bias in the judicial system)[26]. In that case, if the actual percentage of innocent people is θ, we will observe that $(1-\lambda) * \theta = 0.9 * \theta$ percent of the sample have no criminal record.

This misclassification is even more severe if we *interpret* the study as the rate of trustworthy people among a demographic. That is because there is another underlying reasons for misidentifying trustworthy people as untrustworthy by looking at whether they are actually guilty of a crime[27]. That underlying reason is the inequality in society. Some percentage of disadvantaged trustworthy good people can be pushed to doing misdemeanours (like shop-lifting or getting in a fight) repeatedly, be guilty of a crime and hence identified as untrustworthy (ϕ in the diagram). For modelling

[26] This possibility is not even salient to us in the case of the study on white men. That is because we cannot think of an underlying reason for systematic misclassification.

[27] As noted before, if we define 'crime' as 'repeated misdemeanour OR felony' then the possibility of misclassifying untrustworthy people as trustworthy by checking whether they actually committed a crime is *not salient*.

reasons[28], in this paper we assume ϕ is negligible compared to λ.



Babic et al. [2] offers a model in which a Bayesian maintains a prior beta distribution over different levels of noise (λ) in the statistical study ($P^0(\lambda)$) and then use it along with their prior beta distribution over different chances (θ) of having the characteristic in question ($P^0(\theta)$) in the updating procedure. Using $P^0(\theta)$ and $P^0(\lambda)$, they first calculate the joint posterior $P^1(\theta, \lambda)$ and then use that to calculate marginal distributions $P^1(\theta)$ and $P^1(\lambda)$. This procedure is not straightforward and needs complex statistical machinery.

They then show how positing a prior distribution over λ and hence assuming some level of noise (the mean of that distribution) in the sample, even for very small values, leads to a substantial information loss and affects how much the statistical evidence changes our credence. In their model, updating credence upon a noisy large sample can be considered equivalent to updating upon a small noise-free sample. They define the notion of *effective sample size* as the following:

> Given a prior, and a sample of size n, with misclassification rate λ, the effective sample size $n*$ is the sample which would have the same effect on the prior if $\lambda = 0$. ([2], p.19)

So, the effect of noisy statistical evidence with size n on our prior credence is measured by its effective sample size ($n*$):

$$P^1(Q) - P^0(Q) \;\; = \;\; f(\frac{n^*}{a_\theta + b_\theta})$$

Their model shows how *computationally costly* it is for a bounded agent to estimate the effective sample size of a noisy study. It also demonstrates the drastic difference between the size of a noisy sample and its effective sample size when assuming a small

[28] To use the model from Babic et al. [2], we must assume that $P^0(T)$ is probabilistically independent of P^0(misclassification). For this assumption to hold, we need to assume that there is no common cause for being untrustworthy and being misclassified as untrustworthy. One could argue that societal inequality can not only lead to trustworthy individuals being misclassified as untrustworthy (by pushing them towards minor misdemeanours) but also cause individuals to actually become untrustworthy (by pushing them towards severe personality disorders and major crimes). In such cases, our prior credence regarding a person being untrustworthy and being misclassified as untrustworthy would not be probabilistically independent. Therefore, for the purposes of this paper, we will disregard societal inequality as an underlying cause of misclassification and focus on bias within the judicial system.

level of noise, and how rapidly this difference increases as our uncertainty about the noise level grows[29]. For example if the mean of your prior distribution over different levels of noise for a statistical study is such that you consider its misclassification rate to be 30% i.e. $\lambda = 0.3$, and the study has a sample size of $100,000$ then the effective sample size for that study is only 3.7. That is to say that study exerts an effect on your prior credence equivalent to the effect of a noise-free study with a sample as small as ≈ 4 cases.

Sample size (n)	Equivalent noise-free sample size (n^*)
10	1.4
100	3.2
$1,000$	3.6
$10,000$	3.7
$100,000$	3.7
$n \to \infty$	3.8

Actual and noise-free sample sizes for $P^0(\theta) = 0.7$ and $P^0(\lambda) = 0.3$ [2]

In Andrew's example, if Andrew was a black man and hence the possibility of noise in the statistical study about his demographic group was salient to you and the mean of your prior beta distribution over misclassification rate is 0.3, then even if the study is based on a sample as big as $100,000$ cases, then your posterior towards him being untrustworthy after updating on that evidence will be $\frac{3+0.555}{10+3.7} = 0.259$. This means that updating only changes your credence by 4%. This implies that, in this scenario, updating your credence based on demographic evidence is not decision-changing, and doing it will be the waste of cognitive resources.

The important question is whether you could have predicted this without going through the updating procedure. The answer is yes if you consider the fact that the possibility of noise in the statistical evidence is salient to you. Given that the possibility of noise makes the change of credence after updating very minimal, it also makes the possibility of updating being decision-changing extremely unlikely ($P^0(D) \approx 0$). So, it can be said that *typically* when a bounded agent assigns a *non-negligible*[30] mental cost to the cognitive act of updating (c), the RHS of the conditional sentence (1) does not hold i.e. $P^0(D) \ngtr \dfrac{c}{(H_1 - L_1) + (H_2 - H_2)}$ and therefore $EU(U) \ngtr EU(\neg U)$ i.e. the expected utility of updating is not greater than not updating and hence updating is instrumentally irrational. This in turn implies that, typically, when the

[29] Please check table 3 on page 20 of Babic et al. [2] to see this point in numbers.

[30] Non-negligible compared to the utilities involved in the macro decision i.e. $(H_1 - L_1) + (H_2 - H_2)$. This implies that $\dfrac{c}{(H_1 - L_1) + (H_2 - H_2)} \not\approx 0$.

possibility of noise in a statistical study about a demographic is salient, even for very small estimation of noise, updating on demographic evidence is instrumentally irrational. We will show this in more detail with an example in the next section.

Babic et al. [2]'s model of updating on noisy statistical evidence offers two important insights:

1. Bayesian updating on noisy statistical evidence is computationally challenging, making the cognitive cost of performing it significant.
2. Bayesian updating on noisy statistical evidence *typically* results in minimal changes to our prior credence, making it unlikely to be decision-changing.

As it was explained, these two insights together implies that *typically* updating on noisy statistical evidence is instrumentally irrational for a bounded agent. In the following section, we explain how this conclusion can be used to meet the objections raised against Ross [53] in section 3.

1.5.2 Social Egalitarianism and Having Faith in Random Individuals

We have seen that, as bounded agents, whether we decide to pay the mental cost of updating our credence on a piece of demographic evidence depends on the macro-decision that we are about to make based on that credence, the mental cost that we assign to the cognitive act of updating and how likely we think it is that updating our credence will change our mind about the macro-decision. On the other hand, we saw that demographic evidence if paired with *noisy* statistical evidence has a very minimal potential effect on one's credence and is typically extremely unlikely to be decision-changing. Let's see this conclusion and its implication for demographic profiling in a concrete example.

Going back to the setting of the previous example with some modifications: you are a *slightly* anxious person, new to a neighbourhood, deciding how to treat two strangers: Andrew (a random white man) and Mohammed (a random immigrant). You are deciding between two modes of behaviour: being very nice (B) and being neutral ($\neg B$) and you are uncertain whether they are trustworthy. Given your anxious tendencies you assign more absolute value to the outcome of increasing the chance of being exploited compared to the outcome of increasing the chance of making new friends. Given this distribution of utilities you need to be more than 66% certain that someone is trustworthy in order to treat them very nicely. You are an optimist and have a prior credence of above 66% that a random person is trustworthy. So your initial decision is to behave both Andrew and Mohammed very nicely:

	$\boldsymbol{T}$	$\neg\boldsymbol{T}$
→ $\boldsymbol{B}$	+10	−20
$\neg\boldsymbol{B}$	0	0

Deciding on how to treat new people for a slightly anxious person

You know that the rate of crime among white men in this new neighborhood is only 15% while the rate of crime among immigrants is 45%. In this setting and if we assume that committing crime is a perfect indicator of being trustworthy[31] it is obvious that you should *not* update your credence towards trustworthiness of Andrew based on the statistical about crime rate in his demographics. That is because no matter where in the range above 66% your prior credence is and what weight you assign to your pseudo-observations compared to the actual observations of statistical study, updating on that piece of evidence is going to increase your credence towards Andrew being trustworthy. So your posterior will also be above 66% and hence updating is not going to change your mind about how to behave him, i.e. it is not decision-changing. So, since updating will be a waste of cognitive resources and as an instrumentally rational agent, you decide not to update your credence and to attach to your initial decision of behaving Andrew nicely.

How about Mohammed? Can you ignore the statistical evidence about the rate of crime in his demographic group as an instrumentally rational agent? The answer is not as easy as the case of Andrew. If you have a prior credence of 70% towards trustworthiness of Mohammed and take it to be equivalent to 7 pseudo-obsevations and the study is non-noisy and its sample has the size of 100 individuals, then your posterior towards his trustworthiness after updating will be $\frac{7+55}{10+100} = 56\%$ which is below 66% and you will decide to behave him neutrally (not very nicely). So, updating on this piece of statistical evidence can be decision-changing.

The difficult question is how certain you are that it is going to be decision-changing given that you haven't specified a prior distribution? In other words what is your $P^0(D)$? Another important question is what the mental cost (c) that you assign to the cognitive act of updatimg is. Answers to these questions are important because you need to know whether $P^0(D) > \frac{c}{30}$ in order to decide whether updating on this piece of evidence is rational[32] for you based on the conditional sentence (1). I will argue that in this scenario, for a typical agent $P^0(D)$ is too low and c is too high for $P^0(D) > \frac{c}{30}$ to hold and hence typically updating on this statistical evidence is not rational.

Why $P^0(D)$ is typically too low in this scenario? That is because statistical evidence about the crime rate among immigrants can be *noisy*. There are innocent

[31] We have explained that this assumption is rather natural if we interpret 'crime' as 'repeated misdemeanour OR felony'.

[32] $\frac{c}{(H_1 - L_1) + (H_2 - L_2)} = \frac{c}{(0 - (-10)) + (0 - (-20))} = \frac{c}{30}$

immigrants who are *systematically* misclassified as criminals because of the systematic bias in the judicial system. Stereotypes against immigrants and the political pressure for anti-immigrant policies can cause bias in the courts' judgments. We have seen that according to Babic et al. [2], if we consider a statistical study noisy, even for very small estimation of noise, the effect of updating on that statistical study on our credence will be minimal. That does not mean that updating on a noisy statistical study is *never* decision-changing. Of course, if our prior is close enough to 66%, even a minimal decrease of confidence can switch our decision, but a typical person who is undecided on the exact value of their prior credence considers that scenario impossible or very unlikely.

Why c (the mental cost of updating) is typically too high in this scenario? That is because as explained in the previous section, not only Bayesian updating on statistical evidence demands choosing a prior beta distribution which can be cognitively taxing for a bounded agent, but also updating on *noisy* statistical evidence demands calculating the effective size of the sample (n*) which is computationally taxing. Given that statistical evidence about crime rate among immigrants can be considered noisy given the systematic bias in the judicial system, updating on it will be very costly.

Moreover, as discussed in section 4, the mental cost of updating includes both procedural and non-procedural costs. An example of the non-procedural mental cost of updating is the *moral cost* of maintaining a high credence that a random immigrant is untrustworthy or maintaining unequal credences that a random immigrant and a random non-immigrant are trustworthy. A social egalitarian assigns a high disvalue to maintaining unequal mental attitudes (having unequal credences) towards random members of different demographic groups being trustworthy. So, for a social egalitarian c is even higher. It is totally imaginable that a radical social egalitarian assigns a mental cost as big as the subtotal of the value of increasing the chance of making new friends (10) and the disvalue of increasing the chance of getting exploited (20) i.e. 30 utiles to maintaining unequal credences towards trustworthiness of random members of demographic groups. In such a scenario, $\frac{c}{30} = 1$ and the inequality $P^0(D) > \frac{c}{30}$ is trivially false which implies that for a bounded agent with such level of commitment to social egalitarianism, updating on statistical demographic evidence is never instrumentally rational! No matter whether they consider it noisy or not.

Nevertheless, as mentioned, one does not need to be a radical social egalitarian to be justified in ignoring demographic evidence about Mohammed from the perspective of instrumental rationality. As long as they are justified to consider the statistical study about the crime rate among immigrants noisy, they are justified to (i) have a low credence towards it being decision-changing ($P^0(D)$),(ii) assign a high procedural mental cost to updating (c), (iii) believe $P^0(D) > \frac{c}{30}$ does not hold and hence, according to the conditional sentence (1), instrumentally justified to ignore Mohammed's demographics.

This conclusion can be generalized to any statistical study, the data of which is based on *human judgment about a non-observable trait* and confirms a well-known stereotype. I will refer to these studies as stereotypical statistical studies. It can be a study about what percentage of women *performed* above average in a high-stake job, a statistical study about what percentage of men *behaved aggressively* in an experiment, what percentage of Muslims in a society *had radical tendencies*, or what percentage of people of colour in a city were convicted as guilty for a crime in the judicial system. As long as the subjective judgment of the experimenter or another person played a role in the data collection and the result of the study confirms a well-known stereotype, the possibility of noise, i.e., systematic misclassification, in the study is salient.

These studies stand in contrast to those that look for the prevalence of an observable objective trait, like having a physical disease or passing a standardized test, in a population. That is because when subjective human judgment plays a role in the data collection, it can be easily affected by well-known stereotypes about different demographics. Stereotypes like: ”women do not perform well in high-stake jobs,” ”men are aggressive,” ”Muslims have a tendency towards radicalism.” Since the possibility of noise is salient in such studies, a bounded Bayesian is typically instrumentally justified in refusing to update their credence based on the results of them[33].

This result is even more generalizable (more typically held) among social egalitarians who assign more mental cost to updating their credence upon such studies. It is also very generalizable in scenarios in which we receive demographic statistical evidence *after* receiving individualized evidence. That is because in those cases our credence before updating on statistical evidence is more robust, as it based on some individualized evidence and is not easily affected by statistical evidence. That brings us to the worry that we raised against Ross [53] account i.e. whether demographic evidence statistical evidence can ever be relevant given that updating upon such evidence even after receiving individualized evidence (the option **WAIT**) will lead to a non-egalitarian mental attitude (credence function as a non-level playing ground) and excluding such evidence altogether (the option **EXCLUDE**) will be instrumentally irrational as it goes against the principle of total evidence. We have shown the classical formulation of the principle of total evidence, i.e. that the instrumental rationality dictates updating credence function after receiving a piece of evidence and before making a decision based on that credence, is meant for an ideal Bayesian who updates their credence function automatically and without cognitive cost. We also show that if model us as bounded Bayesians and take the mental cost of updating into account then the option of **EXCLUDE** *can be* instrumentally rational and that this option is *typically* rational for *stereotypical* statistical evidence.

Finally, we would like to give a name to this attitude of *ignoring an individual's demographics* in forming credences when the associated statistical study is stereotypical. Buchak [11] defines having *faith* in a proposition as “terminating one's search for

[33] One may argue that this result is in conflict with what Bovens [9] proves about how using statistical data about underrepresented groups in a shortlisting procedure maximizes expected utility (and ignoring it does not). That is not true. My conclusion here, which is based on Babic et al. [2]'s model, is only relevant to statistical studies which measure the *prevalence* of a *binary* trait among a demographic group and *not* those which measure the *distribution* of a *gradable* trait among a demographic group.

further evidence and acting on the supposition" that the proposition is true. This definition is meant for an attentionally and computationally unbounded Bayesian whose credence function gets updated automatically upon the receipt of a piece of evidence and updating is not a matter of decision for them. For such an agent 'acting on the supposition that a proposition is true' requires terminating the *search* for further evidence. For a bounded agent, however, updating is a matter of decision. So, a bounded agent not only can act on the supposition that a proposition is true by terminating to search for evidence, but also they can act on that supposition by *refusing to update* their credence on a piece of evidence that they have already received. So, for a bounded agent refusing to update their credence towards a proposition is also counted as having *faith* in that proposition in Buchak [11]'s account. In the example of Mohammed, refusing to update your credence towards his trustworthiness based on his demographic evidence is in fact having in faith in him and his trustworthiness.

1.6 Conclusion

We have shown that a Bayesian who (i) is aware of their boundedness and (ii) has social egalitarian tendencies (a good type of Bayesian!) can resist demographic statistical evidence affecting their credences in a wide range of cases without violating norms of epistemic or instrumental rationality. This possibility is granted by the fact that such an agent treats *updating* as a costly action that should be done only if it is consequential for their future decisions. A wide range of statistical demographic studies about the prevalence of a trait, especially those whose data collection is based on human holistic judgments, are noisy in the sense that there is some level of systematic misclassification in their data collection. This noise undermines the effect that these studies may have on a Bayesian's credence and thereby on their future decisions. That is why a bounded Bayesian with social egalitarian tendencies, can rationally refuse to pay the cognitive and moral cost of updating their credence on such statistical studies. This refusal to engage with some demographic statistical evidence can be interpreted as having faith in the members of that demographics.

This approach mirrors the suspension of *belief* that Ross [53] proposes, arguing for its rationality as a socially egalitarian response to demographic profiling. However, unlike the suspension of belief—whose rationality is completely insensitive to the utilities involved in decision-making and therefore independent of the context—the rationality of refusing to update credence is sensitive to these utilities, making it more contingent and context-dependent. Nevertheless, we have shown that this dependency can be rationally ignored in many cases where the demographic evidence is stereotypical.

1.7 Appendix

Given that $P^1(X) = P^0(X|E)$ and $P^1(\neg X) = P^0(\neg X|E)$, Table 1 can be simplified as:

	$\boldsymbol{D}$: E is decision-changing	$\neg\boldsymbol{D}$: E is not decision-changing
$\boldsymbol{U}$ **:** Update on E	$10P^0(X|E) - c$	$10P^0(\neg X|E) - c$
$\neg\boldsymbol{U}$ **:** Don't update on E	$10P^0(\neg X)$	$10P^0(\neg X)$

$EU(U) > EU(\neg U)$ *if and only if*

$$P^0(D)[10P^0(X|E)-c]+P^0(\neg D)[10P^0(\neg X|E)-c] > P^0(D)[10P^0(\neg X)]+P^0(\neg D)[10P^0(\neg X)]$$

if and only if

$$P^0(D)[10P^0(X|E) - c] + P^0(\neg D)[10P^0(\neg X|E) - c] > 10P^0(\neg X)$$

if and only if

$$P^0(D)[10P^0(X|E) - c] + P^0(\neg D)[10P^0(\neg X|E) - c] - 10P^0(\neg X) > 0$$

if and only if

$$10P^0(D)P^0(X|E) - P^0(D)c + 10P^0(\neg D)P^0(\neg X|E) - P^0(\neg D)c - 10P^0(\neg X) > 0$$

Since $P^0(D)c + P^0(\neg D)c = c$, the above inequality holds *if and only if*

$$10P^0(D)P^0(X|E) + 10P^0(\neg D)P^0(\neg X|E) - 10P^0(\neg X) > c$$

Replacing $P^0(\neg D)$ by $[1 - P^0(D)]$, and $P^0(\neg X)$ by $[1 - P^0(X)]$, and $P^0(\neg X|E)$ by $[1 - P^0(X|E)]$, we will see that the above inequality holds *if and only if*

$$10P^0(D)P^0(X|E) + 10[1 - P^0(D)][1 - P^0(X|E)] - 10[1 - P^0(X)] > c$$

if and only if

$$10P^0(D)P^0(X|E)+10[1-P^0(D)-P^0(X|E)+P^0(D)P^0(X|E)]-10+10P^0(X) > c$$

if and only if

$$10P^0(D)P^0(X|E)+10[1-P^0(D)-P^0(X|E)+P^0(D)P^0(X|E)]-10+10P^0(X) > c$$

Doing some algebraic gymnastics, the above inequality holds *if and only if*

$$20P^0(D)[P^0(X|E) - \tfrac{1}{2}] > c + 10[P^0(X|E) - P^0(X)]$$

Assuming that the agent knows that E confirms X, then they know that $P^0(X|E) - P^0(X) > 0$. So, they could argue that the RHS of the above inequality is greater than c and therefore the above inequality holds only if the LHS is greater than c i.e.

$$20P^0(D)[P^0(X|E) - \tfrac{1}{2}] > c$$

The agent does not know the value of $P^0(X|E) - \frac{1}{2}$ but they know whatever value it has, it is *less than* 1. This implies that the above inequality holds *only if* $20P^0(D) > c$ which is equivalent to $P^0(D) > \frac{c}{20}$

Chapter 2

Does Political Relevance Make a Difference When Reading Science?

