UNIVERSITY OF CALIFORNIA,
IRVINE

Non-epistemic Values in Model Building, Theory Testing, and Communication in Science

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Logic and Philosophy of Science

by

Alysha Kassam

Dissertation Committee:
Professor Cailin O' Connor Irvine, Chair
Professor James Weatherall
Associate Professor Lauren Ross

2021

DEDICATION

To

my advisor Cailin and mother Noor

for support and guidance

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to express the deepest appreciation to my committee chair and dissertation advisor, Professor Cailin O'Connor. She has spent endless hours proofreading my research papers and providing excellent feedback. Without her, completing the PhD and this dissertation project would not have been possible.

I would like to thank my committee members, Professor Lauren Ross and Professor Jim Weatherall, whose generous assistance and enthusiasm for my research carried me through the project.

In addition, a thank you to Professor Patricia Marino and Pen Maddy for their constant feedback and care. I am so grateful for both their mentorship and friendship throughout the years.

Financial support was provided by the University of California, Irvine.

VITA

Alysha Kassam

2012	B.A. in Psychology, University of California, Irvine
2019	M.A. in Logic and Philosophy of Science, University of California, Irvine
2021	Ph.D. in Logic and Philosophy of Science University of California, Irvine

ABSTRACT OF THE DISSERTATION

Non-epistemic Values in Model Building, Theory Testing, and Communication in Science

by

Alysha Kassam

Doctor of Philosophy in Logic and Philosophy of Science

University of California, Irvine, 2021

Professor Cailin O'Connor, Chair

Scientific practice has long portrayed itself as objective, in the sense that it is guided by epistemic values that are independent of ethical, social and political thought. The worry scientists have long had is that moral or political reasoning undermines science, as it contaminates the search for truth with social, political and ethical priorities and motives. However, there are many ways in which science is responsible to society, as the fruits of science are often used in value-laden settings. For instance, consider how science bears on the distribution of resources (Greenberg 2001, Kitcher 2001), or the evaluation of risks (Beck 1992, Douglas 2009, Hempel 1965), or how it shapes the material conditions of our lives (Winner 1986, Scolve 1995, Kitcher 2001). When one considers this more seriously, a clear separation between science and social concerns starts to seem less plausible. For this reason, feminist philosophers of science have criticized the value-free ideal, pointing out that non-epistemic values (i.e., social, political, ethical values) are not only unavoidable, but also often critical to proper scientific reasoning. Grappling with the notion that non-epistemic values play an important role in scientific reasoning, philosophers have asked themselves: when and how do non-epistemic values serve a

permissible role? The purpose of this chapter is to survey the various responses to this question in the philosophical literature.

In this dissertation, I will first describe the value-free ideal and the challenges that have been lodged against it. The subsequent discussion will center on philosophers' proposed criteria of distinguishing between non-epistemic values that play a permissible versus impermissible role in scientific reasoning. I will then highlight some philosophical perspectives on the types of ethical considerations important to scientific reasoning. Finally, the dissertation project will close with a discussion on how non-epistemic values are often embedded in mathematical modeling work and what implications should be drawn from this.

INTRODUCTION

My dissertation project centers on science policy and the role of scientific expertise in democracy. The project is motivated by the current debate over the ideal of value-free science. Scientific practice has long portrayed itself as objective, in the sense that it is guided by epistemic values that are independent of ethical, social and political thought. The worry scientists have long had is that non-epistemic values undermine the objectivity of science by contaminating the search for truth with social, political and ethical priorities and motives. However, there are many ways in which science is responsible to society, as the fruits of science are often used in value-laden settings. When one considers this more seriously, a clear separation between science and social concerns starts to seem less plausible.

One way in which non-epistemic values play a role in scientific reasoning is through inductive risk considerations. The concept of inductive risk was first expressed by Hempel (1965) and later developed by Douglas (2009) and is the chance that one will be wrong in accepting or rejecting a hypothesis. According to Douglas, the choice of a level of statistical significance requires scientists to consider which kind of error they are willing to tolerate, as changing the level of statistical significance changes the balance between false positives and negatives (Douglas, 2009). For instance, if one wishes to avoid false negatives and is willing to accept more false positives, then she should lower the standard for statistical significance. On the other hand, if one wishes to avoid false positives, then she should raise the standard for statistical significance. In developing a standard for statistical significance, scientists must consider the consequences of false positive and false negative results. Considerations surrounding these consequences often include non-epistemic value judgements.

Based on these considerations, Douglas claims that scientists should not aim to entirely exclude non-epistemic values from their reasoning or, in other words, that the value-free ideal is not a defensible *ideal*. Scientists have ethical responsibilities with respect to society as their decisions have social consequences. Douglas is predominantly concerned with the unintended harm scientists may cause by their negligence. On the basis of this moral concern, Douglas claims that when there are potential risks involved in the confirmation of a hypothesis, scientists should raise their evidential standards to avoid causing harm. By the same token, if a hypothesis supports a social- good, should scientists relax their evidential standards since the acceptance of the hypothesis has positive social consequences?

In the first chapter, I explore this question in relation to the construction and application of mathematical models in the social sciences. I use Hong and Page's (2004) 'diversity trumps ability' result as a key example where academics have dropped their epistemic standards because the model's stated result supports a social-good. I argue that in dropping our epistemic standards, we undermine the conditions for rejecting research that supports social ills on epistemic grounds.

I consider the value-laden assumptions and consequences associated with the use of mathematical models in other domains of scientific inquiry and the ethical obligations modelers have to the general public. Problems arise when non-epistemic values are embedded in mathematical models in such a way that unpacking these value judgements becomes a difficult and sometimes unfeasible task. In her book *Weapons of Math Destruction*, Cathy O'Neil concludes that complex mathematical models can be good at hiding the various ways non-epistemic values are embedded in their construction. According to O'Neil, non-epistemic values can permissibly influence model- building if modelers are transparent about the role such values play. If a model is transparent, then even if it encodes a bias, this bias can nevertheless be evaluated by other modelers and the general public. However, O'Neil finds that many mathematical models are what she calls 'weapons of math destruction'—mathematical models that have negative social consequences because they opaquely encode biases.

The current project expands on this idea by pointing out additional epistemic features of mathematical models that make them difficult to evaluate by the general public. Consider the nuanced epistemic role mathematical models may play in argumentation. Sometimes mathematical models do not tell us something is the case, but instead, provide a plausibility argument for why something *may* be the case, given a certain set of assumptions. Moreover, mathematical models can be difficult to replicate, partly because of the opaque role non-epistemic values play in their construction. I consider how mathematical models may encode prejudice and bias in ways that can be opaque and whether modelers have an ethical obligation to make their modeling work transparent and replicable.

Along these lines, the second chapter considers the concept of transparency—which has received a great deal of attention in the philosophy of science literature, especially as it relates to communicating scientific studies and findings to non-experts. Many philosophers claim that transparency is important for establishing public trust in science (Douglas 2009; Elliott 2017; Kitcher 2011; I. de Melo-Martin and Intemann 2018; Stanev 2017; Williams 2002) whereas some philosophers instead argue that transparency about scientific practices could actually promote unwarranted skepticism (John 2018; Kovaka 2019). The transparency debate has often centered on the notion that skepticism in science is caused by a discordance between non-experts' idealized assumptions on scientific methodology and actual scientific practice.

However, if we want to better understand what role transparency plays in the public's perception of scientific claims, we must realize that the issue is multifaceted and complex. In these later chapters, I argue that claiming we shouldn’t be transparent because non-experts hold a false folk philosophy of science oversimplifies the issue. When determining whether transparency will bolster or hinder public trust in science, we must first consider the way scientific information is being communicated to the public. More specifically, we should ask ourselves: ‘how is the general public receiving scientific information?’, ‘has this scientific topic been politicized?’ and ‘are special interest groups playing a role in the dissemination of the scientific information?’. Answering these questions is essential to determining the effectiveness of transparency. This is because, transparency is only effective if the public receives an accurate, unbiased account of what most scientists actually claim.

The third and final chapter considers whether a clear demarcation between epistemic and non-epistemic values is tenable in the machine-learning context. Machine learning algorithms are often touted as superior to human decision-making due to their ability to be completely objective and have been increasingly used in public and private practices such as predicting recidivism in criminal justice, determining who should be hired, admitted to university, or granted social welfare benefits, evaluating job performance, suggesting who should get a loan, or pay which insurance rate. In these contexts, it has been made evident that the outcomes associated with machine learning algorithms have been worse for racialized people, women and for people in other minority or marginalized communities. It is commonly assumed in the machine learning literature that unfair outcomes are due to issues with data collection: either the data is not representative of the population or the problematic pattern is already pervasive in the population (Johnson, 1). As the assumption goes, non-epistemic values can infect machine learning algorithms if such values are already present in the data on which the algorithm is operating on (ibid). What this assumption implies is that non-epistemic values are usually not inserted into the algorithm itself through the explicit design decisions of the modeler.

Taking inspiration from the values in science literature, I attempt to push back on this general assumption. Helen Longino (1995, 1996) famously argued that epistemic and non-epistemic values play an indistinguishable role in scientific reasoning. According to Longino, what proponents of the value-free ideal fail to recognize is how their choice to adopt seemingly purely non-epistemic values over others is itself a value-laden judgement as it arguably includes an appeal to social and political factors. Longino's analysis is applicable to the machine learning context. Recently, in response to the ubiquitous application of machine-learning algorithms in decision procedures that directly impact peoples' lives, many modelers have developed an interest in algorithmic fairness. A common view in the machine learning literature is that there is an inherent trade-off between accuracy and fairness. However, it is a choice to model assumptions that cast fairness in direct opposition to accuracy. This framing of a tradeoff does not just involve purely mathematical assumptions, but also implicates non-epistemic concerns regarding how to value fairness and accuracy both independently and in relation to each other. Thus, the influence of non-epistemic values in machine learning algorithms extends far beyond simply data collection—non-epistemic values play a role in explicit choices concerning how to define algorithmic accuracy in the first place. Tying this back to Longino's claim, I argue that accuracy is not a purely epistemic notion in the machine learning context and as I discuss, this has important implications for the algorithmic fairness literature.

CHAPTER 1

EPISTEMIC RISK IN THE DIVERSITY TRUMPS ABILITTY MODEL

1.1 Introduction

While the results of science are used in many value-laden settings, scientific practice has often portrayed itself as objective and value-free. More recently, philosophers of science have criticized the value-free ideal, pointing out that non-epistemic values are often critical to proper scientific reasoning. Grappling with the notion that non-epistemic values play an important role in scientific reasoning, philosophers have asked themselves: when and how do non-epistemic values serve a permissible role?