Abstract

I will argue that the mere fact that a claim is politically salient, even though it is well-supported within a peer-reviewed literature, can be enough to justify suspension of judgement on it. I will model a bounded Bayesian who cannot engage with every relevant possibility and show why they should engage with the possibility of an article being biased when its claim is politically salient. Then, I will show why such engagement destabilizes their credence towards the article's claim. Finally, I will use a probabilistic account of belief to show why this destabilization is enough to justify excluding politically salient claims from their set of beliefs.

Suppose that you have the habit of reviewing your social media feed early in the morning, bookmarking the interesting posts, and then returning to them late in the evening to decide which ones to endorse by re-sharing. This morning you come across two interesting peer-reviewed articles: article 1 claims that there is a strong statistical correlation between being a second-generation Muslim in London and having radical tendencies; article 2 claims that there is a strong statistical correlation between being diagnosed with an autoimmune disease and being diagnosed with anxiety disorder. Skimming through them, they seem credible. So, you bookmark both and, as always, postpone the decision on resharing them to the evening. During lunch, you remember what you read in the morning and as you contemplate the decision that you are about to make in the evening, two voices in your head have the following conversation:

Voice 1: How confident are you that the claims of these articles are true?

Voice 2: Quite confident! I would say 85 percent!

Voice 1: So, do you *believe* both of them?

Voice 2: No! I believe the article about the autoimmune disease (article 2) but I do not believe the one about the radical tendencies among the second-generation Muslim population in London (article 1)!

Voice 1 (smirking): Ok! I get it. You are very responsible, both morally and epsitemically!

Voice 2: I wish! But I'm afraid my response had nothing to do with morality! I'm simply attempting to manage my mental space efficiently while upholding a consistent set of beliefs!

In this paper, I will argue for Voice 2's claim i.e. that there is a categorical difference between article 1 and article 2 and that this difference justifies not believing the former while believing the latter and while maintaining a high degree of confidence towards both of them. This categorical difference is the pertinence to a major political controversy. The first article is related to many live political controversies: is radicalism among Muslims a real issue? If it is, is it related to geopolitics of the region in which Muslim population is located or are there some essential ingredients in the culture that give rise to radicalism? Should it affect immigrant policies about Muslims? The second article, however, is unrelated to any major political controversy.

But how does this difference affects whether we should believe them or not? If we use a probabilistic account of belief, like the one offered by Leitgeb [39], whether *believing* a proposition is justified for a Bayesian depends on their domain of credence function. However, as bounded Bayesians with limited memory and processing capacity, we cannot, and it is not instrumentally rational for us to include every logically possible proposition in that domain. Drawing from the literature on evidence gathering inspired by Good [24]'s principle of total evidence, I argue that we need to choose what to include in the domain of our precise credence function according to the decisions under uncertainty that we face, the experiments available to improve those decisions, and our crude comparative probabilistic judgments of the outcomes of those experiments. These comparative probabilistic judgments differ between political and non-political articles, justifying the maintenance of a different epistemic attitude towards them. The background assumptions are that we are bounded, instrumentally rational and socially aware Bayesians in ways that will be explained later on. Here is a rough version of the argument:

Claim 1: As bounded agents our space of live possibilities is coarse-grained at any point in time [10]. As Bayesians, the level of fine-graining of this space determines which propositions are included in the domain of our *precise credence function.*

Claim 2: As bounded Bayesians, making the space of possibilities more fine-grained with respect to a proposition is a matter of *volition* and *costly* for us [64]. This is because we must decide on the degree of relevance of the new proposition to the previous ones and formulate a new credence function that adheres to probabilism. As instrumentally rational agents, we start with a fine-graining of the space of possibilities in accordance with the decision(s) we have to make under uncertainty. Subsequently, we decide on further fine-graining of that space in accordance with whether we find a possibly decision-changing evidence sufficiently likely.

Claim 3: As socially-aware agents making decisions based on peer-reviewed articles, our judgment about whether a strong (possibly decision-changing) piece of evidence against the reliability of an article is sufficiently likely, varies depending on whether the article is relevant to a major political controversy or not.

Subconclusion: As socially-aware, instrumentally rational, bounded Bayesians who are making a decision based on their credence for the claim of a peer-reviewed article, which propositions are included in the domain of *precise credence function*, differs between peer-reviewed articles that are relevant to a major political controversy and those that are not. (Claims 1,2,3)

Claim 4: We *should believe* a proposition if and only if our credence towards it is stable, and whether our credence towards a proposition is stable depends on other propositions included in the domain of our credence function.(Leitgeb's stability thesis[1])

[1] This can be replaced by any other probabilistic accounts of belief which is partition-dependent like Salow and Goodman [55] and Lin and Kelly [42].

Conclusion: As socially-aware, instrumentally rational, bounded Bayesians who are making decisions based on our credence for the claim of a peer-reviewed article, what we *should believe* differs between peer-reviewed articles that are relevant to a major political controversy and those that are not. (Subconclusion, Claim 4)

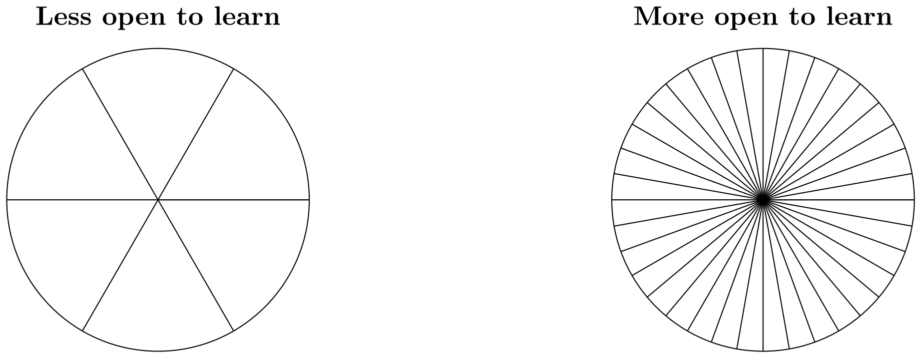
To clarify the position of my claim and its standing in the literature, I want to underscore four key points. First, my argument does *not* rely on any moral considerations, setting it apart from relevant discussions in the moral encroachment literature [3, 8]. Second, my assertion revolves around justified epistemic discrimination between peer-reviewed articles. Specifically, those confirming one side of a political controversy and those that are unrelated to a political debate. This approach distinctly differs from discussions focusing on statistical generalizations and their potential epistemic pitfalls [47]. Third, my claim in this paper pertains to any scientific article, regardless of whether it is theoretical or empirical, and if empirical whether they prove a substantial causal claim or a simple correlation. Fourthly, while I present a new framework, I contend that my conclusion can also find support in other frameworks, such as zetetic epistemology [18] and question-sensitive epistemology [32]. While I do not delve into these frameworks in this paper due to space constraints, I believe that exploring them and contrasting potential arguments in their frameworks with mine could yield interesting insights.

Here is the structure of the paper: I begin by introducing a framework for determining the level of fine-graining of the space of live possibilities for a bounded Bayesian agent (claims 1 and 2 of the above argument). Next, I use this framework to elucidate why we are rationally compelled to fine-grain the space of possibilities differently in the case of peer-reviewed articles relevant to a major political controversy and those that are not (claim 3 and Subconclusion). Then, I introduce Leitgeb's stability thesis and discuss how it is affected by the level of fine-graining of the space of possibilities (claim 4). Finally, I use Leitgeb's thesis to explore how the difference in the level of fine-graining of our space of possibilities for articles relevant and irrelevant to a major political controversy affects the rationality of *believing* them.

2.1 Choosing the Mind's Mesh

2.1.1 What is fine-graining?

An ideal Bayesian agent can learn any logically possible proposition and come up with a posterior credence function by conditionalizing. We, bounded Bayesians, however, do not have that luxury. We are bound to a coarse-grained space of possibilities, and that space determines what we can possibly learn. It is helpful to think of our space of possibilities as a net that we throw into the sea of infinitely many possible worlds. The more fine-grained the mesh of the net, the smaller the size of the fishes we can catch. The more fine-grained the mesh of our partition, as bounded Bayesians, the more detailed information we can learn. This mesh in fact determines what information we


Less open to learn

More open to learn

Figure 1: Fine-grained vs coarse-grained spaces of possibilities

are *open to learn*.

But how does the level of fine-graining of this partition change? Is it a voluntary action, or something that happens to us? I contend that both coarse-graining and fine-graining can be a voluntary act. Coarse-graining the partition entails merely disregarding or forgetting certain distinctions. Forgetting can occur involuntarily, while deliberate disregard requires volition. On the other hand, fine-graining the space of possibilities with respect to a new proposition, of which we have become aware, demands evaluating the degree of dependence between the new proposition and the existing ones, all while adhering to Kolmogorov's rules. This endeavor demands exertion and exacts a cognitive toll. If we accept that cognitively taxing tasks necessitate at least some degree or form of volition, it follows that fine-graining the space of possibilities with respect to a new salient proposition can be a matter of volition[2]. So, it can constitute an action and therefore falls under the norms of instrumental rationality.

Not only, *generating* a new level of fine-graining is cognitively taxing, but also *maintaining* it is costly. A more fine-grained space means being open to learn more and this openness comes with a cost for processing information and even learning itself[3]That is because although having a more fine-grained partition facilitates learning the propositions that have already been considered in the partition, it makes learning the propositions of which we have not been aware more difficult. The more fine-grained the partition the more difficult and costly including a newly salient proposition as we have to make more judgments of probabilistic relevance while adhering to Bayesian norms. Moreover, the more fine-grained the partition, the more the number of relevant unconsidered propositions. In fact this number grows exponentially as the number of carvings of the partition grows. If we assume that the greater the number of relevant unconsidered propositions, the greater the rate of the propositions that become salient to us, then it is not hard to admit that a more fine-grained partition can make processing information and learning newly salient propositions more difficult.

[2] Of course, in some instances, this occurs automatically or subconsciously, which is consistent with my argument, which only assumes the possibility of volition.

[3] This aligns with the literature on the computational complexity of probabilistic inference in Bayesian belief networks, which suggests that maintaining numerous distinctions and calculating various conditional probabilities can result in a combinatorial explosion, rendering the task intractable [14], [15].

Therefore, as instrumentally rational agents, we need to consider the costs of generating and maintaining a new level of fine-graining as we decide to whether fine-grain our space of possibilities with respect to a newly salient proposition. In this paper, I model us as agents who are trying to make some decisions under uncertainty about X. They start off with a coarse-grained partition over $\{X, \neg X\}$ and then decide on further fine-graining with respect to the outcomes of available experiments which are relevant to X. I will call these decisions the *micro-decisions of fine-graining.* I assume that as instrumentally rational agents, we need to account for two general considerations when making these micro-decisions:

1. **Utilities and prior credences involved in the decision(s) under uncertainty we face:** I will refer to these decisions as *macro-decisions.* They matter not only because their states of uncertainty gives us a good starting point for the partition cells, but also because they determine the expected utility (EU) at stake which we need to aim to maximize by finding the optimal level of fine-graining. We should decide to become (and remain) open to learning which propositions by considering how this openness affects our EU.

2. **Available experiments:** When deciding which propositions to be open to learning, we should account for experiments are available to us. I assume an experiment is a set of proposition which represent its possible outcomes. An outcome can change our mind about a decision that we are about to make or not depending on how much it affects our credence towards the states of uncertainty. If an outcome of an experiment changes our credence to the extent that it changes our decision, I will call it 'decision-changing'[4].

I am modeling us as agents who are making decisions and conducting experiments while *not* maintaining a *precise* credence function over all the propositions that are salient to us in these procedures. That is because it is not optimal for us to include every salient proposition in the domain of our precise credence function given the discussed cost of generating and maintaining a new level of fine-graining in the partition over which the precise credence function is defined. I assume that we are aware of this cost and hence very conscious of not over-populating the domain of our precise credence function. The following point is crucial in this picture: it is not the case that we lack any judgement whatsoever about the propositions which are salient to us and yet not included in the domain of precise credence function. We can maintain a partial[5] order over those propositions to represent the order of our degrees of confidence towards them. Of course maintaining a coherent partial order over a set

[4] Buchak [12] refers to the evidence that changes our opinion with respect to a decision as 'opinion-changing'. I believe this term may have some misleading connotations. It may make some readers think that we are concerned with whether considering a piece of evidence has some fatal *epistemic* consequences while we are focused on *instrumental* consequences. I have chosen the term 'decision-changing' instead to avoid this unwanted connotation.

[5] A partial order is a binary relation which is reflexive, antisymmetric, and transitive. A relation R on set Σ is reflexive iff for every member a of Σ, aRa. A relation R on set Σ is antisymmetric iff for every two distinct members a and b of Σ, if aRb then $\neg bRa$. A relation R on set Σ is transitive iff for every three members a, b, and c of Σ, if aRb and bRc then aRc.

of propositions is not cost-free, but it is safe to assume that its cost is *usually*[6] less than maintaining a precise probability function over them. So, if we call the partition over the space of logical possibilities, defined by our attention, salient possibilities, it is intuitive to assume that as instrumentally bounded rational agents who are aware of the cost of including a proposition in a coherent partial order and the domain of a probability function, the partition that we maintain for a partial order of confidence is *usually* less fine-grained that salient possibilities and more fine-grained than the partition for the precise credence function.

Of course, there are many questions to be answered for this picture to be complete: what are the exceptions? Do we hold different partitions for total and non-total orders, given that maintaining the former is more costly? What factors determine the salient possibilities and are they constant during micro-decisions of fine-graining? I do not aim to address these questions here. For the purpose of this paper, I only need the following minimal assumption: *For some sets of propositions $\{p_1, ..., p_n\}$ the cost of fine-graining the space of possibilities with respect to the whole set for the purpose of maintaining a non-total confidence order among them is negligible compared to the cost of fine-graining with respect to any member p_i of the set for the purpose of maintaining a precise probability function*[7]*.*

I need this assumption because in my model of an instrumentally rational agent deciding on further fine-graining of her partition for precise credence function with respect to a proposition, I assume that it is rational for the agent to use a non-total confidence order of some propositions which are not in the domain of her precise credence function to inform the fine-graining decision. To put it formally, I assume that the agent is rational to maintain a partition over some propositions which are not in the partition for precise credence function in order to inform that fine-graining decision. But if the cost of maintaining the partition for the non-total confidence order is not negligible compared to the cost of maintaining the partition for precise credence function, then it is not clear that it is rational for an instrumentally rational agent to undergo that cost to inform her fine-graining decision. So, this minimal assumption is essential for my argument in the next section.

This minimal assumption is true because not only the task of *comparing* confidence can be only negligibly taxing compared to assigning a precise degree of credence, but also observing the coherence constraints over a non-total order is negligible compared to observing probabilism. For example suppose we are about to make a macro-decision based on our uncertainty about whether a coin lands head or tail: $\{H, T\}$. So H and T are already in the domain of precise probability function and hence has corresponding cells in the partition of both precise credence and confidence order. Now suppose 101 possible biases of the coin towards head is salient to

[6] If the order is total in the sense that every two members of the set is comparable, it is not obvious that cognitive cost of maintaining an order over the set is less than the cost of forming a probability function over them.

[7] I did not make this a universal claim, as there may be pairs of propositions for which comparing our confidence is difficult and costly. So, although the cost is still less than forming a precise credence function over them, it is not obvious that the cost is *negligible*.

us: $\{C_0, C_{0.01}, ..., C_1\}$ are salient to us. The cognitive cost of fine-graining the space of possibilities with respect to $\{C_{0.1}, C_{0.5}, C_{0.9}\}$ in order to maintain the non-total confidence order $\{(H \succ C_{0.1}), (H \succ C_{0.9}), (C_{0.5} \succ C_{0.9})\}$ is *negligible* compared to fine-graining the space of possibilities with respect to any of $C_{0.1}$, $C_{0.5}$ or $C_{0.9}$ for the purpose of including them in the domain of precise credence function. That is because in order to fine-grain the domain of precise credence function with respect to a proposition like $C_{0.1}$, we need to decide on the precise degree of dependence between H and $C_{0.1}$ while respecting probabilism.

I use the notation $q \succ r$ to represent that q stands higher in the agent's confidence order than r. I also use the notation $q \succ_s r$ to imply assuming s is true, q stands higher in the agent's confidence order than r. Finally, I use the notation $P(q)$ to represent the agent's *potential* precise credence for q. If they already have the fine-graining with respect to q in their partition for precise credence then $P(q)$ is their actual precise credence for q and if not $P(q)$ is what their precise credence for q would be if they had the fine-graining with respect to q. It is natural to assume that $P(q) > P(r)$ if and only if $q \succ r$. So, an agent can use their confidence order to make *comparative* judgments about $P(q)$ even when they do not have q in the domain of their precise credence function.

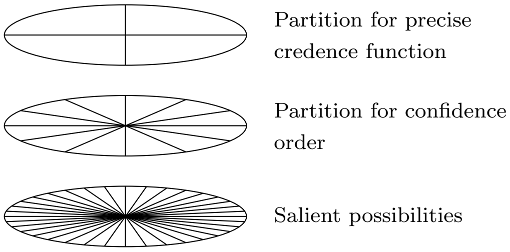

Partition for precise credence function
Partition for confidence order
Salient possibilities

Figure 2: Instrumentally rational agents maintain different partitions over the space of possibilities for different purposes

In what follows, I will give an example of a bounded agent who has to make a macro-decision under uncertainty in the near future, and they are deciding on further fine-graining of their parition for precise credence function with respect to the outcomes of the available experiment. This micro-decision of fine-graining is informed by the order of confidence that they hold over the outcomes of the experiment and other salient propositions. The example will shed light on what I mean by the outcome of an experiment being sufficiently likely given the cost of fine-graining.

2.1.2 Fatima's Example

Suppose Fatima is epileptic and is supposed to take one pill every day to control her seizures. She usually takes it in the morning. It is the evening and she is on her way home from work and she starts wondering whether she forgot taking the pill in the morning. This uncertainty is relevant to a decision that she has to make in future: she has to decide whether to take the pill before going to bed. If she takes the pill and she already took the pill in the morning, she will definitely experience some annoying side-effects tomorrow and if she does not and she also did not take it in the morning, there is a possibility of seizure. Also, suppose she considers a possible seizure twice as bad as the annoying side effects of an extra dosage. Her decision table for this decision looks like:

		$\boldsymbol{X}$: I took the pill in the morning	$\neg\boldsymbol{X}$: I did not take the pill in the morning
→	$\boldsymbol{A_1}$: Take the pill before bed	-10	0
	$\boldsymbol{A_2}$: Do not take the pill before bed	0	-20

Fatima's macro-decision

The small arrow represents the initial inclination of Fatima given her prior credence towards X. As she is walking from the tube station to her place and contemplating this decision, she realizes that she is only 30% sure that she took the pill. So, she is inclined to take the pill before bed. Since, she still has time before making this decision, she decides to do more investigation before making up her mind. Given the above tables, she knows her credence towards X should be more than $\frac{2}{3}$ in order for her to change her mind about the macro-decision. So, she starts wondering if there is a piece of evidence that can pushes her credence above this threshold. Using the terminology introduced in the previous section, she starts wondering whether there is any *decision-changing* piece of evidence that she can learn through an experiment.

It occurs to her that she occasionally but not rarely drinks a glass of water right after she wakes up and sometimes, when she drinks water, she takes her epilepsy pill with it. So, she can inspect her flat to see whether a glass of water was left somewhere.

$\boldsymbol{E}$: A glass of water is left somewhere in the flat.

If she finds one, she will become more confident that she took the pill (more than 30%) but she does not know how confident she will be and more specifically she does not know whether her credence for X will be more than $\frac{2}{3}$. So, she does not know whether E is decision-changing with respect to the macro-decision.

She starts wondering whether she should bother inspecting her flat given that she does not know whether the result will change her mind about taking the pill. She also faces a micro-decision of fine-graining. At this stage, she has a precise credence over $\{X, \neg X\}$ and as a bounded Bayesian, she wonders whether she should pay the cognitive cost of fine-graining to further fine-grain the domain of precise credence function with respect to E. Combining these two decisions, it seems that she has four options to choose from: (i) fine grain with respect to E and inspect the flat, (ii) do not fine-grain with respect to E and inspect the flat, (iii) fine grain with respect to E and do not inspect the flat, and (iv) do not fine grain with respect to E and do not inspect the flat.

I will argue that, as an instrumentally rational agent, options (ii) and (iii) are off the table for her and her real choice is between options (i) and (iv). Let me explain why. It is not rational for her not to fine-grain wrt E and yet does the inspection (option (ii)). That is because if she does not fine-grain the domain of her precise credence function wrt E, she will not know the degree of dependence between her credence for X and her credence for E. So, she will not know how her credence for X changes after observing E and therefore, observing E has no effect whatsoever on her precise credence of X and hence her macro-decision based on X. This implies observing E is not decision-changing for her and hence inspecting the flat is instrumentally irrational for her[8]. That is because she pays the cost of inspection without any gain.

On the other hand, it is not rational for her to fine-grain wrt E and yet does not do the inspection (option (iii)). That is because she pays the cost of fine-graining without any gain[9]. Now, let's see how she should choose between the options (i) and (iv). She does not know whether E is decision-changing with respect to the macro-decision. Assuming that c is the *combined cost of fine-graining and inspection*, she faces the following decision under uncertainty:

	$\boldsymbol{D}$: E is decision-changing				$\neg\boldsymbol{D}$: E is not decision-changing
$\boldsymbol{F}$ **:** Fine-grain and inspect	$\boldsymbol{X}\&\boldsymbol{E}$	$\neg\boldsymbol{X}\&\boldsymbol{E}$	$\boldsymbol{X}\&\neg\boldsymbol{E}$	$\neg\boldsymbol{X}\&\neg\boldsymbol{E}$	$-c$
	$0-c$	$-20-c$	$-10-c$	$0-c$	
$\neg\boldsymbol{F}$ **:** Neither fine-grain nor inspect	-10	0	-10	0	0

Let's review each cell of this table. Fatima does not know whether receiving E is decision-changing[10] (D) i.e. she does not know whether it pushes her credence of X above $\frac{2}{3}$. If it is not decision-changing (column 2) then by choosing to fine-grain and inspect, she pays the cost of fine-graining and inspecting without any gain. So, her overall utility will be $-c$.

But what if E is actually decision-changing? Then the picture will be more complicated. That is because the utility of the outcome depends on whether she learns E or $\neg E$ and whether she learn that evidence in X or $\neg X$. It matters because if she learns that a glass of water is somewhere in the flat (E), she is going to switch to not taking the pill (A_2). Now, if she does this switching while she actually took the pill in the morning (X), then this learning will improve her utility. But if she switches to not taking the pill while she did not actually take the pill in the morning ($\neg X$), then this learning worsens her utility. It can be proved that the utility gain of $X\&E$ is more than the utility loss in $\neg X\&E$ if there is no cost in the picture [11]. However, with a cost in the picture, which state dominates depends on $P(E)$.

[8] One may ask what if she has the distinction with respect to E in her partition for confidence order? Can't she consult her confidence order to check whether E is decision-changing given that whether a proposition is decision-changing depends on an *inequality*? The answer is she definitely can but given the picture that I depicted in the previous section according to which the confidence order is a *non-total* order, there is no guarantee that she can judge whether E is decision-changing with a *negligible* cost.

[9] Here, I assume that the cost and gain is solely evaluated with respect to the expected utility of this specific macro-decision. This assumption is justified when we are focused on instrumental rationality (as opposed to epistemic rationality) and when there is no other other macro-decision in the attention horizon which also depends on X or E.

[10] Proposition D is in fact a proposition about Fatima's potential precise credence of E. D : $P(X|E) > \frac{2}{3}$

[11] The proof follows the same strategy that Good [24] follows to prove the principle of total evidence.

Let's delve into more detail regarding the cells under the column D. Remember that Fatima's prior for X is 0.3. So, she is initially inclined to take the pill (A_1). Starting from the last row ($\neg F$), what happens if she neither fine-grains nor inspects? In that case, she keeps choosing what she was about to choose i.e. A_1. That means, as before, she will get -10 out of the macro-decision when X is true and 0 when $\neg X$ is true and regardless of whether E is true or not[12].

Now, let's focus on the first row (F): what happens if she fine-grains with respect to E and inspect the flat? That way not only she will be able to *learn*[13] E or $\neg E$, but also *she will know that E is decision-changing.* If she finds a glass of water (E) in her inspection, her credence for X will be raised above $\frac{2}{3}$ and she will decide not to take the pill. So, if $X\&E$ is the state of the world, she will get utility 0, but she also undergoes the cost of fine-graining and inspection (c). So, her overall utility in that state will be $0 - c$. And, if she finds a glass of water while she did not actually take the pill ($\neg X\&E$), she will get the utility -20 out of the macro-decision and given that she underwent the cost of fine-graining, her overall utility will be $-20 - c$. On the other hand, if she does not find a glass of water ($\neg E$), she keeps choosing A_1. So, if she is in the state $X\&\neg E$, her overall utility will be $-10 - c$ and if she is in the state $\neg X\&\neg E$, it will be $0 - c$. Using this table:

$EU(F) - EU(\neg F) = P(D)[10 * P(X\&E|D) - 20 * P(\neg X\&E|D)] - c$

If $P(E\&D) \neq 0$ and $P(X|E\&D) > \frac{2}{3}$, it is easy to prove that[14]:

$EU(F) - EU(\neg F) > 0$ if and only if

$$P(E\&D) > \frac{\frac{c}{20}}{\frac{3}{2} * P(X|E\&D) - 1} \tag{2.1}$$

It is important to note that Fatimah can maintain comparative judgments about $P(E\&D)$ and $P(X|E\&D)$ with negligible cost even before she includes E or D in the domain of her precise credence function. That is because, as it was discussed in the previous section, these judgments can be implied by a non-total confidence order and fine-graining the space of possibilities for the purpose of maintaining a non-total confidence order can be cost-negligible.

Looking at the above formula, it is easy to see that under these assumptions, if the fine-graining and inspection cost is 0 then $EU(F) - EU(\neg F) > 0$ is always true. That is because the above biconditional will be reduced to $EU(F) - EU(\neg F) > 0$ if and only if $P(E\&D) > 0$, which is always true under those assumptions. This means if E has non-zero probability and Fatimah thinks it is not impossible that E is decision-changing ($P(D) > 0$) and inspection and fine-graining is cost-free, then she

[12] I would like to emphasize that the fact that in this state E is decision-changing does *not* imply that she received E or that she found it decision-changing. Fatima receives E only if she does the inspection and she *finds* E decision-changing only if she has the distinctions with respect to E in her partition for precise credence function. Only then she can know how X and E are correlated and whether E is decision-changing.

[13] In proper Bayesian way of 'learning'.

[14] Please check the appendix for a more general version of this derivation.

is rationally obliged to fine-grain and do the inspection.

However, fine-graining and inspection is not cost-free. It is plausible to assume that the cost is small relative to the gain of making a better macro-decision but it is *not* zero. In that case the right hand side on the inequality will be a small number. Can Fatimah uses her confidence order to check whether that inequality holds?

Let's make the case more tangible by assigning some real-valued numbers: suppose the cost of fine-graining and inspection is only 0.1 percent of the gain of making a better decision in the worst case scenario($c = \frac{20}{1000}$). Given the macro-decision she faces, she knows that if E is decision-changing then her posterior will be greater than $\frac{2}{3}$. So, if we assume $C_{\frac{2}{3}}$ to be the proposition that 'a biased coin with $\frac{2}{3}$ chance of landing head, lands head' in fact she maintains the following confidence order: $X \succ_{E\&D} C_{\frac{2}{3}}$. This implies $P(X|E\&D) > \frac{2}{3}$. Moreover, she thinks it is not unlikely (more than 10% likely) that she finds a glass of water upon inspection and changes her mind about the macro-decision. If we assume $C_{0.1}$ to be the proposition that 'a biased coin with 0.1 chance of landing head, lands head' in fact she maintains the following confidence order: $E\&D \succ C_{0.1}$. This implies $P(E\&D) > 0.1$. So, she has the following collection of comparative probabilistic judgments:

$$\frac{2}{3} < P(X|E\&D) \leqslant 1$$
$$0.1 < P(E\&D) \leqslant 1$$

Additionally, we may assume that Fatima knows that she never maintains precise credence with more than one decimal points. So if E is decision-changing, her precise posterior credence will be equal or more than 0.7. Given these assumptions, the following set of inequalities hold about her precise credence function:

$$0.7 \leqslant P(X|E\&D) \leqslant 1$$
$$0.1 < P(E\&D) \leqslant 1$$

Since $P(X|E\&D) \geqslant 0.7$ and $\frac{c}{20} = 0.001$, the above fraction $\dfrac{\frac{c}{20}}{\frac{3}{2} * P(X|E\&D) - 1} <$ 0.02, and since $P(E\&D) > 0.1$ and $0.1 > 0.02$, the above inequality holds i.e.

$$P(E\&D) > \frac{\frac{c}{20}}{\frac{3}{2} * P(X|E\&D) - 1}$$

So, in this scenario and with this collection of comparative probabilistic judgments, Fatima is rationally obliged to fine-grain wrt E and do the inspection.

Now, consider a different scenario in which cleaning services have visited Fatima's place. In this scenario she assigns the same cost to inspecting the flat ($\frac{c}{20} = 0.001$), and she has the same comparative probabilistic judgments about $P(X|E\&D)$ and $P(D)$ but given that cleaning services always remove and wash used glasses/dishes she considers finding a glass of water somewhere in the flat very unlikely (less that 0.1% likely). So, she has the following collection of comparative probabilistic judgments:

$$0.7 \leqslant P(X|E\&D) \leqslant 1$$
$$0 < P(E) < 0.001$$

$P(E) < 0.001$ and $P(E\&D) \leqslant P(E)$. Therefore, $P(E\&D) < 0.001$. On the other hand, $\frac{c}{20} = 0.001$, therefore, $P(E\&D) < \frac{c}{20}$. Moreover, $P(X|E\&D) \leqslant 1$, therefore $\frac{3}{2} * P(X|E\&D) - 1 < 1$. This implies $\dfrac{\frac{c}{20}}{\frac{3}{2} * P(X|E) - 1} > \dfrac{c}{20}$. So,

$$P(E\&D) < \frac{\dfrac{c}{20}}{\frac{3}{2} * P(X|E) - 1}$$

This means in this scenario, the opposite of the mentioned inequality holds and hence Fatima is rationally obliged *not* to do the fine-graining or inspection. To summerize informally[15]:

Fine-graining norm

Fatimah should fine-grain her space of possibilities with respect to the proposition that a glass of water is somewhere in the flat (E) and do the inspection if she deems the possibility that she finds a glass of water in the flat and that it changes her mind about the decision she faces ($E\&D$), *sufficiently likely* and she should neither fine-grain nor do the inspection if she deems the possibility of $E\&D$ not sufficiently likely.

2.1.3 Mind's mesh while considering a peer-reviewed article

Going back to the example at the beginning of the paper, suppose I came across two peer-reviewed articles on social media. I skimmed through them and they seem credible. So, I am initially quite confident (say 95%) that the claims of these articles are true. I have to make some decisions based on them in future, one of which is

[15] In the more general case, when there are more than 2 options that can have more than 2 consequences, the relevant formula will ask whether it is sufficiently likely that she receives the evidence that changes her mind about the decision— the expected gain minus the cost has to be worth it. But this simpler model is sufficient to reveal the underlying general dynamics.

resharing them on social media in the evening (I chose X to be the proposition that the article's claim is false so that it better mirrors Fatima's example).

	$\boldsymbol{X}$: The article's claim is false	$\neg\boldsymbol{X}$: The article's claim is true
$\boldsymbol{A_1}$: Do not reshare the article	M	M
→ $\boldsymbol{A_2}$: Reshare the article	L	H

Decision of resharing the article

Given my low prior of X (0.15), I am inclined to reshare both of the articles, as I contemplate this decision, it occurs to me that I am aware of some dubious institutions and individuals. Henceforth, I will refer to this collection of institutions and people as 'red flag bodies'. I can check whether an article is related to them by doing a quick research on its authors and funding sources. I have not done that quick research on the two mentioned articles yet. Also, suppose I am a bounded Bayesian whose space of possibilities, while contemplating the decision of resharing each article, is only fine-grained with respect to the truth/falsity of the main claim of that article. During this contemplation I ask myself: Should I do the research to check the truth of the following proposition and fine-grain my space of possibilities with respect to it?

$\boldsymbol{E}$: The article is related to red flag bodies.

I am not sure how my credence changes if I discover that an article is associated with a red flag body. So, I am not sure whether such discovery changes my decision about resharing them. Following the reasoning in Fatima's case, I need to choose between $\boldsymbol{F}$ (fine-grain wrt E and do the research) and $\neg\boldsymbol{F}$ (Neither fine-grain nor do the research) and that decision depends on my comparative probabilistic judgment of the possibility that E is true and decision-changing ($E\&D$). Follwing the inquality (4), introduced in the appendix in the course of generalizing Fatima's case, $EU(F) > EU(\neg F)$ iff

$$P(E\&D) > \frac{\dfrac{c}{H-M}}{P(X|E\&D)(\dfrac{M-L}{H-M}+1)-1}$$

Now, suppose I value seizing the opportunity of resharing articles with true claims as much as I disvalue resharing articles with false claims, then $H - M = M - L$. This implies that the threshold for my credence towards X passing which changes my opinion of the above decision is 0.5. Assuming that my precise credence only has one decimal place accuracy, this implies that if E is decision-changing my credence is equal or above 0.6 i.e. $P(X|E\&D) \geqslant 0.6$. Finally, assume that I regard the cost of the research on checking the reputation of authors and funding sources 0.1% of the gain of resharing an article with a true main claim ($\frac{c}{H-M} = 0.001$). Given these assumptions the above inequality is reduced to:

$$P(E\&D) > \frac{0.001}{2 * P(X|E\&D) - 1} \tag{2.2}$$

Whether this inequality holds depends on our confidence order inspired by the context of the experiment, the outcome of which is E. I will argue that it is rational to have two distinct confidence orders and hence collections of comparative probabilistic judgments for political and non-political articles. If my non-total confidence order implies the following comparative probabilistic judgments, then following the same pattern of reasoning as Fatima's case, it is not hard to show that given the left collection (political articles), the inequality (3) holds and hence, I should fine-grain my space of possibilities with regard to E for political articles. However, given the right collection, that inequality does not hold and hence I should not fine-grain my space with regard to E for non-political articles[16].

Political article

$$0.5 < P(X|E\&D) \leqslant 1$$
$$0.1 < P(E\&D) \leqslant 1$$

Non-political article

$$0.5 < P(X|E\&D) \leqslant 1$$
$$0 < P(E\&D) < 0.001$$

Looking at the above lists of inequalities, we can see that the only difference between the two collections is the last element i.e. the one about $P(E\&D)$. But why is that so? If I focus on decision-changing red flags, why does the fact that an article is related to a live political controversy affects my comparative probabilistic judgement of coming across a red flag in the reputation of its authors and funding sources?

We need to dig deeper. In the both cases, I'm unlikely to be able to find a content-related red flag. But in the political case, it's significantly more likely I'll be able to find some non-content-related red flag, via finding conflict of interest, likely bias, and more generally using my social awareness. Suppose I focus on the scenarios in which the redflags are decision-changing. The likelihood of me finding an association with a red flag body for an article depends on what is included in the list of 'red flag bodies' that I use as a reference during my research, i.e., looking for red flags among the authors and funding sources. So, if the outcome of the experiment is E, then it is discovered through (at least) a member of this pre-identified list of red flags. This list is informed by both my expertise and my social awareness. Let's refer to the elements of the list that are included because of my expertise as 'Expertise-members' and the elements included because of my social awareness as 'SA-members'.

A decison-changing outcome E of the experiment is determined either through an Expertise-member or an SA-member ($P(E\&D) = P(E_{\text{Exp}}\&D) + P(E_{\text{SA}}\&D)$). The fact that I could not find any decison-changing redflag in the *content* of the articles through skimming them suggests that it is unlikely that they are related to my area of expertise and hence the possiblity of a decision-changing E being discovered through an Expertise-member is very unlikely. Let's say this lead us to make the comparative judgment $P(E_{\text{Exp}}\&D) < 0.0005$. On the other hand if the article is unrelated to any

[16] Check the appendix for an step by step reiteration of the argument.

political controversy, my social awareness is very unlikely to be helpful in finding a decision-changing red flag. Let's say this lead us to make the comparative judgment $P(E_{\mathrm{SA}}\&D) < 0.0005$. So, the possiblity of a decision-changing E being discovered through a SA-member is also unlikely. So, we have a reason to think $P(E\&D) < 0.001$.

That is not true for articles which are relevant to a political controversy and for which SA-members of the red flag list contribute to the discovery of a decision-changing red flag. For those articles I have no reason to think $P(E\&D) < 0.001$. That does not imply that I consider this possibility very likely or even more likely than not. However, it is totally reasonable to assume that I deem the possibility of finding a decision-changing red-flag more than 10% likely, while such an assumption is not reasonable in the case of non-political articles.