Heather Douglas (2009) has argued that non-epistemic values play an indispensable role in scientific reasoning through her discussion of inductive risk. Very roughly, inductive risk is the chance that one will be wrong when accepting or rejecting a scientific hypothesis. When deciding their evidential standard for a hypothesis, Douglas claims that scientists must consider what the consequences of error would be. Moreover, in assessing the consequences of error, non-epistemic values often play a role. For instance, when testing whether a certain pesticide is environmentally safe, a scientist concerned about public safety may raise their evidential standards to avoid causing harm.

Thus, Douglas claims that when there are potential social risks that follow from the confirmation of a hypothesis, scientists should raise their evidential standards to ensure public safety. If Douglas is correct that scientists should consider the bad consequence associated with making erroneous claims, such that it requires scientists to raise their evidential standards in order to avoid causing negligent harm does it then follow that scientists should likewise consider the potential benefits of accepting or rejecting a hypothesis? Suppose the hypothesis in question is in support of a social-good. In this case, should a scientist relax their evidential standards, since the acceptance of the hypothesis has positive consequences?

This paper attempts to answer these questions in relation to the construction and application of mathematical models in the social sciences. I use Hong and Page's 'diversity trumps ability' result as a key example where academics have dropped their epistemic standards because the model's stated results support a social-good. As I argue, this has a consequence. The model also has an unstated result that "highest ability problem solvers cannot be diverse" (Hong and Page, 2004, 16389). This result can be utilized to support a socially pernicious notion that groups of best experts must not be diverse. I will argue that in dropping our epistemic standards in evaluating Hong and Page's model, we have no clear epistemic grounds to dismiss this socially pernicious result since, after all, the model itself has not changed. In other words, in dropping our evidential standards to support the 'diversity trumps ability' result, we make it difficult to reject the model's other result that experts must be uniform. I claim that this shows, more generally, that in dropping our epistemic standards, we undermine the conditions for rejecting research that supports social-ills on weak epistemic grounds.

This paper has two aspects of novelty. First, the paper attempts to answer the question of whether modelers should *lower* their epistemic standards when a model's results support a social good. Second, the paper explores whether inductive risk calculations can be applied to mathematical models more generally.

The paper will proceed as follows. The second section of the paper introduces the Heather Douglas's work on non-epistemic values in science, focusing on her notion of inductive risk. The third section discusses Hong and Page's model and simulation results that support the idea that cognitive diversity is more important than ability when it comes to group problem-solving. The fourth section explores some critiques of the model and exposes how non-epistemic values may be playing a role in the author's and general public's assessment of the evidential sufficiency of the model. The fifth section evaluates whether the role values play is in fact a good thing in the case of Hong and Page's model.

1.2 Heather Douglas's view of Inductive Risk

The view that only epistemic values have a legitimate role to play in science has been importantly challenged. Epistemic values such as predictive accuracy, explanatory power, consistency, etc., have always been thought to play a legitimate role throughout all aspects of scientific reasoning (Kuhn, 1977). More recently, philosophers of science have argued that non-epistemic values (e.g., ethical and political concerns) also play a role in many aspects of science. It will be helpful to distinguish at least four stages at which non-epistemic values may affect science. Non-epistemic values may play a role in the (1) choice of a research problem, (2) gathering evidence, (3) the acceptance or rejection of a hypothesis, and (4) the application of the scientific research results (Weber, 1988). Most philosophers of science believe that non-epistemic values permissibly play a role in choosing a research problem and when applying research results. Thus, the real debate has centered on whether values can play a permissible role at the core of scientific reasoning, or in steps two and three.

One way in which non-epistemic values play a role in the internal stages of scientific reasoning is through considerations surrounding inductive risk. The concept of inductive risk was first expressed by Rudner (1953) and Hempel (1965) and was later developed by Douglas (2009) and is the chance that one will be wrong when accepting or rejecting a scientific hypothesis. There are two ways in which scientists can go wrong when accepting or rejecting a hypothesis. The first type of error consists in concluding that there is a phenomenon or an effect when in fact there is none. This is called a type I error or a false positive result. The second type of error consists in discounting or missing an existing phenomenon or effect. This is called a type II error or a false negative result.

According to Douglas, the choice of a level of statistical significance requires scientists to consider which kind of error they are willing to tolerate, as changing the level of statistical significance changes the balance between false positives and negatives (Douglas, 2009). For instance, if one wishes to avoid false negatives and is willing to accept more false positives, then she should lower the standard for statistical significance. On the other hand, if one wishes to avoid false positives, then she should raise the standard for statistical significance. In order to reduce both types of error, one must devise methods for improving the overall statistical adequacy of the experiment (like, for example, increasing the population size). Oftentimes, scientists do not have the means of increasing the overall statistical adequacy of their experiments, so trade-offs between type I and type II errors, like the ones just mentioned, must be made instead.

In developing a standard for statistical significance, scientists must consider the consequences of false positive and false negative results. Considerations surrounding these consequences often include non-epistemic value judgements. This can be seen from a case in which it is uncertain whether a drug has a serious harmful side-effect. Acting as if there were no such side effect when there is one (type II error) would put the public at more risk than acting as if there were such a side effect when there is none (type I error). Thus, a scientist concerned about public safety will find an excess of false positives and a limited number of false negatives permissible. On the other hand, suppose the potential risk of this drug is very mild. Further suppose that the scientist in question helped develop this pharmaceutical drug and is eager to get it out on the market because of its great health benefits. When testing for the side effect, this scientist will find an excess of false negatives and a limited number of false positives permissible, leading to its under-regulation.

It is important to note that on Douglas's picture, inductive risk is not only present in determining whether the evidence is sufficient to support the conclusion of a research project. Instead, inductive risk is present at all moments in scientific reasoning, including the first stages where scientists are confronted with ambiguous data and must decide what to do with it. In characterizing the data, scientists must ask themselves:

> Should they discard them (potentially lowering the power of their study)? Should they characterize them one way or another? Should they give up on the study until a more precise methodology can be found? Each of these choices poses inductive risks for the scientist, a chance that their decision could be a wrong one and thus that they will incur the consequences of error. (Douglas, forthcoming, p. 7).

In answering these questions, scientists often draw upon non-epistemic values (ibid.). Non-epistemic values thus play a role throughout all stages of scientific reasoning on Douglas's view.

One might try to resist Douglas's inductive risk argument by adopting a Bayesian approach. A Bayesian can claim that scientists do not accept nor reject hypotheses in the way inductive risk arguments describe. Instead, scientists merely assign probabilities to hypotheses (Jeffrey, 1956; Mitchell, 2004). These probabilities represent degrees of belief in a hypothesis and are arrived at by an application of Bayes' Rule, which does not require appeal to non-epistemic values (Parker and Winsberg, 2017). Bayes' rule provides a formula for updating the probability assigned to a hypothesis H in light of new evidence, e. The updating of probabilities is always conditional on the agent's background information B, or all the information the agent has prior to the point of considering how the evidence e should affect her probability assignments.[1]

However, Douglas's inductive risk argument need not be about significance testing and p-values. For instance, Steele (2015) claims that scientists often lack the precise degrees of belief or the probabilities that serve as priors and likelihoods that are needed as inputs to Bayesian analysis. Instead, scientists must decide how to represent these probabilities, and these decisions, like other methodological decisions in science, are subject to inductive risk (Steele, 2015).

Although Douglas claims that non-epistemic values play an important role in activities central to scientific reasoning, she does admit that these non-epistemic values should not interfere with scientific reasoning in such a way that it threatens the objectivity of science. In order to both preserve the objectivity of science while still being sensitive to the fact that non-epistemic values play an indispensable role in scientific reasoning, Douglas distinguishes between direct and indirect roles for values (2000, 2009). Values play a direct role when a scientist considers "the direct consequences of a particular course of action" whereas values play an indirect role when they help scientists decide how to respond to the potential consequences of making erroneous choices or producing inaccurate results (Douglas, 2000, 564-565). Another way Douglas characterizes the distinction is the following: values operate in a direct role when they act "as reasons in themselves" or "as stand-alone reasons" to motivate our choices (2009, 96). In contrast, she says that values act indirectly when they "act to weigh the importance of

[1] More explicitly, Bayes' Rule can be formulated as follows: $p(H|e\&B)= p(H|B) \times p(e|H\&B) / p(e|B)$.

uncertainty, helping to decide what should count as sufficient" reason for a choice (2009, 96). According to Douglas, non-epistemic values can be rightly influential when they play an indirect role and should only rarely play a direct role in activities central to scientific reasoning.

It should be noted that Douglas further claims that scientists should not aim to entirely exclude non-epistemic values from their reasoning or, in other words, that the value-free ideal is not a defensible *ideal*. Scientists have ethical responsibilities with respect to society as their decisions have social consequences. Douglas is predominantly concerned with the unintended harm scientists may cause by their negligence. On the basis of this moral concern, Douglas claims that when there are potential risks involved in the confirmation of a hypothesis, scientists should raise their evidential standards to avoid causing harm. By the same token, if a hypothesis supports a social-good, should scientists relax their evidential standards since the acceptance of the hypothesis has positive social consequences? Douglas is silent on how the positive social consequences of a hypothesis may affect a scientist's evidential standards. For the remainder of this paper, I will explore this question in relation to Hong and Page's modeling work on diverse groups of problem solvers.

1.3 The 'Diversity Trumps Ability' Result

Hong and Page's (2004; Page, 2007) 'diversity trumps ability' results indicate that functionally diverse groups whose members have less ability outperform groups of best individual problem solvers. These results are derived from simulation models, and the authors also develop a mathematical theorem to explain the logic behind the model's results.

In the model, the problem which the agents are trying to solve is represented by a circle of 2000 spots. Each spot on the circle can be considered a candidate answer to the problem. The agents move together along the circle and eventually land on a particular spot. There is a random integer assigned to each spot on the circle and this random integer is considered the epistemic payoff for landing on this particular spot in the circle.

The agents each have a *heuristic* that they use to move forward in the circle. A heuristic consists of an ordered list of non-repeating integers {*h1, h2, h3*}. The way the heuristic works is that from wherever the agent is on the circle, she can ask herself if the spot *h1* moves ahead has a higher score than her current spot on the circle. If so, the agent moves ahead to that spot on and if not, the agent stays at the same spot. They then move on to their next heuristic *h2* and repeat the same process with this heuristic. The process is repeated by returning to *h1* after trying *h3* or until the agent can no longer move to a higher score. From a given starting point on the circle, there is a unique stopping point the agent will fall on.

They measure the performance of an agent with a heuristic Φ by its *expected value.* Formally, for a starting point *v* and heuristic Φ, an agent's expected value E(V ; Φ) =

$$1/n \sum_{\mathrm{i}=1}^{n} \mathrm{V}\ (\Phi(\mathrm{i}))$$

(Hong and Page, 2004, 16386)

It is assumed here that each point on the circle is equally likely to be the starting point. Thus, it follows that for each starting point *i* and agent's heuristic Φ, the average of the epistemic payoff values for all possible starting points is the agent's *expected value.* An agent A exhibits more expertise than an agent B if agent A's expected value is greater than agent B's expected value.

As mentioned, Hong and Page are interested in group performance. A group of agents is represented as an ordered list {a1, a2, … ai}. From a given starting point, the first agent takes the group to the highest spot it can using its heuristic. The second agent goes next and leads the group to the highest spot using its heuristic. After all agents have attempted to locate higher-value solutions, the first agent then searches again. The search finally stops when no agent can locate a higher value. The group's performance is the average score the group receives starting from all spots.

For a class of agents defined by their heuristics, Hong and Page rank all the possible agents by their expected values and create two groups: a group that includes the 10 best agents (or the agents with the 10 highest expected values) and another group that includes 10 randomly selected agents. The model result is that groups with randomly selected heuristics outperform groups of with the best distinct heuristics. According to Hong and Page, the reason random groups outperform groups with the best heuristics is because the random groups are more functionally diverse. What functional diversity means in the model is the following. Consider the following heuristics: {3,7,8} and {3,4,5}. These two heuristics *overlap* in the first spot, because they share the same number, three, in that spot of the heuristic. The diversity between two heuristics is measured by the percent of places that the two heuristics do not overlap. Thus, the diversity percentage between heuristics {3,7,8} and {3,4,5} is lower than the diversity percentage between heuristics {1,2,3} and {4,5,6}. If $D(x_1, x_2)$ is the diversity percentage between two heuristics x_1 and x_2 or the percent of places that the two heuristics do not overlap, then the diversity percentage of a group that includes more than two heuristics is the average of all $D(x_i, x_j)$ where x_i and x_j are heuristics in the group and $i \neq j$.[2]

In one iteration of the computational experiment, Hong and Page compared one group of the 10 best agents to a group of 10 randomly selected agents. The expected values of the

[2] See Singer (2018) for a nice explication of the 'diversity trumps ability' model.

individual agents are first used to form the two groups and then they run the experimental trials. They ran 50 trials, where in each trial the group was randomly assigned a starting point on a circle of 2,000 spots. In this iteration of the experiment, the group of the highest performing agents had a diversity percentage of 70.98, whereas, the diversity percentage of the randomly selected agents was 90.99. Their results were the following. The performance score of the best problem-solvers was 92.56 and the performance score of the diverse problem-solvers was 94.53. They repeated this experiment while varying parameters such as the number of agents per group, the number of spots on the ring, etc. Despite the variations in parameters, Hong and Page repeatedly found the same result that "on average, the collective performance of the randomly selected agents significantly outperforms the group of the best agents" (2004, 16387).

Hong and Page develop a mathematical proof to explain the simulation results. This proof is general, in the sense that it does not rely on contingent features of the simulation (e.g., that there are 2,000 spots on the ring, etc.). The proof relies on four important assumptions: (a) agents are intelligent, (b) the problem is difficult, (c) agents are diverse, and (d) the best agent is unique. What these assumptions mean more specifically is the following. Assumption (a) ensures that all the agents are somewhat competent at the task as it states that no matter which alternative *x* the search process starts with, it does not terminate at an alternative *φ(x)* that is worse than *x*. The idea behind assumption (b) is that the problem of identifying the best answer must be sufficiently difficult such that no agent on its own is always going to be able to solve it. Assumption (c) guarantees that for any potential solution that is not the optimal solution, that there exists at least one agent who can find an improvement to this non-optimal solution. This assumption does not imply that for any particular group of problem-solvers, the group will in fact improve upon a non-optimal solution. For instance, a group that is homogenous in its heuristics may very well get stuck on one solution and this would be because no agent in this group uses a search rule that recognizes an improvement. Finally, assumption (d) says that there is exactly one heuristic that outperforms all the others.

Derived from these four basic assumptions, Hong and Page's proof includes two important lemmas. The first lemma is that as the group size becomes large, the independently drawn collection of agents will find the optimal solution with probability one (2004, 16388). Given that agents drawn independently are unlikely to have common heuristics, it follows that as the group size increases, the probably that the group will get stuck on one non-optimal solution converges to zero. The second lemma is that as the pool of problem-solvers grows large, the best problem-solvers will become similar and in the limit, the highest-ability problem solvers cannot be diverse (Hong and Page, 2004). To get an intuitive sense of this result, consider a set of randomly selected numbers from 1 to 100, each representing a score on an exam. As the set of randomly selected numbers expands, the group of the 10 best scores will become more similar, ultimately including only numbers 91 to 100 in the limit. Subsequently, the group of experts drawn from a large pool of problem-solvers have similar heuristics and often do no better than single best problem solver—who, by assumption (b), cannot always find the optimal solution.

The simulation results and corresponding mathematical proof support the idea that "diversity trumps ability" (2004, 16388). It has been discussed how the concept of diversity is represented in the model, but in order to get a better sense of the results of the model, we must consider what diversity refers to in the real-world for Hong and Page. The type of diversity the authors are concerned with in the model is called 'functional' diversity (or, similarly, 'cognitive' diversity). Functional diversity refers to a diversity of perspectives (ways of representing problems) and a diversity of heuristics (ways of generating solutions to problems). Moreover, functional diversity is influenced by what Hong and Page call ‘identity’ diversity, or the differences in people’s demographic characteristics, cultural identities, ethnicity, training, and expertise.[3] The reason functional diversity is partly caused by identity diversity is because a person’s unique perspective on a problem is often influenced by factors related to social identity and learning history. Although it is easy to confuse functional diversity with identity diversity because of how often they are correlated, the authors note how it is important to keep these forms of diversity separate. Functional diversity is conceptually distinct from its causes (cultural identity, gender, ethnicity) and its symptoms as well (differences in opinions, political affiliation, etc.).

It is difficult to overstate the academic and social impact of Hong and Page’s ‘diversity trumps ability’ results. The results have been cited over 3,000 times and have been utilized to argue for inclusiveness in democratic institutions (Landemore, 2012), university settings (UCLA, 2014), the armed forces (Fisher v. University of Texas, Austin, 2016) and in the sciences (Bright 2017; Martini, 2014; Stegenga, 2016). Helen Landemore (2012), for instance, applies Hong and Page’s ‘diversity trumps ability’ results to support her claim that a randomly selected political committee can be expected to produce smarter results than elected representatives, since random selection maximizes diversity. Along these lines, Jacob Stegenga (2016) has argued for the inclusion of experts and non-experts in science policy debates since epistemic diversity fosters the best results.

[3] For more on the link between functional diversity and identity diversity see: Nisbett & Ross 1980, Robbins 1994, Thomas & Ely 1996.

Outside of academic research, the ‘diversity trumps ability’ result has been used to support identity diversity in public institutions. For example, the UCLA College Diversity Committee discussed the results when arguing for a university policy that:

> takes seriously issues of diversity with respect to race, ethnicity, gender, socioeconomic status, sexual orientation, religion, disability, age, language, nationality, citizenship status and/or place of origin. (UCLA College Diversity Initiative Committee, 2014).

The idea here is that diversity with regard to these factors will produce better epistemic outcomes for the university. For similar reasons, the results have also been cited in the US Supreme Court case ‘Fisher vs. University of Texas, Austin’ (2016), where the results were implemented in arguments supporting gender and racial diversity.

The wide application of the ‘diversity trumps ability’ result assumes that the model suitably applies to real-world problem-solving contexts. In the next section, I question this shared intuition, by raising skeptical considerations surrounding the model’s representational adequacy.

1.4 Critiques of the Model

It is common for modelers to ignore or simplify real-world features of the phenomenon they are interested in. The demand for a model to fully represent the target system is untenable and misses the various epistemic functions models serve in light of utilizing simplifications, abstractions and idealizations (O’Connor, 2017; Weisberg 2013). Given this fact, a model can serve important explanatory functions even when it doesn’t represent the complexity of the target system.

Thus, the explanatory power of a model cannot be determined solely by how well it captures aspects of the real-world. Instead, one must consider the specific claim the model is meant to support and whether the model does a sufficient job of supporting this claim. There are issues with the application of the 'diversity trumps ability' model that stem from the way the problem-solving scenario and diversity are represented in the model. For instance, consider how the model has been utilized to argue for diversity in deliberative politics and science. Does the problem-solving scenario in the model adequately capture the complexity of deliberation or scientific reasoning?

To answer this question, let us look at how the Hong and Page model is applied to support diversity in these problem-solving contexts. Consider, first, Landemore's application of the 'diversity trumps ability' results to deliberative politics. The issue here is that a model characterized by agents finding a place along a circle of numbers cannot capture the complexities of individuals deliberating about policy issues in a meaningful way. First, democratic deliberation is oftentimes geared towards consensus or general agreement amongst deliberators. The types of problem-solving strategies utilized to come to a consensus view involve an exchange of reasons and rational reconsideration of one's preexisting beliefs. But, in the model, agents are assigned a set heuristic that does not change. Moreover, for the iteration of the Hong and Page model discussed here, agents work to solve the problem in a sequential order. Since deliberation requires an exchange of reasons in order to make a joint decision, the model cannot capture this fundamental epistemic property of deliberation.

If we now turn our attention to the model's application in support of diversity in science and university education, it is clear that the model similarly does not capture the complex epistemic properties of these contexts either. Consider how cognitive diversity in academic settings introduces a range of values and reasoning strategies. In order for these academic contexts to achieve its epistemic aims, individuals must be willing and able to articulate their positions in a way that is understandable and palatable to their diverse audience. Thus, one general issue is that the problem-solving context in the model does not address the fact that cognitive diversity in academic settings often introduces increased communication costs. Cognitive diversity often entails differences in what people value and how they rank such values, which may result in preference conflicts and cultural misunderstandings. Moreover, the worry is that such communication errors and value conflicts may outweigh the potential benefits brought by cognitive diversity. Therefore, a skeptical evaluator of Hong and Page's model may find that it does not give us insight into how diversity helps in problem-solving contexts, since communication costs brought by cognitive diversity is unavoidable in real life problem-solving scenarios.

Finally, consider a more general criticism that concerns the way expertise is defined in the model. As Grim et al. claim, genuine expertise seemingly requires being able to perform well on many problems of the same type, not just on a single problem. However, this important characteristic of expertise is not captured in the model. According to Hong and Page (2004), each ring of numbers is supposed to represent a specific problem the group of agents is out to solve. Hong and Page model each of these problem-solving tasks as completely random, in the sense that there is no correlation between the numbers assigned to the positions in the circles of different problem-solving tasks. As a result, different problem-solving tasks are represented by distinct circles and subsequently yield best performing agents with very different heuristics (Grim et al., 2019). An agent that is best-performing on one random landscape will likely do poorly on another landscape. According to Grim et al., this is troublesome, because it means no matter how linked two problems may be, by modeling those problems as distinct and random, best-performing heuristics cannot be expected to carry over from one problem to another similar problem (ibid).