One may raise the objection that the fact that the proposed comparative probabilistic judgments are reasonable does not imply that one is obliged to have them. What if one has no comparative probabilistic judgment whatsoever about $P(E\&D)$? Or what if they think $0.001 < P(E\&D) < 0.1$? Given that the confidence order is assumed to be non-total, these scenarios are possible in my model without violating any rationality norm. The answer is that such possibilities are so marginal that they can be reasonably neglected in the analysis. That is because it is very rare for an agent to not have any comparative judgment whatsoever about the possibility of finding a decision-changing red flag in our research on the reputation of authors and funding bodies. It is also rare (and odd) to think that $P(E\&D)$ falls in the specific interval of $0.001 < P(E\&D) < 0.1$ without any additional reason.

Of course if one reduces the relative value of the cost of doing a research on authors and funding sources c and the cost of sharing false articles on social media $(M - L)$, the threshold introduced for $P(E\&D)$ by the instrumental norm of fine-graining decreases and eventually there will be a point that even for non-political articles we deem $P(E\&D)$ to be greater than that threshold and hence fine-graining to be rational. So, the difference between political and non-political articles diffuses, but I believe the difference remains salient for a wide range of plausible decision-making and experiment scenarios.

One natural question is whether the difference between the spaces of possibilities persists after doing the experiment. If I check the authors and funding sources and find no red flag, should I still maintain the fine-graining with respect to E in that space? The answer depends on whether I expect my reference list of red-flag bodies to change before I make a decision based on the truth of the claim of the article. In many cases, there are more than one decision that depend on that, and if I am socially vigilant, I anticipate my reference list of red flags to keep changing, which may lead to the revelation of a new red flag that is decision-changing with respect to some of the decisions that is going to be made. As long as I deem that possibility sufficiently likely, I am rationally obliged to maintain the fine-graining with respect to E in my space of possibilities.

Now, one may ask why it matters that the level of fine-graining of the space of possibilities differs between these two articles. The answer is that the level of fine-graining of the space of possibilities affects the *stability* of our credence. In the next section, I will review Leitgeb's notion of stability and argue why, in his framework, our credence towards X is only stable for article 2 and not for article 1, and what it implies for maintaining categorical belief towards the claims of these articles.

2.2 Believing peer-reviewed articles

The most natural way to connect categorical belief and degrees of belief (credences), is to determine a threshold like 50 percent and declare any proposition towards which our credence is over that threshold as *believed.* This approach is known as the Lockean thesis in the literature[17]. This thesis faces an important difficulty and that is if we use it to determine the set of the propositions that we believe, then the members of that set may be inconsistent and open under logical entailment. In other words, this approach is in clashes with two important norms of categorical belief: belief consistency and belief closure. The lottery paradox is a famous example to show this clash[18].

In order to fix this difficulty Leitgeb [39] suggests the following thesis:

- An agent should believe a proposition if and only if they assign *a stably high credence* to it.
- An agent assigns a stably high credence to a proposition q if and only if for every proposition r which is consistent[19] with q, their posterior towards q after learning r is above $\frac{1}{2}$.

This thesis fixes Lottery paradox, by declaring the belief that each specific ticket loses irrational. According to this thesis, although our credence towards the proposition that a specific ticket, like ticket 1, loses ($\neg T_1$) is high ($cr(\neg T_1) > 0.5$), it is not stable. That is because there is a proposition like 'ticket 1 wins or ticket 2 wins' ($T_1 \vee T_2$) which is consistent with it and yet upon conditionalizing on it, our credence towards $\neg T_1$ will not be high any more: $cr(\neg T_1|(T_1 \vee T_2)) = 0.5$.

What is important for the purpose of this paper is that in Leitgeb's framework, whether our high credence towards a proposition is stable depends on which propositions are consistent with it and that in turn depends on how fine-grained the space of possibilities is[20].

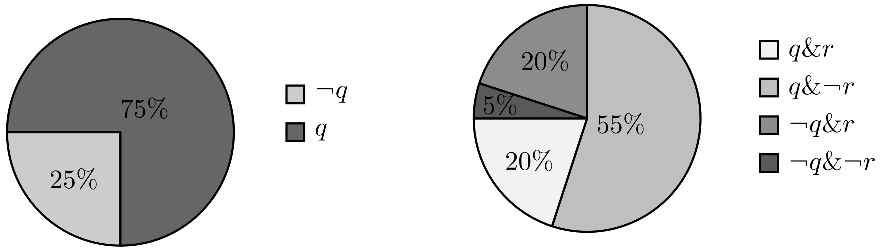
[17] **Lockean Thesis (Traditional):** For any perfectly rational agent, there exists a real number $s \geqslant \frac{1}{2}$ such that whatever categorical belief set B and credal distribution P she adopts, for any proposition X, $B(X)$ just in case $P(X) > s$.[61].

[18] **Lottery Paradox:** suppose there is a fair lottery with 1 million tickets. For such a lottery our degree of belief towards each ticket winning is 1 in 1000000. So, our credence towards the proposition '$Ticket_i$ will win the lottery'(henceforth X_i) is 1 in 1000000 for every i, and according to the Lockean thesis we should not believe any of these propositions and should believe the negation of them ($\neg X_i$). On the other hand, our degree of belief that one of these tickets will win is 1. So, our credence towards the disjunction $X_1 \vee X_2 \vee .... \vee X_{1000000}$ is 1 and hence according to the Lockean thesis, we should believe it. So our belief set will be like $\{\neg X_1, \neg X_2, ..., \neg X_{1000000}, (X_1 \vee X_2 \vee .... \vee X_{1000000}), ...\}$. It is not difficult to see that this set is either inconsistent or not closed under logical entailment.

[19] As I will explain later, the set of propositions an agent deems consistent with q depends on the set of propositions they have chosen to leave open.

[20] The partition-dependency of Leitgeb's notion of stability has been utilized by other researchers

For example if an agent has a credence measure like (a) in Figure 3, then their credence towards q is stably high and hence they should maintain a categorical belief towards it. That is because their credence towards q is above 0.5 and there is no proposition in their space of possibilities which is consistent with q and upon learning (conditionalizing on) which their credence towards q drops to 0.5 or below. However, if they have a credence measure like (b), then their credence towards q is not stably high and hence they should not maintain a categorical belief towards it. That is because although their credence towards q is above 0.5, there is a proposition in their space of possibilities r which is consistent with q ($P(q\&r) \neq 0$) and upon learning (conditionalizing on) r their credence towards q drops to 0.5 or below ($P(q|r) = 0.5 \ngtr 0.5$).



(a) Space of possibilities in which believing q is rational

(b) Space of possibilities in which believing q is irrational

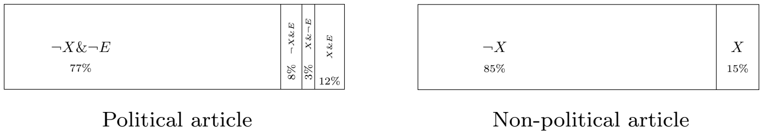
Figure 3: Degree of fine-graining affects the rationality of belief

Now, let's see whether my credence towards the claims of the articles discussed in the previous section is stably high in Leitgeb's framework. My prior towards the truth of the main claims of both articles ($\neg X$) is 0.85. We have seen that given the collection of comparative probabilistic judgments, I should not fine-grain my space of possibilities with respect to E for the non-political article. So, my space of possibilities looks like the right rectangle in Figure 4. In this space my high credence towards X is stable. That is because there is no proposition consistent with X, conditionalizing on which drops my credence to 0.5 or below[21].

However, in the case of the political article, I have to fine-grain my space of possibilities with respect to E and maintain a precise credence over this new more fine-grained space. Obviously, there is no unique way of forming this new credence function. The left rectangle in Figure 4 is just an example of a credence function that respects the prior credence towards X and the collection of probabilistic judgements presented in the previous section. It is easy to see that given this space of possibilities, my high credence towards $\neg X$ is not stable. That is because there is proposition E which is

in the literature. For instance, Gunther [28] employs this aspect of Leitgeb's framework when arguing against the preference for statistical evidence over testimonial evidence in legal proceedings.

[21] One may raise the objection that in the case of non-political articles, although the distinction with respect to E is not present in the partition for precise credence function, it exists in the partition for confidence order. Thus, if one has the comparative probabilistic judgment $P(\neg X|E) < 0.5$, then $P(\neg X)$ is also unstable in the case of non-political articles. The response to this objection is that we assumed the agent does *not* know that E is decision-changing. Given that the threshold for a decision-changing posterior was 0.5, this implies that the agent does *not* have the comparative probabilistic judgment that $P(\neg X|E) < 0.5$.



Political article

Non-political article

Figure 4: Space of possibilities while contemplating resharing an article

consistent with $\neg X$ ($P(\neg X\&E) \neq 0$) and conditionalizing on which drops my credence towards $\neg X$ below 0.5: $P(\neg X|E) = 0.4$.

So, in Leitgeb's framework, maintaining a categorical belief towards the main claim of the non-political article is rational while it is irrational for the main claim of the political article. Given the partition-dependency of probabilistic accounts of belief in general[22] (including Lin and Kelly [42] and Goodman and Salow [25]), I believe that this result is not limited to Leitgeb's framework; rather, it can be produced in other frameworks as well.

Conclusion

In this paper, I imagined a bounded Bayesian agent who is about to make a decision under uncertainty about X and a piece of evidence E (relevant to X) becomes salient to them and they wonder whether it is instrumentally rational for them to fine-grain their space of possibilities with respect to E given their limited cognitive resources. I showed that they should fine-grain with respect to E if and only if they deem the possibility of learning E and changing their mind about the decision they were about to make, suffiently likely. The sufficiency of this likelihood depends on the cost of the experiment and fine-graining. I also showed that whether they decide to fine-grain with respect to E affects the rationality of believing X if they use a probabilistic account of belief like Leitgeb [39]'s.

I then applied this result to compare peer-reviewed articles related or unrelated to a current political controversy. Under certain plausible assumptions regarding the cost and outcomes of conducting an experiment to identify red flags in the reputations of authors and funding sources associated with these articles, I demonstrated that we are rationally compelled to fine-grain our space of possibilities with respect to the outcome of this experiment for political articles, but not for non-political ones. By employing Leitgeb's stability thesis, this suggests that there are scenarios where, despite maintaining equally high credence in the main claims of two articles, we should rationally believe one and not the other.

An important question is whether my result can be exploited by conspiracy theorists. What if a conspiracy theorist can find a way to relate the claim of any article to a political controversy? Does that mean we should not believe the claim of any article? I do not think that is true in my account. That is because, for us to be rational

[22] Wang [62] provides a good summary of some of these accounts.

in fine-graining our space of possibilities with respect to a possible piece of evidence and potentially destabilizing our credence towards the target proposition, we need to have a *reason* to deem the possibility of receiving that evidence and changing our mind about the decision we were about to make based on that article, sufficiently likely[23]. If an article's claim is related to a political controversy through a conspiracy theory, the burden of argument is on the conspiracy theorist to convince us that the experiment of looking for red flags in the reputation of authors and funding sources via their reference list of red flags (identified by their conspiracy theory) has a *decision-changing* outcome. And I believe that burden is quite heavy!

Another related important question is: what if we are uncertain whether an article's claim is related to a political controversy? How should we factor in that uncertainty in my model? I believe that uncertainty can be addressed by considering our uncertainty about whether running the experiment to check for the existence of red flags is decision-changing. The more uncertain we are about whether an article's claim is related to a political controversy, the more uncertain we are about whether looking for red flags in the reputation of authors and funding sources is decision-changing. Consequently, it becomes more difficult for us to justify that the instrumental norm of fine-graining is fulfilled, and thus more difficult to justify the rationality of fine-graining.

Finally, I believe my account can be compatible with accounts like that of Friedman [18], which advocates the view that inquiry excludes rational belief. This is because I have shown that as long as we are about to conduct an experiment with a potentially decision-making outcome which is sufficiently likely, we are rationally obliged to generate/maintain a distinction with respect to that outcome in our space of possibilities. If we consider an inquiry as a collection of experiments, some of which possess the characteristics I specified, then as long as one is engaged in an inquiry, one's space of possibilities is too fine-grained to afford a stable credence towards the proposition that identifies the subject of inquiry for us. Therefore, using a probabilistic account of belief like Leitgeb's, maintaining a categorical belief towards that inquiry proposition is irrational. I believe there is a scope for further investigation of the relation between the notion of *inquiry* in the literature on zetetic epistemology and the notion of *experiment* in the literature on evidence-gathering.

Appendix

Proof of a generalized version of inequality (1):

In this section, I will generalize the result that we got from Fatima's example. Suppose we face the following macro-decision in which $H_1 > L_1$ and $H_2 > L_2$ and our prior for X is low enough to make us choose A_2.

[23] Subjective Bayesians may deny that there are any such constraints, but most theorists think there are at least some constraints on permissible degrees of belief given a fixed body of evidence [65],[56].

		X	$\neg X$
	A_1	H_1	L_2
→	A_2	L_1	H_2

Macro-decision M

Now suppose E is the outcome of an experiment. Doing some algebraic gymnastics on the expected utilities of the above table, it can be shown that E is decision-changing wrt M iff

$$P(X|E\&D) > \frac{1}{\dfrac{H_1 - L_1}{H_2 - L_2} + 1} \tag{2.3}$$

It has been discusssed in the paper that if a bounded and instrumentally rational agent considers the decisions of fine-graining their domain of precise credence function wrt E and doing an experiment with the result E simultaneously, while they do not know whether E is decision-changing, they face the following decision table:

	$\boldsymbol{D}$: E is decision-changing				$\neg\boldsymbol{D}$: E is not decision-changing
$\boldsymbol{F}$: Fine-grain and do the experiment	$\boldsymbol{X\&E}$	$\neg\boldsymbol{X\&E}$	$\boldsymbol{X\&\neg E}$	$\neg\boldsymbol{X\&\neg E}$	$-c$
	$H_1 - c$	$L_2 - c$	$L_1 - c$	$H_2 - c$	
$\neg\boldsymbol{F}$: Neither fine-grain nor do the experiment	L_1	H_2	L_1	H_2	0

It can be proved $EU(F) > EU(\neg F)$ *if and only if*

$$P(E\&D) > \frac{\dfrac{c}{H_2 - L_2}}{P(X|E\&D)(\dfrac{H_1 - L_1}{H_2 - L_2} + 1) - 1} \tag{2.4}$$

Here is the proof:

$EU(F) > EU(\neg F)$ iff

$$\begin{aligned}
&P(D)\,[P(X\&E|D)(H_1 - c) + P(\neg X\&E|D)(L_2 - c)\\
&\quad + P(X\&\neg E|D)(L_1 - c) + P(\neg X\&\neg E|D)(H_2 - c)]\\
&\quad + P(\neg D)(-c) >\\
&P(D)\,[P(X\&E|D)L_1 + P(\neg X\&E|D)H_2\\
&\quad + P(X\&\neg E|D)L_1 + P(\neg X\&\neg E|D)H_2]
\end{aligned}$$

iff

$$P(D)[-c + P(X\&E|D)(H_1 - L_1) + P(\neg X\&E|D)(L_2 - H_2)] + P(\neg D)(-c) > 0$$

iff

$$-c + P(D)[P(X\&E|D)(H_1 - L_1) + P(\neg X\&E|D)(L_2 - H_2)] > 0$$

Assuming $P(D) \neq 0$ the above inequality holds iff

$$P(X\&E|D)(H_1 - L_1) - P(\neg X\&E|D)(H_2 - L_2) > \frac{c}{P(D)}$$

Assuming $P(E|D) \neq 0$ the above inequality holds iff

$$P(X|E\&D)(H_1 - L_1) - P(\neg X|E\&D)(H_2 - L_2) > \frac{c}{P(D)P(E|D)}$$

iff

$$P(X|E\&D)(H_1 - L_1) - (1 - P(X|E\&D))(H_2 - L_2) > \frac{c}{P(E\&D)}$$

Assuming $(H_2 - L_2) \neq 0$ the above inequality holds iff

$$P(X|E\&D)[\frac{H_1 - L_1}{H_2 - L_2} + 1] - 1 > \frac{\frac{c}{H_2 - L_2}}{P(E\&D)}$$

Given what is discussed at the beginning of this section, we know that if E is decision-changing the inequality (3) holds i.e. $P(X|E\&D) > \dfrac{1}{\frac{H_1 - L_1}{H_2 - L_2} + 1}$. Therefore, $P(X|E\&D)[\frac{H_1 - L_1}{H_2 - L_2} + 1] - 1 > 0$ and hence we can divide both sides of the above inequality by it and we will have that it holds iff

$$P(E\&D) > \frac{\frac{c}{H_2 - L_2}}{P(X|E\&D)[\frac{H_1 - L_1}{H_2 - L_2} + 1] - 1}$$

Step-by-step reiteration of the argument for the claim in section 1.3:

For non-political articles: $P(X|E\&D) > 0.5$. That is because given the utilities of the macro-decision, the threshold of the posterior after receiving *a decision-changing evidence* ($E\&D$) is 0.5. Assuming that I only hold precise credences to one decimal point, $P(X|E\&D) \geqslant 0.6$. This implies that $\dfrac{0.001}{2 * P(X|E\&D) - 1} < 0.005$. On the other hand, $P(E\&D) > 0.1$ and $0.1 > 0.005$, therefore,

$$P(E\&D) > \frac{0.001}{2 * P(X|E\&D) - 1}$$

For political articles: $P(X|E\&D) \leqslant 1$, therefore $2 * P(X|E\&D) - 1 < 1$ and $\dfrac{0.001}{2 * P(X|E\&D) - 1} > 0.001$. On the other hand $P(E\&D) < 0.001$. Therefore,

$$P(E\&D) < \frac{0.001}{2 * P(X|E\&D) - 1}$$

Chapter 3

Mind Your Probability Language

Abstract

The notion of probability has multiple interpretations, and one of these interpretations is the propensity interpretation. Given this possible interpretation and the opaque causal structure of the social world, it can be argued that when probabilistic statements are used about social groups, they can conventionally implicate essentialist claims about those social groups. Moreover, in decision-making conversational contexts, the propensity interpretation about social groups can conversationally suggest interventions that are aligned with oppressive social practices. These implicatures render statistical generalizations about social groups vulnerable to exploitation and misinterpretation, potentially perpetuating social injustice. This paper scrutinizes the pragmatics of probabilistic statements in relation to oppressive social practices. It also outlines some strategies to minimize the chance of exploitation and misunderstanding.

Introduction

Much has been written about the harms of promoting essentialist claims about social groups through the use of generic statements [41]. Haslanger [30] discusses the pragmatics of generic statements about social groups and how, given their ambiguity and the obscurity of social structure, they can surreptitiously smuggle unwarranted essentialist claims about social groups into the common ground of conversation, thereby contributing to oppressive social practices. In this paper, I will scrutinize the pragmatics of another linguistic form for expressing generalizations about social groups—probabilistic statements—to examine whether the same dynamics are plausible for them.

I will demonstrate that probabilistic statements, given their ambiguity between multiple interpretations, have the same potential. To establish this, I will first offer an analysis of probabilistic statements about different collections of entities and show that the propensity interpretation is a plausible interpretation of probabilistic statements about social groups and hence can be considered their conventional implicature. Then I will show how this interpretation is equivalent to an essentialist claim. In the next step, I will pinpoint the harm that lies in probabilistic statements about social groups by examining decision-making conversational contexts. I will demonstrate how such statements can conversationally implicate that actions aligned with oppressive practices are the optimal choice. Here is the outline of my argument:

Premise 1: Statistical reports on social regularities contain probabilistic statements about social groups, and these statements carry the conventional implicatures that individual members selected at random from those groups possess an inherent predisposition toward the observed frequencies. (Propensity interpretation)

Premise 2: Given this conventional implicature, uttering probabilistic statements about social groups in decision-making contexts can have the conversational implicature that an action aligned with an oppressive social practice is the optimal choice, even when there is not sufficient evidence to establish it as optimal.

Premise 3: If action A is (i) aligned with an oppressive social practice and (ii) there is not sufficient evidence to establish it as the optimal choice, then uttering statements that has the implicature that A is the optimal choice can be harmful.

Conclusion: Uttering probabilistic statements about social groups can be harmful.

In the final section of the paper, I introduce two strategies to effectively block conventional and conversational implicatures of statistical reports about social groups.

3.1 Pragmatics of Probabilistic Statements

Statistical methods are widely used in science and we often encounter reports like 'An adult man is 14.1% *likely* and an adult woman is 11% likely to smoke cigarettes' or 'The homicide offending *rate* among Black Americans is 10.6 times higher than that of White Americans.'. On the other hand, the notion of probability has multiple interpretations. The first natural question is how we *should* interpret statements that contain terms like 'probability', 'rate', and 'likelihood' in statistical reports. Different interpretations have been discussed in the literature on the philosophy of probability. My focus in this paper will be on the ambiguity of probabilistic statements among the following three interpretations[29]:

1. **Subjective interpretation:** The statement 'event X is p percent probable' means the speaker is p percent certain that event X will happen (has happened).

2. **Frequentist interpretation:** The statement 'event X is p percent probable' means the speaker believes if there were an infinite reference class R of events similar to event X, then the relative frequency of X in R is p.

3. **Propensity interpretation:** The statement 'event X is p percent probable' means the speaker believes event X has a physical disposition which leads to manifesting relative frequency p in a hypothetical infinite reference class of events similar to X.

It is important to note that in the last two interpretations 'the probability of p percent' is regarded as an objective trait (chance) of event X, while in the first interpretation, it is just a measure of the *uncertainty* of the subject who is considering event X. The second interpretation offered above is usually referred to as hypothetical frequentism. A simpler version of it is called finite frequentism which defines probaility of event with respect to a finite reference class and as the relative frequency of occurance of that event in that finite set. My focus in the rest of this paper will be on this simpler version.

When a sentence is ambiguous between multiple meanings, its pragmatics becomes important as well as its semantics. This paper is on pragmatics of probabilistic statements about social groups. By pragmatics, I mean what is not given to us directly but we can *infer* from the sentence given the contextual clues. These clues can come from lexical or syntactical features of the sentence or the context of conversation in which it has been uttered. When the clues are from the feature of the sentence itself, the implicature is conventional and when they are from the conversational context, they are conversational. In what follows I will first argue that probabilistic statements about social groups have the conventional implicature of the propensity interpretation. Then I will show how this propensity interpretation can be translated into essentialist claims and causal claims about social groups.

Let's first review different interpretations of the notion of probability in statistical reports through an example. Suppose I am walking and talking with my friend, Hesam, in the park. Hesam is a hobbyist physicist and a coin collector. He collects coins and has instruments to measure how the mass of a coin is distributed (whether its density is uniform or not). As we are walking, he suddenly takes a dime out of his pocket and says:

> If flipped, it is 50 percent likely that this dime lands heads.

Given what I know about Hesam, I cannot say which of the following propositions is meant by the above statement:

1. Hesam has no information about this coin and he is simply reporting his uncertainty (potential betting behaviour) towards this coin landing heads. (subjective Bayesian interpretation)
2. Hesam has flipped this specific dime several times and 50 percent of times it turned heads. (frequency interpretation)
3. Hesam has measured the mass density of the coin through his instruments and is convinced that the coin has a uniform density and so it has the same *propensity* for landing heads and tail. (propensity interpretation)

So, I ask him which one of the above propositions he meant. He responds:

> Neither! I just recently did a statistical study on the collection of my coins. I have 1000 coins. One day, I selected 100 of them randomly and checked whether they are fair coins. It turned out that all 100 of them were fair. The confidence level of my test was 90%. This coin is just one of my 1000 coins.

After hearing his explanation, my understanding is that Hesam is saying it is 90% likely that a coin in his collection is fair. That is to say, given the confidence level of the test, if Hesam repeats sampling and testing the fairness of sampled coins several times, 90% of the times, all the coins in his sample are fair. So, I interpret Hesam's sentence as:

> It is 90% *likely* that the dime in Hesam's hand is fair.

That does not really help! It only added another layer of probability ('90% likely') to the picture! I am still confused what Hesam means be 'being fair'. To be more precise, I am confused between the following interpretations:

1. Hesam tested the fairness of a sample of 100 out of 1000 coins by flipping each coin in the sample several times.
2. Hesam tested the fairness of a sample of 100 out of 1000 coins by measuring mass density of each coin in the sample.

So, I ask him and he says that he meant the first interpretation i.e. that he did not do any density measurement and only did several flipping of the coins in the sample. So, I infer the *correct interpretation* is:

> It is 90% *likely* that if the dime in Hesam's hand is flipped several times then it will land heads 50% of the times.

However, given *my knowledge of physics*, I am quite certain that a coin which has landed heads 50% of the time through several flips has a uniform mass density. So, although I know, strictly speaking, this is not what Hesam meant. I can safely *infer*:

> It is 90% *likely* that the dime in Hesam's hand has a uniform mass density.

Now, I start worrying about that extra layer of probability and ask myself what '90% likely' means in the above interpretation. Is it chance or Hesam's uncertainty as an experimenter? Is it objective or subjective? The answer is complicated and depends on the sampling process and whether it can be modeled as independent observations. I will return to this issue later. For now, let's assume that the sampling process is truly random and repeatable and hence can be modeled as independent observations. So, we can interpret 90% as objective chance and infer that 'almost certainly the coin in Hesam's hand has uniform density'.

Let's review what the steps of my inference were. The main point of the above reasoning was that I started from the frequency interpretation and ended up with the propensity interpretation. That is to say, I eventually *inferred* something about the physical essence of a random member of a coin collection after receiving the result of a statistical study on that collection and on a trait other than that physical trait. The study is on the trait of 'landing heads 50% of the times upon several flipping' and shows that it is 90% *likely* that a random coin in that collection has that trait. Given my knowledge of physics, I first translated the trait of 'landing heads 50% of the times upon several flipping' into a physical essential trait of 'having uniform density' for the coin. Then, given the repeatability of the study, I interpreted 90% likely as the objective chance of facing this result. Finally, I put my credence equal to this 90% objective chance and inferred that *I am almost certain that a random coin in Hesam's collection has a uniform density.*

Can we do the same line of reasoning for every statistical study? Let's suppose Hesam has a collection marbles. Some of them are shiny and some are opaque. There are too many of them in his collection (2000) and he cannot bother to count how many are shiny and how many are opaque. So, he does a sampling (100 marbles) from his collection through a truly random process and reports:

> With the confidence level of 90%, it is 80% likely that a random marble from my collection is shiny.

What can I infer from this result? The study is on the proportion of shiny marbles in Hesam's collection. But, given the frequency interpretation I can interpret it as a study on the 'probability of being shiny upon random drawing'. It shows that it is 90% *likely* that a marble in Hesam's collection has the trait of *being* 80% *likely to be shiny upon random drawing.* For this study, I can do the last two steps of the previous reasoning, but I cannot do the first step. That is to say, assuming that Hesam's method of sampling is random and repeatable, I can go ahead and interpret that '90% likely' as an objective chance and then put my credence equal to it and declare that I am almost certain that a marble in Hesam's collection has the trait of 'being 80% likely to be shiny upon random drawing', but that is as far as I can go. I cannot translate 'being 80% likely to be shiny upon random drawing' into any essential physical trait of a marble in Hesam's collection. Unlike the case of coin, in which , I could use my background knowledge in physics to translate 'landing heads 50% of the times upon several flipping' into 'having uniform mass density', here there is no background knowledge that lets me do that kind of translation and inference.

3.1.1 Conventional Implicature: Propensity Interpretation

Nevertheless, if I am a proponent of propensity interpretation, I can go one step further. How? By inferring just the existence of a physical trait underlying the observed frequency without articulating what that trait is i.e. I will infer that although I cannot describe that physical trait in terms of any familiar physical trait, I know that the observed frequency stems from a physical propensity shared by all the marbles in Hesam's collection i.e. *there is an underlying physical propensity shared by all the 2000 marbles in Hesam's collection that makes them manifest this trait of 'being shiny* 80% *of the times upon random drawing'.* The only thing is that with my limited knowledge of the world, I cannot translate this physical propensity into any familiar physical trait.

This is not an intuitive inference. It is hard to infer the existence a (mystical) shared trait among Hesam's marbles that causes this specific (80%) frequency of being shiny upon random drawing (and that is exactly why the propensity interpretation is not popular in the literature). So, I am not claiming that an inference to underlying causal properties will generally go. My claim is that in some contexts our background knowledge of the world will make this interpretation plausible and hence we can argue that the propensity interpretation is the *conventional implicature* of the probabilistic statement in those context.

But what are those contexts? When can we say that the propensity interpretation is a plausible reading of a probabilistic statement? My focus is on scenarios in which we use probabilistic statements to describe the prevalence of a trait in a population:

- **Probabilistic statement:** A random member of population C has a p percent likelihood of having trait A.
- **Propensity interpretation 1:** A random member of population C possesses an inherent predisposition toward the observed relative frequency of p.
- **Propensity Interpretation 2:** There is a shared trait among the members of population C that causes (and explains) the manifestation of the observed frequency p.

The two formulations of the propensity interpretation are equivalent. I included both because each is useful for a specific claim I will make later. Given the above notation, here is my claim to which I will henceforth refer as CI (Conventional Interpretation) claim:

CI claim: The propensity interpretation is the conventional implicature of the probabilistic statement if and only if EITHER (1) there is a well-established scientific theory that explains the relationship between being a member of C and manifesting the relative frequency p in that population, OR (2) no such theory exists, but 'being a member of C' is considered to have explanatory power in the scientific discourse related to trait A.

Let's see whether this claim holds true for the cases that we have seen so far. In the case of coin collection, the propensity interpretation was plausible (and hence a conventional implicature of the probabilistic statement) because physics explains the causal relationship between being a coin and manifesting the relative frequency of 50% through uniform mass density. However, in the example of marble collection, there is no such theory, and 'being a marble' has no explanatory power in the scientific discussions related to 'being shiny'. This is because the scientific discussions related to 'being shiny' fall under optics, wherein it is clear 'being a marble' has no explanatory power in this context.

Given the above examples, it is not hard to see the reason behind the first disjunct of CI claim i.e. why the existence of a theory that explains the relationship between 'being a member of C' and 'manifesting the relative frequency p' makes the propensity interpretation plausible. What is more difficult to grasp is the reason behind the second disjunct of CI claim i.e. why if 'being a member of C' has an explanatory power in a relevant scientific field then then it makes the propensity interpretation plausible.

I think the best way to see it is through scrutinizing the second formulation of the propensity interpretation. That formulation declares the existences of a trait shared among all members of C that causes and explains the observed relative frequency of

A. So, we are expecting 'being a member of C' to have an explanatory power. In the absence of a clear theory that explain where that explanatory power comes from, it is not plausible to us that 'being a member of C' has such an explanatory power. However, if being a member of C plays a role in the scientific theories relevant to the trait A, then it becomes plausible to us that it may be able to explain the observed relative frequency of A.

Let's see this through an example of a probabilistic statement about a social group:

> The homicide offending *rate* among Black Americans is 10.6 times higher than that of White Americans.

What is the propensity interpretation of this probabilistic statement:

- **Propensity interpretation 1:** A random Black American has an inherent predisposition toward being 10.6 times more frequently involved in committing homicide compared to a random White American.
- **Propensity Interpretation 2:** There is a shared trait among Black Americans that causes (and explains) their involvement in committing homicide 10.6 times more frequently than White Americans.

Are these interpretations plausible? I think the answer is yes. That is because the second disjunct of the above claim holds true i.e. while there is no widely-accepted scientific theory that explains why being a Black American causes more involvement in committing homicide, being a member of an ethnic group like Black American is widely used in *social sciences* which is the field in which causes and prevalence of crimes like homicide is discussed. So, given that Black Americans is considered a social group which is widely used in theorizing in social sciences, 'being a Black American' has an explanatory power and therefore the propensity interpretation is plausible to us. That is why (both of) the above propensity interpretations are considered conventional implicatures of the probabilistic statement.

I believe we can generalize this point by saying that the propensity interpretation of any probabilistic statement about *social groups* is also its conventional implicature. This is because the second disjunct of the CI claim holds true for probabilistic claims about social groups. Why? Because the causal structure of the social world is opaque. Unlike the natural sciences, robust causal theories in the social sciences are rare, and the available ones rarely apply to groups as large as social groups. If we define a social group as one that bestows a social identity—such as race, gender, religion, etc.—on its members, then, given the strong bond between the members of a social group, they are often treated as a whole in social science theorizing. So, being a member of a social group like Black Americans has some explanatory power, even though no robust causal theory is available.

If a collection of people does not constitute a social group, reporting the statistical prevalence of a socially-relevant trait among them does not have such conventional implicature. For example for the statement 'the rate of speeding violations among sedan drivers is more than SUV drivers', the propensity interpretation is not plausible, and hence is not conventionally implicated. That is because sedan drivers and SUV divers are not social groups and speeding violation is a socially relevant trait.

The next question is: why should we care about this conventional implicature of probabilistic statements about social groups? Why should we care that they imply that a random member of a social group possesses an inherent predisposition toward an observed relative frequency? The reason is that many probabilistic statements about social groups concern the prevalence of a pernicious trait, such as being involved in committing homicide, within a social group. To imply that a random Black American possesses an inherent predisposition toward a higher frequency of involvement in homicide is to imply an essentialist claim about an ethnicity, which is harmful to social discourse, especially in a society known for its vast discrimination and stigma against that ethnicity.

But one can question the harm of essentialist claims about social groups by saying that they are just about the essence of a social construction and not a biological or cultural entity. Saying that a Black American possesses an inherent predisposition toward a higher frequency of involvement in homicide does not mean that that predisposition is inherent to their biology or culture. It just means it is inherent to their social identity which is shaped by historical injustice and discrimination against them. But social identities can change by changing our social practice. So, saying that a pernicious trait is essential to a social identity does not imply that it cannot be removed from that social identity.

To meet this objection, I will switch to the second formulation of the propensity interpretation which is a causal claim and will argue that in policy-making conversations, propensity interpretation, and thereby probabilistic statements, can *conversationally implicate* rationalizing actions which are aligned with oppressive social practices.

3.1.2 Conversational Implicature: Optimal Intervention

Let's consider an imaginary statistical A study shows that 'the rate of major crimes among immigrants in neighborhood X is higher than that of non-immigrants.' Immigrants are a social group. So, according to the analysis in the previous section, the propensity interpretation is the conventional implicature of this statistical report. According to the first formulation of the propensity interpretation, this report implicates that immigrants in neighborhood X have an inherent predisposition toward being more frequently involved in crime compared to non-immigrants. This is an essentialist claim about immigrants in neighborhood X, which attributes a pernicious trait to them and is therefore harmful. However, we have seen that an objector may question the harm of uttering this claim by arguing that the essence of a social group is at least partly shaped by social practices surrounding that social group. Therefore, attributing a pernicious trait to the essence of a social group is not necessarily vilifying their common biology or culture; it can also be interpreted as blaming the harmful social practices surrounding that social group.

In order to meet this objection, I will show that the harm of the propensity interpretation of the probabilistic statements about social groups is not just polluting the social discourse by conveying abstract essentialist claims about social group. The harm is in fact rationalizing actions which are aligned with oppressive social practices through conversational implicature.

In what follows I will present some examples of conversation to illustrate my point. In discussing these examples, I assume the interlocutors are trying to be efficient in their communication. So, in fact I assume they follow Grice [27]'s general cooperative principle according to which in a conversation one should only contribute what is required by their accepted goal of the conversation. This principle is accompanied by four maxims: (i) *Maxim of Quality:* one should only contribute what they believe is true; (ii) *Maxim of Quantity:* one should be as informative as required; (iii) *Maxim of Relation:* one should only contribute relevant information; (iv) *Maxim of Manner:* one should be as clear as required [16]. So, in Grice's framework

> An utterance U from speaker S conversationally implicate proposition p if and only if an interlocutor cannot make sense of U as *cooperative* without assuming proposition p to be believed by S.

As the first example suppose the following conversation happening in a city council:

Criminal Immigrants [City Council]

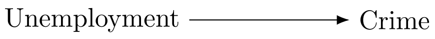
H: The rate of major crimes in neighborhood X is rising.
S: So is the rate of immigrant residents!

The context of this conversation which is happening at a city council conveys that H is concerned about neighborhood X and is trying to find the best explanation for the increasing crime rate to make the best intervention to fight it. He is asking S to help him in this decision-making. So, the purpose of conversation is to find the best intervention to decrease (or at least stop the increase of) crime rate in that neighbourhood. In response, B reports a statistical correlation between the population of immigrants and the rate of crime. I claim that in this context what B says *conversationally implicates* that 'immigrants are the cause of the increase in the crime rate and the best intervention is controlling their population in the neighbourhood'.

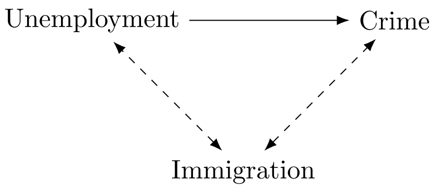
Why? Because first given the maxim of relation, the reported correlation should be relevant to the purpose of conversation which is identifying the optimal intervention on the rate of major crimes in that neighbourhood. On the other hand, in order to find the optimal intervention H needs to first identify the most plausible causal model of the situation in that neighborhood. So, the reported correlation is only relevant if it helps H in that identification. This mean S's report of that statistical correlation conversationally implicate something about the *causal* structure of the crime situation in neighbourhood X.

But what is that implicature? In order to answer this question we need to review how one can contribute to identification of a causal model? There are two ways to make that contribution: (i) by introducing variables (nodes) into the model and (ii) by identifying the existence and direction of causal relations (arrows) between the variables. If what S utters did not have a propensity interpretation, then its only contribution would be introducing the variable immigration into the causal model. However, given the analysis in the previous section and the fact that immigrants constitute a social group, the propensity interpretation is plausible for this statistical report. What does the propensity interpretation imply for the causal model? It implies that there is a trait shared among immigrants that causes manifesting a higher frequency of crime. This implies the more immigrants in a neighbourhood, the more prevalent that trait thar causes high frequency of crime and hence the more the frequency of crime. So, the propensity interpretation implies that there is a causal path from immigration to crime rate. This path can be direct or indirect.

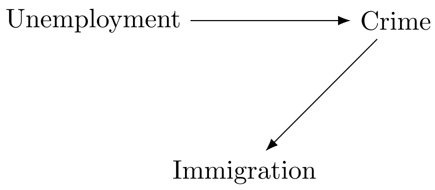
Suppose before this conversation H's causal model was that 'unemployment causes crime'.