The issue then is not the fact that the ‘diversity trumps ability’ model is idealized. Instead, the issue is that the model is *highly* idealized, such that, it is unclear how the model applies to the various problem-solving contexts it is meant to capture.

Given that these representational shortcomings are difficult to ignore, one may wonder why modelers and the academic community have, nevertheless, applied the model so widely. One way of diagnosing the situation is through considerations surrounding inductive risk. Recall that on Douglas’s view, scientists must consider both the epistemic and non-epistemic consequences of error when characterizing data and applying research results. As discussed, Douglas claims that scientists should raise their evidential standards in order to avoid causing negligent harm. But notice how the consequences of a ‘false positive error’ when characterizing and applying Hong and Page’s results are marginal, since the results support diversity, a social-good. On Douglas’s view, if one is willing to accept more false positives and wishes to avoid false negatives, then she should lower the standard for statistical significance. If we apply this line of reasoning to mathematical models, where the question under consideration is not of empirical adequacy but instead one that concerns when a model can be used and how, one way to lower one’s epistemic standards is to utilize a model that does not capture important features of the target system. The application of Hong and Page’s results may be an instance where the academic community has dropped their epistemic standards because the results support a social-good.

To see why, consider how highly idealized models of this sort are usually used to support much weaker claims. One way highly idealized models are used is to show a proof of possibility, or that some phenomenon can in principle be generated from a set of starting conditions (O'Connor, 2017; Weisberg 2007). Another way these models are used is to highlight the important causal factors of a phenomenon by highlighting the minimal conditions under which the phenomenon occurs (ibid). Despite this variation in epistemic goals, modelers generally agree that highly idealized models cannot be used to directly tell us truths about the social world. Thus, the use of the 'diversity trumps ability' model to directly support empirical claims and policy decisions is a deviation from the epistemic norms.

Admittedly, just because there are apparent issues with the application of the Hong and Page results does not necessarily imply that evaluators have dropped their epistemic standards. In response to this worry, consider how the model has two results: one stated result, that diversity outperforms experts, and an unstated result, that the group of experts must be non-diverse (Hong and Page, 2004). According to Hong and Page their

> …results provide insights into the trade-off between diversity and ability. An ideal group would contain high-ability problem solvers who are diverse. But, as we see in the proof of the result, as the pool of problem solvers grows larger, the very best problem solvers must become similar. In the limit, the highest-ability problem solvers cannot be diverse (16389).

Suppose the results of the Hong and Page simulation centered on the discussion of the trade-off between diversity and ability, such that, the authors instead claimed that the simulation results and corresponding mathematical proof showed that the highest ability problem solvers cannot be diverse. In this counterfactual scenario, would the Hong and Page model still be utilized in critical policy debates and decisions (e.g., for the US Supreme court or for a university diversity requirement)?

If it is assumed that the group of experts is homogenous, then it follows from this that when selecting an expert, one should look for an expert of a certain type. This is a dangerous implication of the model since it can legitimize disproportionately recruiting or hiring people from a particular social category. For instance, consider how although there have been reported gains in faculty diversity in the past two decades, the number of underrepresented minorities and women in tenure and tenure-track positions has only marginally improved and still remains disproportionately low (Finkelstein *et al.*, 2016). However, if we assume that there is one type of best expert and that functional and identity diversity overlap, then perhaps the fact that there are far more white male academic experts is justified. Along these lines, given that there are currently more white male academic experts, one can utilize this fact to justify *continuing* to disproportionately hire men—after all, we should expect groups of experts to be homogenous anyways.

Assuming, as we are, that the reaction to the Hong and Page model would be different if the discussion centered on the trade-off between ability and diversity, on what grounds would the model be challenged? Here the results of the model undermine the social-good of diversity in problem-solving scenarios and along these lines, the model can be utilized to support policy measures that threaten inclusivity. Given the consequences of erroneously accepting the model and its results, evaluators will likely raise their evidential standards. In raising their evidential standards, the simulation results could be challenged on the same grounds I have previously discussed in this paper, i.e., the model does not capture many of the important epistemic features diversity produces in problem-solving contexts.

The main moral to be drawn from this section is that given the positive social implication of the 'diversity trumps ability' model, Hong and Page, as well as those who have utilized this model since its publication, have arguably adopted unusually low epistemic standards. If my analysis is correct, an interesting question confronts us: given the positive social implications of the model, were the authors and academic community justified in dropping their epistemic standards? In the next section, I will consider this question in more detail and will eventually conclude that despite the model's initial plausibility, there are apparent dangers in dropping our epistemic standards in this case.

1.5 Inductive Risk and Mathematical Models

Given the positive social implications of the diversity trumps ability model, was the academic community justified in dropping their epistemic standards? Answering the central question of this section will require us to consider the epistemic features of mathematical models more generally. A particular property that is relevant here is the flexibility surrounding the application of mathematical models. As we will see, a single model can serve a variety of explanatory roles in various arguments, even when applied to the same target system.[4] Mathematical models are flexible, in the sense that they are representational structures that can provide multiple distinct conclusions. In this section I will argue that this flexibility allows for mathematical models to take on a life of their own, such that, it is difficult to calculate their inductive risk.

One way in which a single model can serve distinct explanatory roles is through its application to a variety of target systems. One recognized example of this is the use of the

[4] For more on the explanatory plurality of models see: Downes (1992), O'Connor (2017), and Jhun, Palacios, and Weatherall (2017).

signaling games to model between and within organism communication. The standard signaling game as described by David Lewis (1969), is a model of information transfer between two agents. This model of information transfer between organisms has been utilized to develop a theory of convention and meaning (Lewis 1969) as well as the emergence of language (Huttegger 2007; Huttegger & Zollman 2011; Harms 2004; Skyrms 2010). In addition, signaling games have been applied to various biological and cognitive systems to better understand their function including: the perceptual system (O'Connor, 2014), genetic information transfer (Calcott, 2014; Godfrey-Smith, 2014), and neural interactions (Cao 2014; Skyrms, 2010).[5]

It is also the case that a model can generate multiple distinct conclusions when applied to the same target system. For example, consider how the signaling game model, applied to a single target system—between organism communication, can generate distinct conclusions. One conclusion that is derived from signaling games is that it is possible to understand the convention of meaning without dissolving into circularity or regress (Lewis, 1969). This is a 'how-possibly' type of conclusion, as it shows that it is in principle possible to derive meaning from convention.[6] Moreover, signaling games provide a distinct 'how-actually' conclusion. Signaling games offer a framework for analyzing how a conventional language *actually* emerges from interacting agents that are less than fully rational (Huttegger, 2007; Skyrms 1996; 2000; 2004). Relatedly, the signaling game model helps explain how interacting agents spontaneously learn to

[5] Skyrms (2010) only briefly comments that neural interactions can count as a signaling system and so, the extent to which he is committed to the applicability of signaling games to neural systems is unclear. Regardless, neurons are a potential candidate for the Lewis-Skyrms model.

[6] The 'how-possibly' conclusion derived from the signaling game model is first described by David Lewis in *Convention* (1969). Lewis's account is a response to Quine, who claimed that it is impossible to derive meaning from convention. Quine's argument takes the following form. Conventions arise by agreement between agents. However, in order to arrive at an agreement, agents must have some rudimentary language. Now the origins of this rudimentary language must be explained. Thus, according to Quine, conventions of meaning cannot be generated without turning into a regress or circularity. Signaling games provide a framework in which meaning is derived from convention.

signal (Skyrms, 2010) and how conventional language is maintained in a population (Huttegger, 2007).

It is important to note that empirical theories can similarly support multiple distinct claims when directed at a specific target phenomenon. For instance, consider two distinct consequences derived from Einstein's theory of special relativity. Special relativity predicts that the time lapse between two events is not invariant from one observer to another, and instead, depends on the relative speeds of the observers' reference frames. This prediction was later confirmed by the Hafele-Keating 'clock' experiment (1971). Another implication of special relativity theory is the relativity of simultaneity, or the idea that the simultaneity of two events is dependent on the reference frame of an observer. More specifically, two events happening at two different locations can occur simultaneously in the reference frame of one observer may, nonetheless, occur non-simultaneously in the reference frame of another observer.

Nevertheless, mathematical models are especially flexible for a number of reasons. Consider how mathematical models can be generated from a minimal number of assumptions, without capturing the complexity of the target system. Such models are often explanatory in virtue of leaving out real-world features, as they better capture the essential causal factors of the target system by doing so. Since these models prioritize causal transparency over complexity and nuance, we often see the same model applied to a variety of target systems—as long as the target systems share some minimal causal structure. Moreover, as discussed, mathematical models can derive remarkably distinct conclusions when applied to a particular target system. The 'diversity trumps ability' model is a perfect example of this. Here we see the model derives two very different conclusions in its explanation of the role of diversity in problem-solving contexts.

Let’s now return to the question postponed: given the positive social implications of the model, were the authors and academic community justified in relaxing their epistemic standards? One way of tackling this question is to consider a similar counterfactual to the one posed previously. Suppose after publishing their simulation results, Hong and Page went on to retract the take-away that diversity trumps ability and instead, focused on the understated result that as the pool of problem solvers grows large, the very best problem solvers become less diverse. If Hong and Page re-described their results in this way, what reasons can we give for why the model is inadequate?

The worry is that in dropping their epistemic standards in response to the ‘diversity trumps ability’ result, the academic community has subverted the grounds to dismiss this socially pernicious result. At this point, one might respond that this isn’t a problem for Hong and Page for the following reason: it is common for mathematical models to have extraneous or artificial results and it is usually implied that these results are irrelevant to the model’s main conclusion. An example of an extraneous result derived from the model is that if an agent possesses all of the heuristics, the group the agent is in cannot improve. The idea here being, given that the agent possesses all of the heuristics, it can always maximize how it goes around the circle. This result is extraneous because it is unrelated to the question of what role diversity and ability play in a group’s problem-solving performance. In fact, this result implies that there is no reason to have groups of problem solvers in the first place, because there is some individual that can outperform everyone else.

Consider how we cannot easily dismiss the socially pernicious result of the model because it is not an extraneous result. The result that the group of highest ability problem-solvers cannot be diverse is essential to the main diversity trumps ability result. What makes the diverse group outperform the group of best problem-solvers is partly due to the fact that the group of best problem-solvers is homogenous. Recall that the single best problem solver cannot always find the optimal solution from every possible starting point. Therefore, the homogenous group made up of the best problem-solvers is likely to get stuck on a non-optimal solution, which allows for the diverse group to reliably outperform the homogenous group.