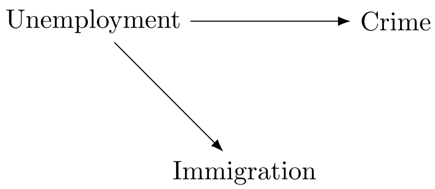
After S mentions that statistical correlation, he adds the variable 'immigration' to his model. He can make eight different models using this new variable and different combinations of causal relations[1].

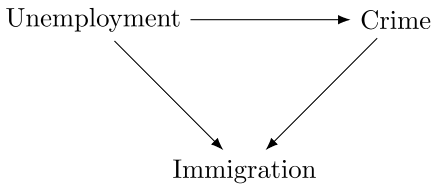


However, given the implicated propensity interpretation those models in which immigration has no causal effect (direct or indirect) on crime will be eliminated. That is because, according to the (second formulation of the) propensity interpretation there is a shared trait among immigrants that causes their involvement in crime. So, the propensity interpretation implies *elimination of the following three models* from the collection of 8 models.



[1] I assume in the context of that neighborhood crime causing immigration and immigration causing crime are both plausible. Similarly, unemployment causing immigration and immigration causing unemployment are both plausible.





In the remaining 5 models, there is a direct or indirect causal path from immigration to crime. So, in all of them, intervening on the immigration rate is effective, but the question is whether it is also optimal in all cases, given that in some of these five models, unemployment is a more proximal and therefore more stable cause. If we assume that it is common knowledge between H and S that (i) doing an intervention on unemployment is more costly than an intervention on immigration and (ii) the cost of intervention is more important than its stability across different contexts, then S's utterance suggests that they believe intervening on immigration is not only effective but also the *optimal* option in all five models. Thus, the conversational implicature of S's statistical report is that 'intervening on immigration is the optimal policy'.

This implicature is unwarranted given the available evidence. An observed correlation is insufficient to establish causation or justify an optimal intervention. Moreover, it aligns with oppressive social practices against immigrant which leave them vulnerable to discrimination and exploitation. Therefore, presenting that statistical report about immigrants in that conversation effectively endorses an unwarranted oppressive social practice as the optimal course of action, which is harmful.

Let's consider another example. H and S are in the management team of a firm and are discussing ways to improve promotion rate in their firm. The common knowledge is that the majority of employees in their firm are women.

Submissive Women [Board Meeting]

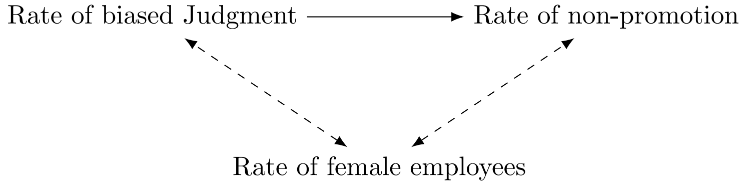
H: The rate of promotion is quite low in our firm.

S: Did you know that, statistically, most women are submissive?

Again the context of conversation shows that H is concerned about the current state of promotion rate and looking for the best intervention to improve it. Suppose, before the conversation, H has the following simple causal structure in mind:

Rate of biased judgment → Rate of non-promotion

S's utterence is *cooperative* only if it helps H in finding the best course of action. Given that S's statistical report is about women, it introduces a node (rate of female employees) into H's causal model[2]:



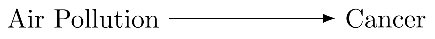
Also, given that women constitute a social group, the statistical report carries a propensity interpretation suggesting that 'women have a shared trait that *causes* being submissive (and not asking for promotion) with high frequency'. This implies that only causal models with a direct or indirect path from the node 'rate of female employees' to 'rate of non-promotion' are suitable for identifying the optimal intervention. Assuming it is common knowledge between S and H that intervening on the bias rate in the firm is more complicated and costly than addressing the rate of female employees, S's statement suggests that the optimal point of intervention is controlling the rate of female employees. This suggestion is not only unwarranted based on the available evidence in the hypothetical scenario, but it also aligns with the historically oppressive practice of excluding women from competitive workplaces. Therefore, while S's statement may appear as an innocent report of statistical correlation, it carries a harmful conversational implicature.

Finally, let's examine a conversation in which the statistical report does not carry a propensity interpretation or a harmful conversational implicature.

Cancer and Asthma [City Council]

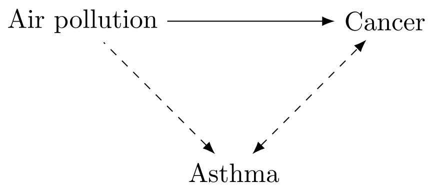
H: The rate of cancer diagnosis is raising in this neighborhood.
S: So is the rate of Asthma!

In this conversation, which is happening in a city council, H is concerned about the rate of cancer diagnosis in a neighborhood and is looking for the best strategy to decrease it. Suppose, before the conversation, he has the following simple conjecture about the causal structure of the situation:



S's utterence is cooperative only if it helps H in identifying the causal structure and finding the optimal strategy. His statistical report adds the node 'Lung disease' to his model. Also, assume that it is common knowledge between them that lung disease cannot cause air pollution. However, there is no common knowledge about the causal effect of asthma and cancer on each other (for example they may suspect that chronic lung disease can contribute to the chance of developing lung cancer).

[2] I assume that, in the context of that firm, both the bias rate affecting the female employees' rate (by influencing the firm's reputation and the number of female applicants) and the female employees' rate affecting the bias rate (due to societal bias against women's competency) are both plausible. Similarly, the non-promotion rate affecting the female employees' rate and the female employees' rate causing the non-promotion rate are both plausible.



The important point is that S's statistical report does not have the conventional implicature of propensity interpretation here. That is because propensity interpretation is not plausible for it. Reviewing CI claim can help us to see this point. There is no well-established scientific theory that explains the relation between 'having cancer' and 'having lung disease'. Moreover, 'having cancer' is not (at least commonly) used as an explanans in the scientific discourse about asthma. So, 'having cancer' does not have explanatory power in that scientific discourse. This means none of the disjuncts required by CI claim holds true for this statistical report.

3.2 Mitigating Strategies

3.2.1 Interpreting Confidence Level: Drop the Pretense and Nip it in the Bud!

All scientific statistical reports include a confidence level or p-value, depending on whether they report the prevalence of a trait or a causal relationship. It is a sign that a generalization has taken place (from a sample to a population or from a small set of experiments to a bigger set) and it represents how reliable that generalization is. In the examples of Hesam's statistical reports, we have seen that a confidence level of 90% means that if Hesam's sampling is repeated infinitely many times then 90% of the times, the same prevalence will be observed. But can statistical sampling be repeated? If not, what does confidence level mean? Is it an objective chance or the subjective uncertainty judgment of the experimenter? There is no consensus on the answer to this question in the literature. Given this ambiguity and lack of consensus in the literature about the correct interpretation of confidence level and p-value, we need to make a *decision* on how to treat them. Spiegelhalter [58] maintains that the best choice is to treat them as objective chances. That is to say according to Spiegelhalter [58] we will be better off if we pretend statistical reports describes objective facts about the world which are independent from the observer/experimenter. In the following, I will argue against him and show that, given the potential harms of implicatures in statistical reports about social groups, the desirability of such pretense is at least questionable, if not deniable.

Spiegelhalter [58] argues that in many fields, including quantified social and behavioral science, probability of an observation should *not* be interpreted as chance of that observation. That is because the notion of chance is only meaningful for *repeatable* events and observation in social science is happening under so many assumptions that it is often unrepeatable. He believes any application of probability in social sciences involves the subjective judgment of the experimenter. However, he also believes, *pragmatically* speaking, probability should be interpreted as chance.

The example that he offers is a randomized controlled trial on hospitalized patients to test the effect of dexamethasone on treating Covid 19. The experiment shows that:

> Among those on mechanical ventilation, the age-adjusted daily mortality risk was 29% lower in the group allocated dexamethasone compared with the group that received only standard care (95% confidence interval of 19–49%). The P value — the calculated probability of observing such an extreme relative risk, assuming a null hypothesis of no underlying difference in risk — can be calculated to be 0.0001, or 0.01% [58].

To put it shortly, this report is saying that 'assuming the null hypothesis, it is extremely *likely*[3] that dexamethasone reduces mortality risk by 29%'. Spiegelhalter [58] says it is wrong to interpret this result as 'assuming the null hypothesis, the *chance* of reducing mortality risk by 29% is very high'. That is because this interpretation means that if we repeat this experiment infinitely many times, the proportion of times that we get a 29% mortality risk reduction is very high. However, repeating this experiment under exactly the same conditions is impossible. It is also impossible to repeat it under *similar but distinct* conditions such that the observations remain *independent*, i.e., that the time and place of the experiment does not affect its outcome[4].

I do not intend to argue for or against this claim here. The philosophical controversies surrounding the interpretation of probability— both in general and specifically within the domain of social sciences— are vast topics that lie beyond the scope of this paper. What concerns me is Spiegelhalter [58]'s next claim i.e that, although wrong, it is pragmatically acceptable to interpret p-value as chance in social sciences. I will refer to this claim as 'pragmatic approach'. He does not provide any argument to justify this approach. However, he applies de Finetti's exchangeability principle to argue that this approach is mathematically justified in many cases of statistical studies in social sciences.

He argues that according to the exchangeability principle, if the order of a sequence of events does not affect our uncertainty towards each of them then we can act as if those events are independent and each has a true underlying chance. This implies that if an experimenter's judgment of a sequence of repeated experiments is not affected by the order of them, then we can act as if each experiment has a true underlying chance. If we assume that in social science, how the experiment is repeated and in what order does not affect the experimenter's judgment, then we can assume that we can use p-value to calculate the *chance* of the result of experiment being true. So, in

[3] The precise number can be calculated using the p-value.

[4] He believes the best interpretation of the above report is that '*the chance of the experimenter judging* the mortality risk reducing by 29% is very high'. That is to say, if you ask the experimenter to judge infinitely many experiments similar to this experiment, with a very high frequency, they will judge it to be a 29% reduction in mortality risk. He also argues that given de Finetti's exchangeability theorem, this interpretation mathematically implies that in a very high proportion of such occasions (similar judgment), there will be 29% decrease in the mortality risk

the above example we can assume that the chance of mortality risk being reduced by 29% is very high.

If I have to guess Spiegelhalter [58]'s reasons for the desirability of his proposed pragmatic approach to probability, they would be the following two reasons: (i) chance is an objective feature of the world that science, including social sciences, should strive to discover. So, when we interpret the result of statistics in social sciences as the chance of some social phenomena, we treat it as a testable intersubjective proposition. The subjective interpretation does not have that potential. (ii) The chance interpretation makes us less prone to error. That's because, cognitively speaking, it is easier and more natural for us to imagine why it respects the rules of probability theory (Kolmogorov's laws). This is because the whole edifice of probability theory was developed to model situations involving chancy events like gambling tools.

However this pretense i.e. acting as if p-value and confidence-level and therefore numbers reported in statistical reports are all objective chances have some important undesirable outcomes as well. Such pretense opens the way for the propensity interpretation of statistical reports about social groups. This in turn opens the way for conversational implicatures which rationalize actions and interventions aligned with some oppressive social practices. It is not at all obvious that the harm that these implicatures have for the social discourse is outweighed by the aforementioned benefits of treating p-value as an objective chance.

What is evident is that if we care about the societal harms caused by the implicatures of social statistical reports, the first and most effective step is to abandon the pretense that we are discussing an objective chance when referring to the p-values of these reports. This approach preemptively eliminates both the propensity interpretation and its potentially harmful conversational implicatures.

How? Let's review an example we've seen before. Suppose there's a statistical study that reports, with a confidence level of 0.01%, that the rate of homicide offending among Black Americans is 10.6 times higher than that of White Americans. If we treat the confidence level as an indicator of objective chance, then the chance that this relative rate is true is 99.9%. This implies that, according to the principal principle, we should be almost certain that the reported relative frequency is true. Once that reported relative frequency becomes the object of our credence, we are doomed!

That's because we must consider the plausibility of the propensity interpretation, and everything that follows from it. However, if we deny (or suspend judgment on) interpreting the confidence level as an objective chance, we can avoid treating those bare relative frequencies as the immediate object of our credence. In this denial mode, the object of our credence is the confidence level itself (and the reported relative frequency (10.6) is an object of the confidence level function, rather than our credence function). Thus, we are do not have to be epistemically in touch with those relative frequencies, and hence, the propensity interpretation is not immediately plausible, and its harmful implicatures do not immediately follow.

So, the first strategy to mitigate the potential social harms of social statistical reports is to be clear about the ambiguity of the notion of confidence level and drop the pretense that it is an objective chance. This way we can block the conventional implicature of probabilistic statements about social groups and thereby the following possible conversational implicatures about optimal interventions.

3.2.2 Get Your Causal Lens at the Ready

The next question is whether there is a way to block the potentially harmful conversational implicatures without delving into p-values and complex statistical jargon. The most natural strategy is what Haslanger [30] refers to as *meta-linguistic negation.* That is, when we mention a statistical correlation which may be interpreted as making an essentialist or causal claim about a social group, we should explicitly deny that possible interpretation. For example, in conversations about crime rates in a neighborhood, to block the harmful implicature about optimal intervention, S can explicitly say that 'I do not mean to say that immigrants inherently possess a trait that causes their higher frequency of involvement in crime' (denying the propensity interpretation) or 'I do not mean to say that immigration has a causal effect on crime rate'. This assertion undermines the harmful implicature of their previous assertion about the optimal intervention.

The problem with this strategy is that although it is simple, it can be ineffective, especially in contexts with more sophisticated interlocutors. This is because an interlocutor with a nuanced perspective about the essence of social groups may interpret your denial of the essentialist claim as merely denying attribution of the pernicious trait to their biological or cultural essence, rather than their essence as a social construction. Thus, they may still take your claim as suggesting there is an *indirect* causal path between immigration and crime rate. This interpretation, given the available causal structures, still leads to the same harmful implicature about optimal intervention.

The same holds for denying the causal claim simpliciter. An interlocutor with a nuanced perspective on causation may interpret it as denying only a direct causal path (rather than both direct and indirect) and continue interpreting your statistical report as a propensity that is instantiated through indirect causal paths. In the example of immigrants, such an audience, even after meta-linguistic negation of the causal claim, would take your statistical report as suggesting that there is an indirect path between immigration and crime rate and make the same harmful inference about the optimal intervention.

Given this complication about the nature of interlocutors—i.e., how sophisticated they are with respect to the metaphysics of social groups and the notion of causation—the best strategy is to explicitly mention possible alternative causal structures which are consistent with your assertion of statistical correlation yet do not include any direct or indirect *causal path* from membership in that social group to the mentioned pernicious trait. For example, in the case of crime and immigration, the best strategy is saying something like 'In this neighborhood, crime and immigration can be the effects of a common cause, such as a certain type of unemployment!' This way, the interlocutor knows that you did not suggest anything about the direction of arrows in the possible causal structures and merely meant to add a variable to the picture to make a more informed decision.

Concluding Remarks

In this essay I have shown that, unlike most probabilistic statements about natural phenomena for which propensity interpretation seems implausible, probabilistic statements about social groups have a plausible propensity interpretation. That is because this interpretation to be plausible demands a level of opaqueness in the causal structure of reality, described through the probabilistic statement, which the natural world does not afford but the social world does. This plausibility makes the propensity interpretation the conventional implicature of probabilistic statements about social groups.

On the other hand, the propensity interpretation of probabilistic statements about social groups can be taken as essentialist claims about them, uttering which can be harmful for social discourse. I showed where exactly the harm of these claims lies by identifying and analyzing some conversational contexts in which these claims (and the probabilistic statements suggesting them) can be interpreted as rationalizing actions aligned with oppressive social practices. In this analysis, I used Grice [27]'s cooperative principle and showed that introducing such actions as the optimal choice is the conversational implicature of probabilistic statements in the context of those conversations.

Finally, I introduced two general strategies to block this harmful implicature. One was through blocking the conventional implicature and making the propensity interpretation implausible by getting a clear perspective of confidence level interpretation in statistical reports about social groups. The other was through blocking the conversational implicature by offering alternative causal structures that are consistent with the statistical report and yet do not suggest the oppressive action is the optimal choice.

Chapter 4

Choosing the Mind's Mesh

Abstract

In this chapter, I present a refined version of the model that served as the foundation for both Chapter 1 and Chapter 2 and review some of their content. This more developed model helps show how the findings of these two chapters connect to each other.

4.1 What is Fine-graining and Why is it Voluntary?

An ideal Bayesian agent can learn any logically possible proposition and come up with a posterior credence function by conditionalizing. We, bounded Bayesians, however, do not have that luxury. We are bound to a coarse-grained space of possibilities, and that space determines what we can possibly learn. It is helpful to think of our space of possibilities as a net that we throw into the sea of infinitely many possible worlds. The more fine-grained the mesh of the net, the smaller the size of the fishes we can catch. The more fine-grained the mesh of our partition, as bounded Bayesians, the more detailed information we can learn. This mesh in fact determines what information we are *open to learn*.

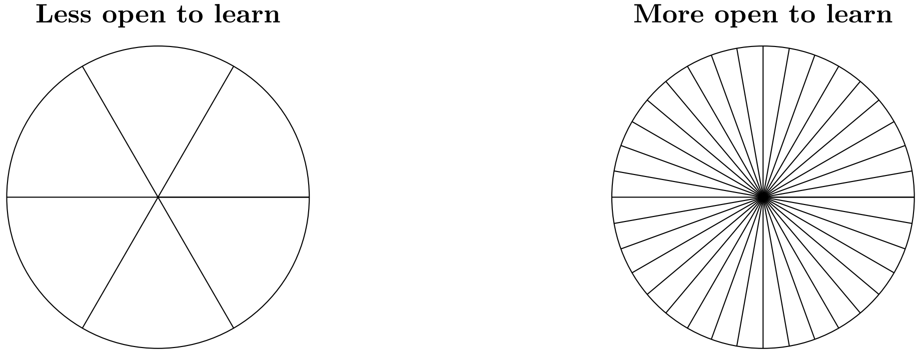


Figure 1: Fine-grained vs coarse-grained spaces of possibilities

But how does the level of fine-graining of this partition change? Is it a voluntary action, or something that happens to us? In this chapter, my focus is on fine-graining the space of possibilities with respect to a proposition which has become salient to us. This task demands evaluating the degree of probabilistic dependence between the newly salient proposition and the existing one while making sure that the resulting function is consistent with probabilism. This demands consulting background knowledge and some mathematical calculation and hence cognitively costly[1]. If we accept that cognitive tasks like estimation and mathematical calculation necessitate some degree or form of volition and intention, it follows that fine-graining the space of

[1] A subjective Bayesian may object by saying that we do not need to consult objective facts when making *prior* judgments about those probabilistic dependencies, as priors do not need to be based on objective facts, but most theorists think there are at least some constraints on permissible conditional priors [65],[56].

possibilities with respect to a newly salient proposition can constitute an action and therefore falls under the norms of instrumental rationality.

Not only, *generating* a new level of fine-graining is cognitively taxing, but also *maintaining* it is costly. A more fine-grained space means being open to learn more and this openness comes with a cognitive cost[2]. That is because with a more fine-grained space of possibilities, not only do we find a greater number of propositions relevant and salient to us, but we also have to pay a higher cognitive cost to integrate those newly salient propositions with the old ones by assessing their probabilistic dependencies, should we decide to fine-grain our space of possibilities with respect to them.

Therefore, as instrumentally rational agents, we need to consider the cognitive costs of generating and maintaining a new level of fine-graining as we decide to whether fine-grain our space of possibilities with respect to a newly salient proposition.

4.2 Micro-decisions of fine-graining

In this paper, I model us as inquisitive agents who are aware of specific decisions under uncertainty that we have to make in the future. To fix notation, I assume we will need to make a future choice between A_1 and A_2, under uncertainty about proposition X. I refer to this decision as the macro-decision C.

	$\boldsymbol{X}$	$\boldsymbol{\neg X}$
$\boldsymbol{A_1}$	H	L
$\rightarrow \boldsymbol{A_2}$	M	M

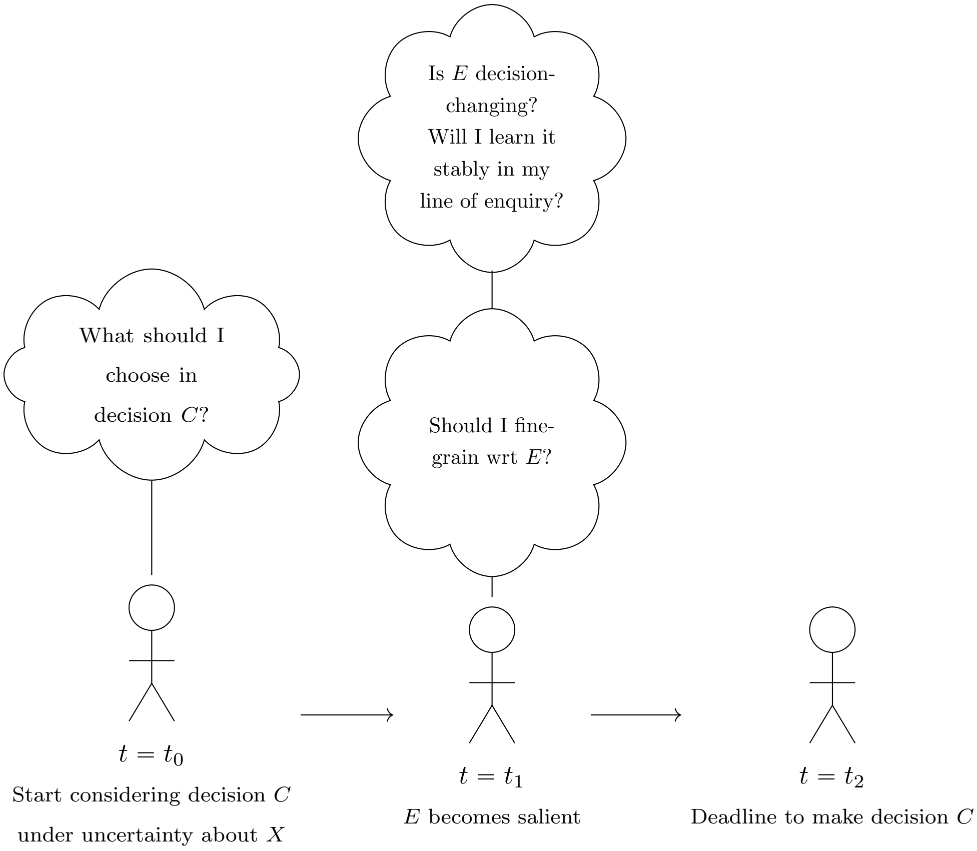
Macro-decision C

In the above decision table, for simplicity, I assumed that the decision has the following distribution of utilities in which H represents a high utility, L a low utility and M a medium utility. I also assumed that our prior towards X is low enough to make us inclined to choose A_2. This inclination is represented by the small arrow on the left hand side of A_2.

We start off at $t = t_0$ with a decision under uncertainty C in mind that we have to make at a later time ($t = t_2$) and a coarse-grained partition over the possible states of that decision i.e. $\{X, \neg X\}$. We have a prior inclination towards one of the options in the decision C based on our prior credence towards X. As time passes and at $t = t_1$ which is before the time we have to actually make the decision, a new proposition like E, relevant to X, become salient to us. We have not learned E yet at this stage.

[2] This aligns with the literature on the computational complexity of probabilistic inference in Bayesian belief networks, which suggests that maintaining numerous distinctions and calculating various conditional probabilities can result in a combinatorial explosion, rendering the task intractable [14], [15].

It is just a possibility that has become salient to us[3]. At this stage, given that our space of possibilities is not fine-grained with respect to E and fine-graining is a matter of volition then as bounded agents, we face the decision of whether fine-grain our space of possibilities with respect to it or not. I will call this cognitive decision the *micro-decision of fine-graining.*


Is E decision-changing?
Will I learn it stably in my line of enquiry?
What should I choose in decision C?
Should I fine-grain wrt E?
$t = t_0$
Start considering decision C under uncertainty about X
$t = t_1$
E becomes salient
$t = t_2$
Deadline to make decision C

This decision is itself a decision under uncertainty which depends on two factors: (i) our view about the rough strength of evidence E and (ii) our vision of our line of enquiry and its learning opportunities. To be more precise, I will show that these micro-decisions are decisions under uncertainty about the following two propositions:

(i) $\boldsymbol{D}$ **:** *Learning E is decision-changing*

- **Informal description:** If I learn E while having the fine-graining with respect to it, then my credence towards X changes to the degree that my initial choice in decision C will change[4].
- **Semi-formal description:** If I were to fine-grain (F), my credence of X conditional on E will be over the threshold credence P^{th} for switching from A_2 to A_1 (this is a counterfactual sentence).

[3] My discussion here sidesteps the problem of awareness growth and its subsequent literature [10], [43], as from the beginning, I assume that the agent's algebra is very coarse-grained and does not include all logical possibilities.

[4] When we do not have the fine-graining in our algebra we can't say whether updating will change our mind. So, we need to talk about the counterfactual scenario wherein we had the fine graining.

- **Formal notation:**
  - $\boldsymbol{D} :$ $F \Box\!\!\rightarrow P_a(X|E) > P^{th}$
  - P_a: actual precise credence function[5]
  - $P^{th} = \dfrac{M - L}{H - L}$

(ii) $\boldsymbol{S}$ **:** *In my line of enquiry, I will learn E and the effect of learning it on decision C is stable*

- **Informal description:** In my line of enquiry contains resources that facilitate learning E and it does *not* contain any resources that facilitate learning propositions that undermine the effect of learning E on my credence of X. That is to say, I will not learn any other proposition that changes my credence of X in a way that will switch my choice in decision M back to my initial choice.

If E is a strong (decision-changing) piece of evidence which we will learn stably in our line of enquiry ($D\&S$) then fine-graining with respect to E will help us achieve better expected utility with respect to the decision C. However, if E is not strong enough to be decision-changing ($\neg D$), then fine-graining with respect to E does not change our expected utility with respect to C and only incurs some cognitive cost. Similarly, if we cannot learn E, then fine-graining with respect to E will be pointless. So, in fact the decision of fine-graining is a decision based on how well-off we are after making the decision C (at time $t = t_2$) and under uncertainty about propositions D and S:

	$\boldsymbol{D\&S}$	$\boldsymbol{D\&\neg S}$	$\boldsymbol{\neg D}$
Fine-grain	Better decision-making with respect to C while undergoing the cognitive labour of fine-graining	No improvement with respect to C while undergoing the cognitive labour of fine-graining	No improvement with respect to C while undergoing the cognitive labour of fine-graining
Don't fine-grain	Status quo	Status quo	Status quo

If we represent the cost of fine-graining (F) with respect to E as C_F and the improvement in our expected utiliy with respect to the decision C after updating upon a piece of evidence which is decision-changing, as ΔEU_C then the following decision table represents the micro-decision of fine-graining:

	$\boldsymbol{D\&S}$	$\boldsymbol{D\&\neg S}$	$\boldsymbol{\neg D}$
$\boldsymbol{F}$	$+\Delta EU_C - C_F$	$-C_F$	$-C_F$
$\boldsymbol{\neg F}$	0	0	0

Micro-decision of fine-graining

It is not hard to show[6] that in the above decision table $EU(F) > EU(\neg F)$ *if and only if* $P(D\&S) > \dfrac{C_F}{\Delta EU_C}$. In English:

[5] I will discuss the difference between the actual and potential precise credence functions later in section 0.3

[6] Please see the appendix for the proof and more details about ΔEU_C

> We should fine-grain our space of possibilities with respect to E if and only if we deem the possibility that learning E is decision-changing—and its effect on decision M is stable in our line of enquiry—to be *sufficiently likely*, i.e., more likely than the ratio determined by the cognitive cost of fine-graining (C_F) and the improvement in the expected utility of the decision (ΔEU_C).

Given that $P(D\&S) < P(D)$ and $P(D\&S) < P(S)$, we can extract the following *necessary conditions* for the rationality fine-graining from the above biconditional:

- **First:** $EU(F) > EU(\neg F)$ *only if* $P(D) > \dfrac{C_F}{\Delta EU_C}$
  - We should fine-grain our space of possibilities with respect to E *only if* we deem the possibility that learning E is decision-changing to be *sufficiently likely*
- **Second:** $EU(F) > EU(\neg F)$ *only if* $P(S) > \dfrac{C_F}{\Delta EU_C}$
  - We should fine-grain our space of possibilities with respect to E *only if* we deem the possibility that the effect of learning E on decision M is stable in our line of enquiry—to be *sufficiently likely*

The first question that comes to mind is that how we can judge whether the possibility $D\&S$ is sufficiently likely when we do not have the fine-graining with respect to it in our space of possibilities. In this model, I assume that a bounded agent can make some rough probabilistic judgments with a negligible cognitive cost. That is because the cognitive cost of making a partial confidence order among a set of propositions is negligible compared to forming a coherence credence function over them. I will discuss this point in detail in the next section.

To sum up, so far I made the following assumptions in this picture:

1. Our enquiry is spanned over a limited period of time which starts at the time we start considering the dicision C under uncertainty about X and ends when we actually make that decision. During this period, new proposition, relevant to X becomes salient to us.

2. The nature of enquiry can be passive or active. So, the cost of learning a proposition can be negligible compared to the cost of fine-graining (C_F) or the improvement in the expected utility (ΔEU_C)

3. We can use a newly salient proposition in the decision C only if we fine-grain our space of possibilities with respect to it.

4. We can have some rough (comparative) probabilistic judgments about propositions with a negligible cognitive cost.

5. We have a vision of our line of enquiry which informs us on the rough likelihood of learning some propositions (S).

6. Given the coarse-grained nature of our algebra and the limited length of our enquiry, the epistemic utility of conditionalization (through improving the accuracy of credence function) is negligible compared to the cost of fine-graining (C_F) and the expected utility improvement with respect to decision C (ΔEU_C).

In the next section, I will provide more details about the 3th assumption.

4.3 Comparative Confidence

Is it possible to make the above cognitive micro-decisions under uncertainty without maintaining a precise credence towards D or S? I believe it is reasonable to assume that we can make some rough probabilistic judgments about D and S with a negligible cognitive cost. To be more precise, it is reasonable to assume that as bounded agents we can maintain a partial[7] confidence order over propositions that are salient to us and have not yet been included in the domain of our credence function (like D and S) with a negligible cognitive cost. Of course, maintaining a coherent partial order over a set of propositions is not cost-free, but it is plausible[8] to assume that its cost is often negligible compared to the cost of maintaining a precise probability function over them. So, if we call the partition over the space of logical possibilities, defined by our attention, salient possibilities, it is plausible to assume that as bounded instrumentally rational agents who are aware of the aforementioned cognitive costs, the partition that we maintain for a partial confidence order is less fine-grained that salient possibilities and more fine-grained than the partition for the precise credence function.

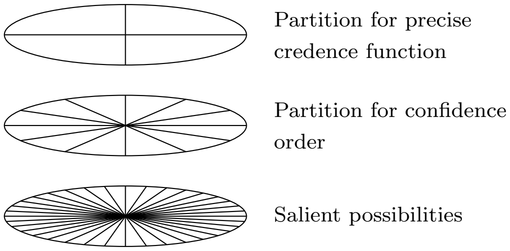

Partition for precise credence function
Partition for confidence order
Salient possibilities

Figure 2: Instrumentally rational agents maintain different partitions over the space of possibilities for different purposes

This assumption is true because the task of *comparing* confidence while observing the coherence constraints over a non-total order is negligibly taxing compared to assigning a precise degree of credence while observing probabilism. For example suppose we are about to make a macro-decision based on our uncertainty about whether a coin lands head or tail: $\{H, T\}$. So H and T are already in the domain of precise probability function and hence has corresponding cells in the partition of both

[7] A partial order is a binary relation which is reflexive, antisymmetric, and transitive.

[8] If the order is total in the sense that every two members of the set is comparable, it is not obvious that cognitive cost of maintaining an order over the set is negligible compared to the cost of forming a probability function over them.

precise credence and confidence order. Now suppose 101 possible biases of the coin towards head is salient to us: $\{B_0, B_{0.01}, ..., B_1\}$ are salient to us. The cognitive cost of fine-graining the space of possibilities with respect to $\{B_{0.1}, B_{0.5}, B_{0.9}\}$ in order to maintain the non-total confidence order[9] $\{(H \succ B_{0.1}), (H \succ B_{0.9}), (B_{0.5} \succ B_{0.9})\}$ is *negligible* compared to fine-graining the space of possibilities with respect to any of $B_{0.1}$, $B_{0.5}$ or $B_{0.9}$ for the purpose of including them in the domain of precise credence function. That is because in order to fine-grain the domain of precise credence function with respect to a proposition like $B_{0.1}$, we need to decide on the precise degree of probabilistic dependence between H and $B_{0.1}$ on one hand, and T and $B_{0.1}$ on the other hand, while respecting probabilism.

It is important to note that I use the notation $P(q)$ to represent the agent's *potential* precise credence for q as opposed to their actual precise credence $P_a(q)$. If they already have the fine-graining with respect to q in their partition for precise credence then $P(q)$ is their actual precise credence for q i.e. $P(q) = P_a(q)$, and if not, $P(q)$ is what their precise credence for q would be if they had the fine-graining with respect to q. It is natural to assume that $P(q) > P(r)$ if and only if $q \succ r$. So, an agent can use their confidence order to make *comparative* judgments about $P(q)$ even when they do not have q in the domain of their precise credence function.

So, one can meaningfully use $P(D)$ and $P(S)$ although D and S are not in the domain of their *actual* precise credence function. Moreover, they can make the judgment $P(D) > \frac{C_F}{\Delta EU_C}$ and $P(S) > \frac{C_F}{\Delta EU_C}$ by referring to their comparative confidence order. I also use the notation $q \succ_s r$ to imply assuming s is true, q stands higher in the agent's confidence order than r. This notation becomes important when it comes to making the above probabilistic judgments[10].

One may object that there is something fishy about including D in the domain of the function P given that D is a proposition about the function P itself. The answer lies in the fact that D is a *counterfactual* proposition about P_a and not P: *If I were to fine-grain (F), then $P_a(X|E)$ will be over the threshold credence P^{th} for switching from A_2 to A_1* ($F \mathrel{\Box\!\!\rightarrow} P_a(X|E) > P^{th}$).

4.4 Fine-graining or Updating?

Roughly speaking, there are two ways that a propositions can become salient to us: with a high probability and with low probability.

In the first scenario, E becomes salient to us with such a high probability that we can declare it as learned. In such a scenario, our dilemma is whether it is worth undergoing the cognitive labor of fine-graining to let our credence of X become updated on E and affect decision C. In such occasions, *the fine-graining decision is in fact an updating decision*, and these are occasions that I discuss in the first chapter. I evaluate whether it is worth updating our credence towards 'a random member of a population

[9] I use the notation $q \succ r$ to represent that q stands higher in the agent's confidence order than r.
[10] Please see the appendix for more details.

having a trait' (X) upon 'the report of a statistical study about the prevalence of that trait in that population' (E), given the cognitive cost of fine-graining. I will show that they are rational to do so only if they are sufficiently confident that that statistical report will *change their mind* about the decision they are going to make about that random person ($P(D) > \frac{C_F}{\Delta EU_C}$). Then I will use Babic et al. [2]'s account of updating on noisy statistical evidence to show that if the statistical report is aligned with a stereotype, then a statistically sophisticated agent, i.e. an agent who can engage with the details of statistical method and more specifically Bayesian statistics, cannot be sufficiently sure that it is decision-changing and hence is not going to update their credence of X on E.

In the second scenario a proposition E becomes salient to us with not very high probability. So, it is just a possibility and we do not want to declare it as learned yet and we do not care about updating our credence of X on it yet. But we do care about the possibility of updating our credence of X on it in case it becomes salient to us with enough certainty later. So, our dilemma is whether we should get the domain of our credence function ready by fine-graining it with respect to E to pave the way for our credence of X getting updated on E in case we learn E. Again this decision has to be made based on our probabilistic judgment of D (whether learning E is decision-changing) and S (whether we can learn E in our line of enquiry). This is what I discuss in chapter 2. In that chapter X is the *content* of a statistical report like '40% of immigrants in this neighbourhood have been involved in crime' and E is the proposition that 'the statistical study is related (funded/affiliated) to some political bodies'. So, I am making a decision based on X later (like whether I should share it on social media) and I have not yet learned E. It is just a possibility. My micro-decision is whether to open up the algebra of my credence function to learning this possibility by fine-graining that algebra with respect to it. As was shown in the model, I argue that we should fine-grain only if we are sufficiently confident that we can learn such association in our line of enquiry ($P(S) > \frac{C_F}{\Delta EU_C}$). I will show that if the statistical report provides support for one side of a major political controversy, then we have reason to be sufficiently confident that we will learn E in our line of enquiry and therefore we should fine-grain with respect to it. Then I will use a probabilistic account of belief [39] to show that this fine-graining undermines rationality of maintaining *categorical belief* towards the content of that statistical report.

So, the Chapter 1 shows if you are a statistically sophisticated agent you will not update your credence function upon a statistical report that is aligned with a social stereotype. And Chapter 2 shows if you are a socially aware agent you will suspend judgment[11] on the *content* of a statistical study which provides support for one side of a political controversey.

[11] By suspending judgment, I mean not including it in the set of categorical beliefs.

4.5 Reviewing Chapter 1 and Chapter 2 and Their Relation

I wrote the first two chapters of this thesis as independent papers. Although both were grounded in similar intuitions, I initially failed to recognize their connection. It was only after simplifying the model presented in this chapter that I realized they are based on exactly the same framework. As a result, their findings are generalizable to one another and can be seamlessly merged.

I will illustrate this relation here with an example. Consider the following propositions:

- ***T***: The statistical study Y shows that 40% of immigrants in this neighborhood have been involved in crime.
- ***N***: 40% of immigrants in this neighborhood have been involved in crime.
- ***R***: A random immigrant in this neighborhood has been involved in crime.
- ***B***: The statistical study Y is associated with political bodies.