Notice, then, how the socially pernicious result is epistemically on par with the 'diversity trumps ability' result, as it is either the case that both results are derived from the model, or neither result is derived. Hong and Page make this explicit in their discussion of the parameters needed to derive the simulation results. Hong and Page claim that the 'diversity trumps ability' result

> …relies on the size of the random group becoming large… At the same time, the group size cannot be so large as to prevent the group of best problem solvers from becoming similar… As the group size becomes larger, the group of the best problem solvers becomes more diverse and, not surprisingly, the group performs relatively better (2004, 16389).

In other words, when group sizes are too large, groups of expert problem solvers are no longer homogenous and subsequently, they perform better than the diverse groups.[7] This importantly shows the interdependence between the two results—without the result that the group of best experts are homogenous, the intended 'diversity trumps ability' result cannot be derived.

[7] Notice how a different notion of group size is being invoked here. Previously, we saw Hong and Page claim that "as the pool of problem solvers grows larger, the very best problem solvers must become similar" (16389). This refers to the set of problem solvers that the members of the diverse group and expert group are selected from. Here Hong and Page are discussing the group size of the diverse group and expert group themselves.

The discussion thus far illustrates why we cannot reject the socially pernicious result in isolation. However, perhaps the academic community must instead reject the model entirely in this hypothetical. For the academic community to reject the Hong and Page model only after the result that the group of expert problem-solvers must be non-diverse is later emphasized would be problematic also. In step with the logic of Douglas's view on inductive risk, to reject the simulation results in this scenario where the model stays the same but different morals are drawn would be for values to directly contribute to the weight of the evidence. The fact that the model now supports a claim that grates against our ethical intuitions serves as evidence to reject it and recall that on Douglas's picture, in the phases of science where evidence is interpreted, and hypotheses are tested, values shouldn't play a direct role of this sort. If values played a direct role in the assessment of evidence, a scientist's preference for a particular outcome could act as a reason for that outcome, or for the rejection of a disliked outcome (Douglas, forthcoming). Such a situation would impede critical evaluation of research, as there would be no shared standards for determining the validity and empirical adequacy of another's work. Thus, in order to avoid ad hoc theory rejection, the academic community should have rejected the model at an earlier stage.

One may think that the arguments presented in the section merely show that Hong and Page and the academic community have miscalculated the inductive risks involved in the 'diversity trumps ability' model. They erroneously assumed that the 'diversity trumps ability' model has low inductive risk and so they dropped their epistemic standards. However, the point is not that the academic community merely miscalculated the model's inductive risk. It is instead that inductive risk calculations cannot be appropriately conducted for mathematical models like the 'diversity trumps ability' model. This is because, there is flexibility in what results can be derived from the model. This flexibility makes some risks unforeseeable—like the risk of concluding Hong and Page’s model supports the socially pernicious result.

As we saw in the case of the diversity trumps ability model, mathematical models often derive distinct results, even when applied to a single target system. Moreover, the differences in the model’s results can be stark, such that, in dropping our epistemic standards in response to the result that supports a social-good, we undermined the conditions for rejecting a result that supports a social-ill on weak epistemic grounds. Since the flexibility of derived results is a feature of mathematical models in general, the example of Hong and Page’s model illustrates why the academic community should not drop their epistemic standards in their evaluation of mathematical models, even when supporting a social good.

1.6 Conclusion

To conclude, I would like to reemphasize the aspects of novelty presented in this paper. The paper has shown that mathematical models generate inductive risks like those described by Heather Douglas. In the case of the ‘diversity trumps ability’ model, this resulted in modelers lowering their epistemic standards because the model supports a social-good. However, the paper also illustrates that inductive risk calculations are difficult to do for mathematical models. For instance, consider how mathematical models can be applied to multiple target systems or can generate such distinct results even when applied to a single target system. It is for this reason that the risks in generating and applying mathematical models are often unforeseeable.

Second, the paper attempts to answer the question of whether modelers should lower their epistemic standards when a model’s results support a social-good. Through the example of Hong and Page’s diversity trumps ability model, I argued that dropping our epistemic standards is problematic, as it can result in situations where we are no longer able to evaluate claims based on independent epistemic grounds. This implication was made evident through the discussion of the counterfactual situation in which Hong and Page's model supports the idea that the group of best problem solvers cannot be diverse. The issue here was that to dismiss the model because of the newly emphasized socially pernicious result would be for values to play a direct role in our evaluation of the model. Thus, in order to avoid ad-hoc theory rejection, we should keep our epistemic standards high regardless of what results the model supports.

CHAPTER 2

WILL TRANSPARENCY BOLSTER OR HINDER PUBLIC TRUST IN SCIENCE?

2.1 Introduction

The concept of transparency has received a great deal of attention in the philosophy of science literature, especially as it relates to communicating scientific studies and findings to non-experts. Many philosophers claim that transparency is important for establishing public trust in science (Douglas 2009; Elliott 2017; Kitcher 2011; I. de Melo-Martin and Intemann 2018; Stanev 2017; Williams 2002) whereas some philosophers instead argue that transparency about scientific practices could actually promote unwarranted skepticism (John 2018; Kovaka 2019). Despite their disagreements, both camps similarly claim that transparency plays a particularly important causal role in shaping the general public's perception of science.

Non-experts hold a particular folk philosophy of science, or a set of idealized assumptions concerning how science should work that sometimes conflicts with actual scientific methodology (John 2018; Kovaka 2019). For example, non-experts often assume scientific consensus is much more common than it actually is. Another common folk view of scientific methodology is that the social structures of science should always encourage debate when, in fact, shutting down certain positions is often epistemically useful and important. The worry is that if scientists are completely transparent, they will expose non-experts to practices that conflict with their folk philosophical assumptions, thus causing more skepticism in science. It is important to note that some defenders of transparency agree that skepticism in science is often caused by a discordance between actual scientific practice and non-experts' assumptions. For example, Douglas (2015) and Elliott (2017) claim public trust in science only has the hope of being successful if transparency is coupled with proper philosophy of science education.

However, if we want to better understand what role transparency plays in the public's perception of scientific claims, we must realize that the issue is multifaceted and complex. In this paper, I argue that claiming we shouldn't be transparent because non-experts hold a false folk philosophy of science oversimplifies the issue. When determining whether transparency will bolster or hinder public trust in science, we must first consider the way scientific information is being communicated to the public. More specifically, we should ask ourselves: 'has this scientific topic been politicized?' and 'are special interest groups playing a role in the dissemination of the scientific information?'. Answering these questions is essential to determining the effectiveness of transparency. This is because, transparency is only effective if the public receives an accurate, unbiased account of what most scientists actually claim.

In order to support the claim that the transparency debate in philosophy is mistakenly focused on non-experts' knowledge of scientific methodology, I first outline Stephen John's argument against transparency. John uses Climategate—an incident where climate scientists' emails were leaked—to support his claim that transparency can hinder public trust in science when scientific methodology diverges from the public's folk philosophy of science. I argue that this example doesn't serve as a general argument against transparency. This is because, it isn't clear that what caused climate skepticism was a transparent presentation of climate science methodology. Instead, special interest groups and conservative media presented the climate scientists' emails and methodological approaches as unscientific and pernicious. I then discuss how we still see skepticism in science in cases where experts' methodological approaches fit with the public's idealistic assumptions of proper scientific methodology. This further supports my argument that agreement with a folk philosophy of science isn't the most important factor when it comes to transparency. I then consider an example where transparency bolstered public trust in science, despite methodological disagreements between experts and non-experts. Finally, I discuss why answering the two related questions: 'has the scientific topic been politicized?', and 'are special interest groups playing a role in the dissemination of scientific information?' is fundamental in determining the relevance of transparency.

2.2 John's argument against transparency

Stephen John claims that transparency can destroy non-experts' trust in science. To support this claim he first discusses how non-experts learn from experts. John claims that learning from experts can be modelled as a two-premise inference. The idea here is that our everyday practices of deference to scientific experts are grounded in these two general assumptions. The first premise is what he calls the sociological premise. According to the sociological premise:

> Institutional structures are such that the best explanation for the factual content of some claim (made by a scientist, or group, or subject to some consensus) is that this claim meets scientific 'epistemic standards' for proper acceptance (John, 77).

The sociological premise states that institutional structures of epistemic groups typically ensure that its members only assert and accept claims when those claims meet community-based standards. However, standards for acceptance between different epistemic communities greatly varies and so the sociological premise alone does not tell us when we are warranted in accepting the claims of a particular epistemic communities. For example, the sociological premise holds true for a group of highly regarded astrologers—they only assert claims that meet astrological standards for assertion. Despite this fact, it is obvious that consensus amongst astrologers about some claim does not carry as much epistemic weight as a report of consensus amongst scientist about a claim. This leads us to what John calls the epistemological premise or the idea that we all implicitly assume that

> if some claim meets scientific epistemic standards for proper acceptance, then I should accept that claim as well (John, 77).

Unlike other domains of inquiry, most people living in modern societies agree that the epistemic standards characteristic of scientific research communities should govern our beliefs. The scientific community can be contrasted with the astrological community, where we do not hold an analogous assumption when it comes to the astrological standards for proper acceptance.

John's account of how non-experts learn from a scientific community rests heavily on the notion that the institutional structures of science ensure that only claims that meet rigorous epistemic standards are accepted. For example, some rigorous epistemic standards of science include replicability, peer-review, statistical significance—all of which require transparency between scientists. Since transparency within the scientific community is required to meet epistemic standards, it may seem "unreasonable—or even unethical" for scientific experts to withhold information from non-experts (p. 80).

Nevertheless, John argues that scientists need not follow the norm of transparency when communicating to non-experts. According to John, transparency about knowledge production does not necessarily promote the flow of true beliefs among non-experts. The reason why has to do with the fact that non-experts often have false ideas about how scientific methodology works. In other words, non-experts can have false beliefs about what epistemic practices make the sociological premise true. These incorrect assumptions non-experts have about how science should work is what John calls a “folk philosophy of science” (p. 80). Some examples of a folk philosophy of science include the belief that there is usually consensus in the scientific community and that lack of consensus indicates there is no fact of the matter. Another example is that the social structures of science should always encourage debate and discussion when, in fact, shutting down certain views is sometimes epistemically useful.