When we see the statistical study Y, we may face *two macro-decisions*. One is based on our uncertainty about the proposition R (how should we treat a random immigrant in this neighborhood given that they might have been involved in crime) and the other is based on our uncertainty about the proposition N (should we share it with others?). So, we already have a precise prior credence[12] towards R and N. We also face two micro-decisions: (i) whether to fine-grain (and then update) with respect to T and (ii) whether to fine-grain with respect to B[13].

Chapter 1 shows that if you are statistically sophisticated, you will not be sufficiently confident that T is decision-changing (i.e. for T, $P(D) < \frac{C_F}{\Delta EU_C}$). This implies that your confidence towards the *content* of that statistical report will not be high. So, if you are statistically sophisticated you will have a low prior credence towards N and according to the probabilistic account of belief that I am using [39] you should suspend judgment on it as your credence of it is equal or below the threshold 0.5.

On the other hand, Chapter 2 shows that if you are socially aware then you are rather confident (or at least not too inconfident) that you *learn* B i.e. that the study has been associated with political bodies (i.e. for B, $P(S) > \frac{C_F}{\Delta EU_C}$. This implies that

[12] In the model, we assume that maintaining precise credence towards the states of macro-decisions has negligible cognitive cost. A decision under uncertainty qualifies as a macro-decision based on where it falls on the timeline of inquiry.

[13] I have not discussed the order of these micro decisions of fine-graining and I do not think it matters. I think one can even do them simultaneously by considering the macro-decisions that they are going to make simultaneously i.e. the macro-decision about the random member of a population and the macro decision about sharing/accepting the content of the statistical report. If they do so, they start off by a 4-cell algebra which is fine-grained with respect to a proposition about a random member of the population (R) and a proposition about the truth of the *content* of statistical report (N). After making the two micro-decisions, if they are socially aware OR statistically sophisticated they end up with an 8-cell algebra which is fine-grained with respect to a proposition about whether the study is associated with political bodies (B) and *not* fine-grained with respect to the statistical report itself (T).

you are rather confident that your confidence in T will be undermined and you will never learn it *stably* in your line of enquiry[14] (i.e. for T, $P(S) < \frac{C_F}{\Delta EU_C}$). So, as a socially-aware agent you should not fine-grain (or update) your credence of R based on T.

Combining these two result we can infer:

> If you are statistically sophisticated OR socially aware, you should refrain from updating your credence based on a statistical study that aligns with social stereotypes and reinforces one side of a political controversy. Furthermore, you should suspend judgment on its content.

4.6 Appendix

Suppose in the following decision, we are initially (before fine-graining) inclined to choose A_2.

	$\boldsymbol{X}$	$\neg \boldsymbol{X}$
$\boldsymbol{A_1}$	H	L
$\rightarrow \boldsymbol{A_2}$	M	M

Macro-decision C

That means our expected utility with respect to C before fine-graining is:

$EU_C^{\text{Before F}} = M$

If we learn E stably (S) and E is decision-changing (D) then our expected utility with respect to C after fine-graining is:

$EU_C^{\text{After F}} = P(X|E\&D) * H + P(\neg X|E\&D) * L = P(X|E\&D) * (H - L) + L$

Therefore, $\Delta EU_C = P(X|E\&D) * (H - L) - (M - L)$

So, $\dfrac{C_F}{\Delta EU_C} = \dfrac{C_F}{P(X|E\&D) * (H - L) - (M - L)} = \dfrac{\frac{C_F}{H-L}}{P(X|E\&D) - \frac{M-L}{H-L}}$

On the other hand, the threshold for switching from A_2 to A_1 is $P^{th} = \dfrac{M-L}{H-L}$. Therefore,

$$\frac{C_F}{\Delta EU_C} = \frac{\frac{C_F}{H-L}}{P(X|E\&D) - P^{th}}$$

If we can estimate an α for which the following comparative confidence holds: $X \succ_{E\&D} B_\alpha$ in which B_α is the chance of a coin with the bias α towards head to land head, then we can make the probabilistic judgment about $P(X|E\&D)$ by consulting that comparative confidence. If we have that estimation then we can say

[14] Here, for simplicity, I assumed that the two macro-decisions share the same enquiry timeline. This assumption is, of course, open to dispute. However, even if we reject it, while the result of Chapter 2 may not be easily generalizable to Chapter 1, the assumption that for T, $P(S) < \frac{C_F}{\Delta EU_C}$ remains plausible. Thus, the conclusion holds.

$\alpha < P(X|E\&D) \leqslant 1$. This will give us a range for $P(X|E\&D) - P^{th}$ which is the denominator of the fraction $\dfrac{C_F}{\Delta EU_C}$ according to the above formula and therefore a range for $\dfrac{C_F}{\Delta EU_C}$.

The next step is estimating a β and γ for which the following comparative confidence holds: $B_\gamma \succ D \succ B_\beta$ in which B_γ and B_β are the chances of coins with the biases γ and β respectively towards head to land head. We can then use that comparative confidence to make the probabilistic judgments $\gamma > P(D) > \beta$.

Having a range for both $P(D)$ and $\dfrac{C_F}{\Delta EU_C}$, we can make the fine-graining decision.

Conclusion and Further Research

This work stems from three intuitions across epistemology, philosophy of language, and political philosophy. These intuitions guided my exploration of probability in social contexts:

- **Epistemology:** We can leverage our cognitive boundedness to claim control over our epistemic attitudes. Specifically, the fact that we cannot judge probabilistic dependence between propositions instantly and without cognitive cost allows us to have control over what we include in the domain of our credence function and the set of categorical beliefs.
- **Philosophy of language:** Statements expressing statistical correlations can imply more than an innocent correlation, depending on the context of conversation. This possibility arises from the inherently ambiguous nature of the notion of probability.
- **Political and moral philosophy:** Statistical studies about social groups can be exploited in policy-making to justify policies targeting minorities and oppressed social groups. They can also serve as Trojan horses in political controversies, smuggling essentialist claims about these groups into social discourse. These possibilities appropriately evoke moral discomfort in using, believing, and sharing the findings of such studies.

The first insight inspired my model of fine-graining the space of possibilities, which led to a rationality norm for using new evidence and engaging with new possibilities for a bounded Bayesian. I then used a probabilistic account of belief [39] to translate the norm of engaging with new possibilities into a norm for maintaining categorical belief.

Leitgeb [39]'s stability thesis provides a formal basis for the intuition that whether we should believe a proposition X depends on *how easily* our high credence in X can be undermined. My account builds on Leitgeb's thesis by introducing the variable of 'our line of enquiry' into the framework. The ease with which our high credence in a proposition is undermined depends on the information we encounter, and the information we encounter is shaped by our line of enquiry.

On the other hand, our line of enquiry itself depends on the decision(s) we are going to make under uncertainty. We do not gather evidence without purpose, and that purpose is usually tied to a decision under uncertainty. We aim to gather as much relevant information as possible for that decision, and this endeavor defines our line of enquiry. So, if a piece of information is not sufficiently relevant to the decision at hand, it is natural to assume that we should ignore it as bounded agents with limited resources. This implies that if a piece of information is not relevant enough to the decision we are about to make, it should not be considered a threat to our high credence in other propositions, and therefore to our set of categorical beliefs.

Thus, what we should believe depends on the decisions we are about to make. More specifically, whether we should believe a proposition X depends on the decisions we are about to make under uncertainty about propositions relevant to X. Accordingly, my account can also be viewed as a formal (probabilistic) explanation of the thesis of pragmatic encroachment[36].

I also believe that the model of fine-graining I proposed can offer insights into the doxastic voluntarism debate. I argued that, given the cognitive cost associated with fine-graining our space of possibilities, fine-graining is a voluntary act. Using a probabilistic account of belief Leitgeb [39], I demonstrated that fine-graining our algebra influences what we should believe. This suggests that we have control over our set of beliefs through the structure of our algebra. It is worth exploring how the type of control we have over our algebra corresponds to the four types of control identified by Alston [1]. Based on the proposed model of fine-graining, our algebra is undoubtedly influenced by the decisions under uncertainty that we are considering and the sources of information we encounter. This aligns with what Alston [1] refers to as "indirect influence," which he categorizes as the fourth type of control.

I am also interested in exploring the question of fine-graining with respect to epistemic accuracy. It is well-established that conditionalization (and, by extension, further fine-graining) always increases the accuracy of the credence function, regardless of how fine-grained the current algebra is[48]. However, it is worth investigating whether we should fine-grain with respect to a newly salient proposition if we prioritize the accuracy of certain propositions in the algebra over others. If we have this asymmetric preference for promoting the accuracy of propositions already included in the algebra, how does that influence the fine-graining decision?

Another potential avenue of exploration involves the implicatures of probabilistic statements about social groups. I propose examining how these implicatures relate to the implicatures of generic statements about social groups. This connection is particularly relevant because majority generics represent a significant category of generics [40]. Majority generics about social groups can often be interpreted as probabilistic statements about them, leading to the aforementioned harmful implicatures.

Furthermore, the strategies discussed for mitigating the harm of these implicatures could be explored in the context of generic statements. I believe the second strategy, namely offering alternative causal models, will be especially effective for generic statements, given recent causal accounts of their semantics [59], [52].

My thesis aimed to explain why a deliberate attempt to avoid engaging epistemically (through suspension of judgment) and instrumentally (through refusing to update on them) with statistics aligned with social stereotypes is rational. It is interesting to explore whether this avoidance can be considered as an attempt to *alter the essence* of certain social groups, like ethnic groups, over time. For example, if we adopt Hu [34]'s "thick constructivist" account of race, according to which the essence of an ethnic group is determined by the set of statistical regularities attributed to that group, then refusing to engage with these statistical regularities may change the essence of the ethnic group over time, thereby altering the oppressive social practices directed towards that group.

Bibliography

[1] William Alston. “The deontological conception of epistemic justification”. In: *Arguing About Knowledge*. Routledge, 2020, pp. 324–350.

[2] Boris Babic et al. “Normativity, epistemic rationality, and noisy statistical evidence”. In: *The British Journal for the Philosophy of Science* 75.1 (2024), pp. 153–176.

[3] Rima Basu. “Radical moral encroachment: The moral stakes of racist beliefs”. In: *Philosophical Issues* 29.1 (2019), pp. 9–23.

[4] Rima Basu. “The wrongs of racist beliefs”. In: *Philosophical Studies* 176.9 (2019), pp. 2497–2515.

[5] Rima Basu and Mark Schroeder. “Doxastic wronging”. In: (2019).

[6] David Blackwell. “Equivalent comparisons of experiments”. In: *The annals of mathematical statistics* (1953), pp. 265–272.

[7] Renée Jorgensen Bolinger. “The rational impermissibility of accepting (some) racial generalizations”. In: *Synthese* 197.6 (2020), pp. 2415–2431.

[8] Renée Jorgensen Bolinger. “Varieties of moral encroachment”. In: *Philosophical Perspectives* 34.1 (2020), pp. 5–26.

[9] Luc Bovens. “Selection Under Uncertainty: Affirmative Action at Shortlisting Stage”. In: *Mind* 125.498 (2016), pp. 421–437. DOI: 10.1093/mind/fzv157.

[10] Richard Bradley. *Decision theory with a human face*. Cambridge University Press, 2017.

[11] Lara Buchak. “Can it be rational to have faith?” In: *Contemporary Epistemology: An Anthology* (2019), pp. 110–125.

[12] Lara Buchak. “Instrumental rationality, epistemic rationality, and evidence-gathering”. In: *Philosophical Perspectives* 24 (2010), pp. 85–120.

[13] Rudolf Carnap. “On the application of inductive logic”. In: *Philosophy and phenomenological research* 8.1 (1947), pp. 133–148.

[14] Gregory F Cooper. “The computational complexity of probabilistic inference using Bayesian belief networks”. In: *Artificial intelligence* 42.2-3 (1990), pp. 393–405.

[15] Paul Dagum and Michael Luby. “Approximating probabilistic inference in Bayesian belief networks is NP-hard”. In: *Artificial intelligence* 60.1 (1993), pp. 141–153.

[16] Wayne Davis. “Implicature”. In: *The Stanford Encyclopedia of Philosophy.* Ed. by Edward N. Zalta and Uri Nodelman. Spring 2024. Metaphysics Research Lab, Stanford University, 2024.

[17] Heather Douglas. *Science, policy, and the value-free ideal.* University of Pittsburgh Pre, 2009.

[18] Jane Friedman. “Inquiry and belief”. In: *Noûs* 53.2 (2019), pp. 296–315.

[19] James Fritz and Elizabeth Jackson. “Belief, credence, and moral encroachment”. In: *Synthese* 199.1 (2021), pp. 1387–1408.

[20] Marilyn Frye. *Politics of reality: Essays in feminist theory.* Crossing Press, 1983.

[21] Georgi Gardiner. “Evidentialism and moral encroachment”. In: *Believing in accordance with the evidence: New essays on evidentialism* (2018), pp. 169–195.

[22] Tamar Szabó Gendler. “On the epistemic costs of implicit bias”. In: *Philosophical Studies* 156.1 (2011), pp. 33–63.

[23] Irving John Good. “On the principle of total evidence”. In: (1966).

[24] Irving John Good. “On the principle of total evidence”. In: *The British Journal for the Philosophy of Science* (1967).

[25] Jeremy Goodman and Bernhard Salow. “Epistemology normalized”. In: *Philosophical Review* 132.1 (2023), pp. 89–145.

[26] Hilary Greaves and David Wallace. “Justifying conditionalization: Conditionalization maximizes expected epistemic utility”. In: *Mind* 115.459 (2006), pp. 607–632.

[27] H Paul Grice. “Meaning”. In: *The philosophical review* 66.3 (1957), pp. 377–388.

[28] Mario Gunther. “Probability of guilt”. In: (Manuscript, available online: https://philpapers.org/re

[29] Alan Hájek. “Interpretations of Probability”. In: *The Stanford Encyclopedia of Philosophy.* Ed. by Edward N. Zalta and Uri Nodelman. Winter 2023. Metaphysics Research Lab, Stanford University, 2023.

[30] Sally Haslanger. “Ideology, generics, and common ground”. In: *Feminist metaphysics: Explorations in the ontology of sex, gender and the self.* Springer, 2010, pp. 179–207.

[31] Carl G Hempel. “Science and human values”. In: *Ethical Issues in Scientific Research.* Routledge, 2015, pp. 6–28.

[32] Daniel Hoek. “Questions in action”. In: (2022).

[33] Janina Hosiasson. “Why do we prefer probabilities relative to many data?” In: *Mind* 40.157 (1931), pp. 23–36.

[34] Lily Hu. “What is “race” in algorithmic discrimination on the basis of race?” In: *Journal of Moral Philosophy* 1.aop (2023), pp. 1–26.

[35] Simon M Huttegger. *The probabilistic foundations of rational learning.* Cambridge University Press, 2017.

[36] Jonathan Jenkins Ichikawa and Matthias Steup. “The Analysis of Knowledge”. In: *The Stanford Encyclopedia of Philosophy.* Ed. by Edward N. Zalta and Uri Nodelman. Fall 2024. Metaphysics Research Lab, Stanford University, 2024.

[37] Simon Keller. “Friendship and belief”. In: *Philosophical Papers* 33.3 (2004), pp. 329–351.

[38] Brian Kim. “Pragmatic encroachment in epistemology”. In: *Philosophy Compass* 12.5 (2017), e12415.

[39] Hannes Leitgeb. *The stability of belief: How rational belief coheres with probability.* Oxford University Press, 2017.

[40] Sarah-Jane Leslie. “Generics: Cognition and acquisition”. In: *Philosophical Review* 117.1 (2008), pp. 1–47.

[41] Sarah-Jane Leslie. “The original sin of cognition”. In: *The Journal of Philosophy* 114.8 (2017), pp. 395–421.

[42] Hanti Lin and Kevin T Kelly. “A geo-logical solution to the lottery paradox, with applications to conditional logic”. In: *Synthese* 186 (2012), pp. 531–575.

[43] Anna Mahtani. “Awareness growth and dispositional attitudes”. In: *Synthese* 198 (2021), pp. 8981–8997.

[44] Matthew McGrath. “Being neutral: Agnosticism, inquiry and the suspension of judgment”. In: *Noûs* 55.2 (2021), pp. 463–484.

[45] Andreas Mogensen. “Racial profiling and cumulative injustice”. In: *Philosophy and Phenomenological Research* 98.2 (2019), pp. 452–477.

[46] Sarah Moss. “IX—Moral encroachment”. In: *Proceedings of the aristotelian society.* Vol. 118. 2. Oxford University Press. 2018, pp. 177–205.

[47] Jessie Munton. “Beyond accuracy: Epistemic flaws with statistical generalizations”. In: *Philosophical Issues* 29.1 (2019), pp. 228–240.

[48] Graham Oddie. “Conditionalization, cogency, and cognitive value”. In: *The British Journal for the Philosophy of Science* 48.4 (1997), pp. 533–541.

[49] Michael Pace. “The epistemic value of moral considerations: Justification, moral encroachment, and James”Will to believe”'. In: *Noûs* 45.2 (2011), pp. 239–268.

[50] Anne Phillips. “Unconditional equals”. In: (2021).

[51] Joseph Raz. *Practical reason and norms.* OUP Oxford, 1999.

[52] Robert van Rooij and Katrin Schulz. “A Causal Power Semantics for Generic Sentences”. In: *Topoi* 40.1 (2021), pp. 131–146.

[53] Lewis Ross. “Profiling, neutrality, and social equality”. In: *Australasian Journal of Philosophy* 100.4 (2022), pp. 808–824.

[54] Richard Rudner. “The scientist qua scientist makes value judgments”. In: *Philosophy of science* 20.1 (1953), pp. 1–6.

[55] BJ Salow and J Goodman. “Belief revision from probability”. In: (2023).

[56] Miriam Schoenfield. “Permission to believe: Why permissivism is true and what it tells us about irrelevant influences on belief”. In: *Contemporary Epistemology: An Anthology* (2019), pp. 277–295.

[57] Mark Schroeder. “The ubiquity of state-given reasons”. In: *Ethics* 122.3 (2012), pp. 457–488.

[58] David Spiegelhalter. “Does probability exist?” In: *Nature* 636 (2024), p. 19.

[59] Michael Strevens. “Ceteris paribus hedges: Causal voodoo that works”. In: *The Journal of philosophy* 109.11 (2012), pp. 652–675.

[60] Sarah Stroud. “Epistemic partiality in friendship”. In: *Ethics* 116.3 (2006), pp. 498–524.

[61] Michael G Titelbaum. *The Stability of Belief: How Rational Belief Coheres with Probability, by Hannes Leitgeb*. 2021.

[62] Minkyung Wang. “Aggregating Credences Into Beliefs: Threshold-Based Approaches”. In: *Logic, Rationality, and Interaction: 9th International Workshop, LORI 2023, Jinan, China, October 26?29, 2023, Proceedings*. Ed. by Natasha Alechina, Andreas Herzig, and Fei Liang. Springer Nature Switzerland, 2023, pp. 269–283.

[63] Zina B Ward. “On value-laden science”. In: *Studies in History and Philosophy of Science Part A* 85 (2021), pp. 54–62.

[64] Gregory Wheeler. “Less is more for Bayesians, too”. In: *Routledge Handbook of Bounded Rationality.* Routledge, 2020, pp. 471–483.

[65] Roger White. “Problems for dogmatism”. In: *Philosophical studies* 131 (2006), pp. 525–557.