John highlights this point through the example of Climategate, an incident where 1,000 emails between climate scientists at the Climate Research Unity (CRU) of the U.K.’s University of East Anglia were leaked (John, 2018). Climate skeptics claim that these emails show scientific misconduct that amounts to fabrication of anthropocentric climate change. More specifically, skeptics claimed that these emails showed that climate scientists at the University of East Anglia were engaging in unscientific practices including confusing correlation with causation, refusing to publish papers by certain authors, refusing to include certain data sets in their analysis and making assertions based on uncertain modeling work. Since non-experts have a false folk philosophy of science, they were easily convinced that these practices are unscientific when, in fact, such practices are typical and acceptable within the scientific community. For example, as John states, inferring causation from sufficient types and kinds of correlations is justifiable scientific procedure (Papineau 2012); refusing to publish certain research is often necessary to promote progressive research projects (Lakatos 1978); and discarding recalcitrant or unusual data sets is a warranted response to uncertainties in data collection (McAllister 2012).

As a result, non-expert trust in science is often warranted yet fragile—if a non-expert learns a scientist reaches her results in a way that deviates from a folk philosophy of science, then the non-expert is likely to remove her trust. It is for this reason John claims that demands that scientists should be transparent about the research process to non-experts are not well grounded. As long as non-experts hold a false folk philosophy of science, transparency can be epistemically harmful, as we saw in the example of Climategate.

One may wonder why the solution to the problem John presents is to limit transparency between experts and non-experts. One might think that the natural solution is to combat non-experts' folk philosophy of science. If non-experts knew how scientific theorizing really worked, then experts could be transparent with non-experts in the same way experts are transparent between one another. John agrees that the long-term goal should be to change non-experts' folk philosophy of science through science education reform. Unfortunately, when it comes to climate science, we cannot wait for a better world as immediate action is needed to mitigate anthropocentric climate change. Thus, John concludes that the long-term benefits of transparency does not outweigh the short-term costs.

It is important to note that John isn't alone in focusing on non-experts' misguided views on scientific methodology when discussing public skepticism in science. For instance, Karen Kovaka has similarly claims that

> dispelling misconceptions about the nature of science may force people to re-evaluate their rejection of climate change and ultimately help them change their minds (2019, p. 3).

Along these lines, Heather Douglas argues

> …we need a public that has a better understanding not just of scientific facts but, more importantly, of the nature of science… understanding the nature of science is crucial for both being able to properly process science in the news, and more importantly, for engaging with scientific and technological controversies. (2015, p. 10).

Notice how Douglas argues for the importance of an accurate folk philosophy of science in public understanding. And in particular, both John and Kovaka claim that skepticism in climate science can be mitigated if the general public had a more accurate conception of scientific methodology. However, there are reasons to doubt that we can derive general conclusions about transparency's effect on non-experts based on the Climategate example. In the next section, I will explore some reasons for why we may doubt the Climategate example can generalize in the way John claims it does.

2.3 Climategate: a peculiar case of transparency

To determine whether or not the Climategate example generalizes, we must ask ourselves whether non-experts' false folk philosophy of science sufficiently explains public mistrust in climate science after the leaked emails, or if there were other factors involved. If there were other factors that influenced the public's mistrust of climate science, then the central issue may not be a matter of forced transparency on the part of climate scientists.

To answer this question, we must consider how public perception changed after Climategate and for what reasons. Climate change denialists first broke the story and argued that the emails showed global warming was a scientific conspiracy and that scientists manipulated climate data and suppressed critics. The mainstream media then picked up the story at the same time negotiations over climate change mitigation began in Copenhagen in December 2009. Due to the timing of the leak, scientists, policymakers, and public relations experts claim that the release of the emails was intended to undermine the aims and objectives of the climate conference. Although the American Association for the Advancement of Science (AAAS). The American Meteorological Society (AMS), and the Union of Concerned Scientists (UCS) released statements supporting the scientific consensus that the Earth's mean surface temperature had been rising for decades in response to the controversy, in many ways the damage had already been done.

It is important to note that the way the emails were discussed by conservative media outlets was not neutral. Instead, quotes from the emails were taken out of context and were made to appear more controversial than they really were. For example, the most quoted phrase took words from an email written by Phil Jones, which referred to a graph he was preparing for the World Meteorological Organization on the status of climate change in 1999. Jones wrote:

> I've just completed Mike's *Nature* trick of adding in the real temps to each series for the last 20 years (i.e., from 1981 onwards) and from 1961 for Keith's to hide the decline. (Washington and Cook, 2011, p. 44)

The graph showed three series of paleoclimate reconstructions based on tree ring, coral, ice core, and lake sediment samples along with historical and instrumental records. The 'trick' refers to a technique for combining data series. Climatologist Michael E. Mann published a paper on temperature trends in *Nature* in 1998 which combined various proxy temperature records and related them to actual temperature records. It included a figure later notably called the 'hockey stick' graph which clearly distinguished between this proxy and instrumental data. According to Mann, the 'trick' is simply a concise way of showing these two kinds of data together, while still clearly labeling the two types of data. Thus, he claims that there is nothing hidden or inappropriate about this method and that this method of combining proxy and instrumental data had be corroborated by numerous statistical tests and matched thermometer readings taken over the past 150 years. The phrase 'the decline' referred to the divergence problem in dendroclimatology—a well-known issue that while thermometer records indicate a substantial warming trend up until the late 20th century, many tree rings from these areas do not display a corresponding change in their maximum latewood density (Oreskes and Conway, 2010).

The phrases ‘trick’ and ‘the decline’ were misquoted to mean a “trick” to hide the decline in measured global temperatures. This was of course entirely false, since 1998 had been the warmest year on record. However, this did not stop conservative news outlets, like Fox News, and politicians to misrepresent what Jones meant. For instance, in 2009 US vice presidential candidate Sarah Palin said the phrase showed a “highly politicized scientific circle” that had “manipulated data to hide the decline in global temperatures” and Senator Jim Inhofe of Oklahoma said “of course he means hides the decline in temperatures” (Pearce, 2010).

In fact, climate change deniers have been conducting a smear campaign against climate scientists and their work since the 1960s. In their book *Merchants of Doubt,* Naomi Oreskes and Eric M. Conway identify the role of special interest groups in both the global warming and tobacco smoking controversies. The book draws a number of parallels between these two cases and focusses specifically on how special interest groups and think tanks set up by the oil and tobacco industries actively created controversy on these issues. The general strategy of these special interest groups was to present legitimate scientific methodology as flawed. As Oreskes and Conway claim, the idea here is that “if they cannot contest the scientific facts, then the next best thing is to go after the scientific methodology of those claiming things they don’t like” and this is exactly what we see in the Climategate example, a full-blown attack against the efficacy of climate research (Oreskes and Conway, 143). It is no surprise that non-experts were skeptical of the research methods exposed in the Climategate case, after all, their methods had been under attack for a number of decades prior to the scandal.

Thus, transparency did play a role in inducing skepticism, as it allowed skeptics to make effective, if spurious, arguments to the public about the illegitimacy of climate science. However, the large role special interest groups played in mischaracterizing emails and portraying climate science as illegitimate indicates that transparency alone did not cause climate skepticism, like John claims. Instead, the important factor here is the industrial actors and how these actors utilized transparency as fodder for their anti-climate science agenda. For this reason, it is unclear whether John's example of Climategate generalizes as he claims it does. At the very least, the Climategate example doesn't show why transparency is harmful or dangerous in the absence of nefarious industrial actors. Most scientific debates don't include special interest groups that spread misinformation like in the case of climate change, or at least not to the same degree. More importantly, when we think of transparency, we think of an *unbiased* presentation of the scientific facts and as discussed this isn't what we find in the Climategate example.

Furthermore, evidence indicates that knowledge of scientific facts and methodology isn't necessarily the most salient driving factor when it comes to trust in climate science. For instance, a PEW study (Funk et al., 2019) found that among those with a 'high' degree of environmental science knowledge, 44% said environmental scientists provide fair and accurate information about their research, whereas, among those that have a 'low' degree of environmental science knowledge, 25% said environmental scientists provide fair and accurate information about their research. Presumably, those with knowledge about environmental science would likewise know that the methodology often deviates from folk philosophy of science norms. However, we see that the percentage of trust in environmental science between these two groups isn't as different as we might expect.

Compare this statistic with how trust in environmental science deviates along political lines. In the same PEW research study (Funk et al., 2019), 70% of democrats said they have a mostly positive view of environmental scientists, whereas, only 40% of republicans said the same. Moreover, 47% of democrats agreed with the claim that environmental scientists provide fair and accurate information whereas only 19% of republicans agreed.

It is important to note that when controlling for each of these variables, political affiliation seems to be a better predictor of trust in environmental science than scientific literacy. For example, when controlling for education level, democrats on average trust the results of climate scientists more than their conservative counterparts (Gauchat, 2010; Hornsey et al., 2016). Moreover, although increased education level, in general, positively correlates with more trust in climate science, it should be noted that this main trend must be qualified by a moderate effect: research in the U.S. using representative samples suggest that the link between scientific knowledge and trust in climate science is more positive among Democrats than Republicans (Hornsey et al., 2016). This seems to indicate that when controlling for education, politics still plays a significant role.

In this section, I discussed how climate change skeptics spread misinformation in order to fracture public trust in science. Due to this misinformation campaign, it is unclear that transparency was the main driver of public skepticism in climate science. I then argued that even if transparency played a significant role in public mistrust in climate science, given the unique role special interest groups played in climate change denial, the pernicious effect transparency possibly had in this case can’t generalize to other scientific debates. John’s argument against transparency relies on the fact that there is a discrepancy between non-experts’ idealized views and actual scientific methodology, such that, mistrust in science is generated when research methods that don't correspond to a folk philosophy of science are made transparent. Finally, I argued that political leaning is more strongly correlated with trust in environmental science than scientific education.

In the next section, I will further question John's account by looking at the example of vaccines, where research methods corresponded to the public's folk philosophy of science and yet, nevertheless, we see public mistrust in science. Through a careful examination of this example, I hope to show that public mistrust in science is a complex issue caused by a set of factors and thus, John's account of mistrust in science oversimplifies the issue.

2.4 Is science skepticism usually rooted in a methodological disagreement between experts and non-experts?

According to John, the communicative obligations scientists have towards non-experts should be grounded in claims about what will further non-experts' epistemic interests (p. 82). More specifically, we saw him argue that when the scientific community and the general public's standards for acceptance diverge, scientists shouldn't be transparent about their methods. In this section, I will argue that when we seek to bring about conditions necessary for warranted trust, we shouldn't myopically focus on the general publics' views about scientific methodology. Scientific skepticism stems from many factors and focusing on non-experts' methodological misconceptions oversimplifies the issue. In supporting this argument, I will look at the case of vaccine safety, where scientists did everything right vis a vis the folk philosophy of science, but we still see skepticism in part because of influencers and propogandists.

On John's account, in cases where scientific methodology corresponds to common folk philosophy of science assumptions (e.g., debate and discussion should always be encouraged, refusal to publish and engage with certain research projects is impermissible, large sample sizes indicate validity, etc.) we should expect transparency to at least minimally curb scientific skepticism. If we do find cases where scientists are operating within the "extremely idealized… normative models of scientific inquiry" that non-experts are "routinely exposed to" and nevertheless see mistrust in the science, then this seems to imply that other factors besides a folk philosophy of science helped to generate this mistrust (John, 81).

There is reason to believe that the measles, mumps, and rubella (MMR) vaccine safety research is an example where scientific methodology corresponds to a folk philosophy of science. For example, earlier studies determining the efficacy of these vaccines reported that:

> clinical studies of 284 triple seronegative children, 11 months to 7 years of age, demonstrated that M-M-R II is highly immunogenic and generally well tolerated. In these studies, a single injection of the vaccine induced measles hemagglutination-inhibition (HI) antibodies in 95%, mumps neutralizing antibodies in 96%, and rubella HI antibodies in 99% of susceptible persons (CDC report, 2020).

Furthermore, it has been reported that the efficacy of measles, mumps, and rubella vaccines was established in a series of double-blind controlled field trials which demonstrated a high degree of protective efficacy by the individual vaccine components (Cutts 1991, Hilleman 1967, 1968; Weibel 1967).

Compare the clinical studies conducted to establish the efficacy of the MMR vaccine, which included a large sample size and double-blind controlled field trials, to the studies published by vaccine skeptics. For example, in 1998, Andrew Wakefield and 12 of his colleagues[8] published a case series in *Lancet* which suggested that measles, mumps, and rubella (MMR) vaccine may predispose children to behavioral regression and pervasive developmental disorder. Despite the small sample size (n=12), the uncontrolled design, and the speculative conclusions, the paper was very influential.

Instead of dismissing these methodologically flawed studies conducted by vaccine skeptics, the scientific community responded in a way that corresponds to a folk philosophy of science—they engaged in debate and discussion with anti-vaxxers. In fact, the scientific community took the anti-vaccination movement surprisingly seriously and conducted many carefully performed scientific studies to determine whether there is a link between vaccines and autism. One of these studies was supervised by the CDC and the Danish Medical Research Council in November 2002. The study followed more than 500,000 children over 7 years and found no link between MMR vaccination and autism (Center for Disease Control and Prevention, 2020). Also included is an April 2006 study conducted by the National Institution of Child Health and Human Development of NIH and the CDC that assessed data from 351 children with autism spectrum disorders and 31 typically developing children. This study similarly found no link between MMR vaccination and autism (ibid.). More recently, a study from September 2008 published in Public Library of Science was conducted to determine whether the results from an earlier study that claimed to find measles virus RNA in the intestinal tissue of a specific group of autistic children could be confirmed. The results of this earlier study could not be confirmed, once again corroborating the fact that there is no link between MMR and autism (ibid.).

[8] Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. Wakefield AJ, Murch SH, Anthony A, Linnell J, Casson DM, Malik M, Berelowitz M, Dhillon AP, Thomson MA, Harvey P, Valentine A, Davies SE, Walker-Smith JA

Lancet. 1998 Feb 28; 351(9103):637-41.

Unfortunately, we still see widespread skepticism concerning the safety of vaccines. A recent survey from the Pew Research Center show that 9 percent of Americans think the MMR vaccine is not safe and another 7 percent is unsure (Pew Research Center, 2015). Moreover, among the people who are actually parents of minors, the number of vaccine skeptics is at 13 percent. In fact, only 80 percent such parents agreed that the MMR vaccine is safe (ibid.).

Here we see a case where the scientific experts' studies and discourse resemble non-experts' folk ideals of good scientific methodology—large sample sizes, peer reviewed studies, engaging in with vaccine skeptics, etc. Compare this to how small the sample size is for Wakefield's original study, how it wasn't peer reviewed and how anti-vaccine 'scientists' never formally engaged with articles that demonstrate the safety of vaccines. If non-experts' methodological ideals strongly influence trust in science, then we might expect non-experts to trust scientific studies that fit these methodological ideals. However, evidence indicates quite the contrary: a recent report evaluated the effectiveness of messages designed to reduce parental misperceptions and increase vaccination rates, including messages about peer review studies with large sample sizes establishing vaccine safety, demonstrated that these messages were not only ineffective, they even reduced the intention to vaccinate in some groups of parents (Nyhan et al., 2014).

The discussion thus far implies that a folk philosophy of science was not the primary cause of vaccine safety skepticism. Epidemiologists were conducting their research according to a folk philosophy of science and nevertheless, we still see public mistrust in vaccine safety.

What then is causing non-expert's mistrust if it is not their folk philosophy of science? This is a difficult question without a single straightforward answer. One plausible answer is that mistrust in vaccines stem from a discrepancy between epidemiologists' and the general public's assessment of risk (Hicks, 2017). An epidemiologists' assessment of risk is determined by the frequency of a hazard across an entire population given the implementation of the vaccine. This risk is then compared to the overall public health benefits in order to determine whether the risk is worth taking. For example, a mandatory vaccination policy that promised to prevent 5,000 mumps cases, even while leading to 50 cases of serious side effects, might still be considered worthwhile (Hicks, 2017). Compare this assessment with the way a worried parent may assess the risk of a vaccine. When assessing risk, parents tend to focus on their child instead of an overall social balance of costs and benefits. More specifically, in determining whether to vaccinate their children, parents may narrowly focus on the increased risk of exposing their child to harsh side effects and ignore the public safety risks of not vaccinating their child (Hicks, 2017).

Another answer to this question has to do with the way anti-vaxxers exploited the way people share information and learn from one another (O'Connor and Weatherall, 2019). For instance, after learning that autism rates were particularly high in a Somali community in Minneapolis, anti-vaccine supporters distributed fliers that reported a link between vaccinations and autism (ibid). In fact, Andrew Wakefield visited Minneapolis many times between 2010 and 2011 to speak with Somali parents of autistic children. Following these visits was a drop in vaccination rates in the Somali community from 92 percent in 2004 and 42 percent in 2014 (O'Connor and Weatherall, 2019). As O'Connor and Weatherall claim, anti-vaxxers were taking advantage of conformity effects by pushing their views on a close-knit group that is already susceptible to their message (ibid).

Notice how public skepticism in science is a complicated issue often generated by more than one factor. For the MMR vaccine, we saw that both the discrepancy in risk assessment between parents and epidemiologists as well as the way anti-vaxxers exerted influence by targeting close-knit communities contributed to vaccine mistrust. While it is plausible that a folk philosophy of science plays a role in vaccine skepticism, the discussion suggests that it isn't the most important factor.

In what follows, I will push back on the claim that mitigating methodological disputes between scientists and laypeople requires scientists to be less transparent. Medical research for an AIDS cure during the 1980s involved a direct clash between experts and non-experts' methodological assumptions. As will be discussed, receptivity to the methodological critiques and further transparency from the scientific community is eventually what quelled non-experts' skepticism.

2.5 An example of transparency fostering public trust in science

Let us now turn to the example of AIDS research in the United States during the height of the epidemic. Philosophers studying methodological disputes between experts and non-experts tend to focus on the deleterious role non-experts' methodological assumptions play in knowledge production. However, the example of AIDS research shows that non-experts can in certain circumstances become genuine participants in the construction of scientific knowledge in virtue of their methodological disagreements with experts. This is meant to illustrate that divergences between expert and non-expert methodological assumptions alone isn't enough to produce scientific skepticism.

Non-experts in the form of gay rights activists disagreed with the methodological approach of AIDS research. More specifically, they challenged the use of randomized clinical trials and questioned the exclusion of subjects and requirements that subjects avoid participating in multiple trials. This was a direct challenge to orthodox scientific methodology when it comes to drug trials. In challenging assumptions about the randomized clinical trial standard, activists forced the scientific research establishment to design clinical trials that could serve AIDS patients, most of whom suffered from multiple health problems that needed simultaneous treatments, and gained the inclusion of a more diverse subject population in antiretroviral trials (Epstein, 1996).

From the activists' perspective, experts' emphasis on methodological purity reflected a dangerous abstractedness from pressing social realities (ibid). In the medical literature, the dispute concerning drug trials between experts and activists has often been characterized as a clash in values. For instance, Feinstein (1983) distinguishes between the two conceptions of such trials as the "pragmatic" and "fastidious" perspectives. Activists took the pragmatic approach and thought that the trial design should "incorporate the heterogeneity, occasional or frequent ambiguity, and other 'messy' aspects of our ordinary clinical practice" (Feinstein 1983, p. 545). In their view, medical research should respond to the pressing public health issues at hand and given the urgency of the AIDS epidemic, they felt that there was no time to wait for "pure" subjects. On the other hand, experts initially took a more fastidious methodological approach as they feared that the pragmatic strategy would yield a "messy" answer (ibid). They instead preferred "using homogenous groups, reducing or eliminating ambiguity, and avoiding the spectre of biased results" (ibid).

Once activists established that their methodological critiques were credible, they were able to gain representation on NIH and FDA advisory committees, on institutional review boards at local hospitals and research centers, on community advisory boards established by pharmaceutical companies, and on a national board created by the Clinton administration, responsible for overseeing the entire course of AIDS research (Epstein, 1996). Eventually, AIDS activists became full partners in bringing effective antiretroviral drugs to the market in 1996. Through their efforts, they revolutionized how pharmaceutical sciences are practiced and today, patients of most diagnoses are involved in research through formal advisory boards.

There are many lessons to glean from the role activists played in AIDS research. By being receptive towards non-experts' methodological disagreements, researchers were able to increase the external validity of the AIDS clinical trials and bolster trust. As AIDS activists themselves pointed out, methodological disagreements between experts and non-experts stemmed from different value-laden assumptions and not from a lack of understanding by non-experts. For this reason, John may object that the example doesn't serve against his argument, since the methodological dispute wasn't rooted in non-experts' folk philosophy of science.

However, notice how experts couldn't have learned of these valid methodological critiques if transparent and honest communication wasn't fostered between themselves and non-experts. In other words, John's suggestion that scientists should consider the general publics' views about scientific methodology *before* deciding whether to be transparent or not could have backfired here. If researchers followed John's advice, it is plausible that they would have excluded dissenting activists from the conversation, fearing transparency may run contrary to non-experts' epistemic interests.

A final important note about the relationship between experts and non-experts for AIDS research is the following. Activists started as science novices in their initial interactions and involvement with experts. Yet, through the course of their conversations with experts, they developed detailed and sophisticated scientific knowledge. Overtime, activists found themselves comfortably conversing with researchers about "viral assays", "reserve transcription", "cytokine regulation" and "epitope mapping" (Epstein, 1996, p. 419). As a result, researchers felt even more compelled to consider activist arguments on their merits. This seems to imply that AIDS activists weren't actually non-experts, as they gained fluency in the relevant medical and epidemiological concepts. But, once again, notice how an atmosphere of transparency and honest communication is what helped activists develop the necessary knowledge to converse comfortably with experts.

The example of AIDS research reveals how transparency can strengthen trust in science, despite methodological disagreement between experts and non-experts. In the next section, I will consider a set of other important sources for skepticism in science that will help us better understand when transparency is helpful versus not.

2.6 Other factors that relate to transparency

The discussion of Climategate and vaccine skepticism is meant to show that the public's folk philosophy of science isn't the most salient factor to consider when determining the efficacy of transparency. Instead, the extent to which transparency initiatives are effective in fostering trust in science depends less on how careful scientists are when communicating their work, and more instead on whether special interest groups are acting out of bad faith. Thus, we should ask ourselves two related questions: 'has this scientific topic been politicized?', and 'are special interest groups playing a role in the dissemination of the scientific information?' when deciding whether transparency on the part of the scientists will be effective or not. I will consider each of these questions in turn.

First, we may ask ourselves 'has this scientific topic been politicized?'. Consider how 70% of Democrats have a positive view of environmental researchers while only 40% of Republicans do. Furthermore, 47% of Democrats trust environmental scientists, whereas, only 19% of Republicans do (Funk et al., 2019). This discrepancy between Democrats and

Republicans when it comes to trust in climate science is no accident—65% of Republicans trust FOX News, including its reporting on climate science (ibid). In contrast, 61% of Democrats *distrust* FOX News and this includes its reporting on climate science. Citing survey data, climate scientist and author of *Climatology versus Pseudoscience* Dana Nuccitelli writes:

> Republicans who watch Fox News are more than twice as likely to deny human-caused climate change than Republican non-viewers, and 62 percent of Republicans watch Fox News… [this data] suggests that the presence of Fox News and other conservative media outlets may be the primary explanation for why climate denial is more prevalent in the United States than in other developed countries (Bulletin of the Atomic Scientists, 2019).

A common strategy in the politicization of science is to have actors emphasize the inherent uncertainty of science to cast doubt on the existence of a scientific consensus. This results in citizens dismissing credible scientific information and undermines the positive role that science can play in informing political debates on issues with substantial scientific content (Dietz, 2013). This strategy has been shown to be extremely effective. Bolsen and Druckman (2015) tested the effects of politicization of scientific information by surveying 2,484 Americans about two relatively new energy technologies. The first survey asked participants about CNTs, which are tiny graphite tubes that convert sunlight into electricity and thus offer a novel method to obtain energy from an alternative source (N=1,256). Surveys suggested that nearly half of the US population knew virtually nothing about CNTs at the time. The other survey concerned hydraulic fracturing or fracking (N=1,228). Although more Americans knew about fracking than CNTs, survey data indicated that most Americans in 2014 were "largely unaware and undecided about this issue" (Boudet et al., 2014, p. 63).

Participants were then assigned to one of 6 different conditions, though I will discuss just the first 3 conditions. For condition 1, no information about either technology was given and participants were simply asked to report on the extent of support for the use of CNTs or fracking. For condition 1, only 17% of the participants supported fracking and 22% CNTs. For condition 2, they presented the technologies as if there was general scientific support for their use. This condition was meant to test the effect of scientific consensus *without* politicization. Here they found that 95% of participants supported fracking and 94% supported CNTs. Finally, the third condition added a statement that accentuated politicization such as

> yet, importantly, politics nearly always color scientific work, with advocates selectively using evidence. This leads many to say it is unclear whether to believe scientific evidence related to debates over CNTs [or fracking]. Some argue the process leads to pollution that harms the environment, while others disagree, pointing to evidence that there are minimal or no negative environmental consequences (p. 754).

With the politicization condition only 4% of participants supported fracking and 17% CNTs. What this seems to imply is that the general public is likely to trust the consensus views of scientists and the politicization of science plays an important causal role in generating skepticism.

This leads to the second question ‘are special interest groups playing a role in the dissemination of scientific information?’. As we saw with the examples of climate change and vaccine skepticism, special interest groups did play a role in manufacturing doubt. This third question is directly related to the politicization of science because often times special interest groups have political motives. For example, politically motivated think-tanks such as the American Enterprise Institute and the George Marshall Institute have been active in promoting a message that is at odds with the consensus view on climate change (Gelbspan 1997, 2004; Oreskes and Conway, 2010). These organizations have helped scientists who disagree with the mainstream view get media attention and airtime on right-leaning political news networks like FOX News (ibid).

It is important to note that special interest groups are often utilized by the private sector as well. When scientific claims contradict the financial motives of certain industries, special interest groups are hired to manufacture doubt. Returning to the example of climate change, the message of scientific uncertainty was reinforced by the public relations campaign of corporations that have a stake in the issue. For example, ExxonMobil has spent at least $16 million between 1998 and 2005 to fund a network of 40 think tanks and special interest groups to manufacture doubt about climate science (Union of Concerned Scientists, 2007).

We cannot determine what effect the general public's misconceptions about scientific methodology have on their appreciation of transparent scientific information until we are sure that scientific information is being disseminated in an objective, honest way. Thus, the philosophical debate concerning transparency shouldn't center on how non-experts process scientific information. When determining whether transparency will be effective in promoting trust in science, we should first answer the two more basic question presented in this section. Moreover, even if transparency might be used by propogandists in these nefarious ways, the right advice isn't to make science less transparent. Instead, we should be dealing with these malicious actors more directly.

2.7 Conclusion

I have shown that the existence of bad faith actors serves as a defeater for John's interpretation of the Climategate case study. I then presented examples that instead support the idea that how scientists communicate their work isn’t what matters most when it comes to public trust in science. More specifically, the vaccine example indicates that skepticism occurs even when scientists are working in ways that correspond to non-experts’ idealized methodological assumptions. Although this doesn’t serve as a knockdown argument against John’s general claim, it illustrates how there are other more influential factors for science skepticism besides non-experts’ methodological misconceptions. Alternatively, the example of AIDS research shows that contrary what John claims, transparency can actually strengthen trust in science when experts and non-experts disagree on methodology. In general, both examples suggest that focusing on one influential factor, such as non-experts’ methodological assumptions, will not help us derive a comprehensive account of the efficacy of transparency.

Transparency as a method of garnering trust in science can only work if the public is receiving their information from scientists themselves. However, most Americans do not receive their scientific knowledge directly from science journals and organizations. Instead, news media is the common source of scientific information and as discussed, many popular news outlets in the United States either lean conservative or liberal. This is problematic, since the politization of science undermines the impact of positive consensual scientific information. Additionally, special interest groups with political or corporate interests manufacture doubt by making it seem as though there isn’t scientific consensus in cases where there is.

All of these factors are relevant to the transparency debate. Thus, instead of myopically focusing on the general public’s methodological assumptions or scientists’ communication mishaps, we should expand our analysis to include the epistemic and political environment in which scientific information is communicated. This will hopefully allow us to develop more direct and comprehensive solutions to the issue of public mistrust in science.

CHAPTER 3

DIAGNOSING THE ACCURACY-FAIRNESS TRADEOFF IN MACHINE LEARNING ALGORITHMS

3.1 Introduction

Machine learning systems have been increasingly utilized in human decision making in both the public and private sector. For instance, machine learning algorithms are often used to produce predictions in various domains including job candidates' outcomes, susceptibility to loan payment, likelihood of recidivism, etc. Due to their ability to track and use massive amounts of personal data, machine learning algorithms have generated novel ethical concerns about fairness, privacy, and the control of information. More recently, scholars have noted that the outcomes associated with machine learning systems are often worse for racialized people, women, and other marginalized minorities.

Machine learning algorithms applied to such problems use training data to produce a function that takes inputs and produces predictions (Kleinberg et al., 2019). The way such algorithms generate prediction is via a training model that usually involves fitting a curve to a set of training data points for which classification labels are already known (Cooper and Abrams, 2021). After the training phase, the algorithm is fed new data points and provides classification labels for these new data points. To better understand this process, consider the following example. Suppose we are interested in predicting a job applicant's future performance in a new company. In building such a model, the training process uses data to produce classifiers that best optimize some objective function. When the algorithm receives a new job applicant's data after the training process, it can classify whether that applicant should be hired or not. As a result, accuracy measures how often a machine learning algorithm correctly predicts or infers a decision outcome after training (ibid.).

In response to the ubiquitous application of machine-learning algorithms in decision procedures that directly impact peoples' lives, many modelers have developed an interest in algorithmic fairness, and a common view in the machine learning literature is that there is an inherent trade-off between accuracy and fairness. For example, in the criminal justice context, the accuracy of a decision is defined by how it best maximizes public safety, whereas, the fairness constraints aim to reduce racial disparities in decision outcomes (Corbett-Davies et al., 2017). The pre-existing machine learning literature presents strategies for dealing with this inherent tension between accuracy and fairness by specifying the conditions under which the trade-off dissolves, arguing for why fairness should be scarified in favor of accuracy, or something in between (Dutta et al. 2020; Chen, Johansson, and Sontag 2018; Bakker et al. 2019; Menon and Williamson 2018). However, very few scholars have actually challenged the assumptions that casts fairness at odds with accuracy.

In this paper, I examine how the way accuracy and fairness are defined and operationalized results in the inherent trade-off between them. I refer to the current debate in the values in science literature concerning how epistemic and non-epistemic values should be distinguished in scientific reasoning to better understand why accuracy and fairness are considered to be at odds with one another. There is no reason to assume that fairness and accuracy *must* be in tension with one another. Instead, modelers choose to incorporate assumptions in their models that cast fairness in direct opposition to accuracy and what guides these choices are oftentimes non-epistemic values. More specifically, I focus on the way accuracy is defined in the literature and argue that although accuracy is meant to be a purely epistemic notion that is conducive to truth and rational belief, in this context, it incorporates non-epistemic considerations.

The paper will proceed as follows. The second section will discuss the current literature on the accuracy-fairness tradeoff in machine learning algorithms. The third section will discuss how claims concerning the inherent tradeoff between accuracy and fairness typically fail to acknowledge the fact that accuracy is a socially shaped concept. More specifically, I will argue that accuracy in machine learning algorithms is always understood with respect to some socially, ethically, or politically shaped goal or objective. Here I will go on to claim that the choice in goal or objective to be maximized in the algorithm are based on sociopolitical, or non-epistemic, values in the sense outlined in the philosophy of science literature. In the following section four, I will argue that non-epistemic values also play a role in how accuracy is operationalized in the algorithm. Finally, section five concludes that once we recognize the role non-epistemic values play in defining and operationalizing accuracy, the accuracy-fairness tradeoff must be understood differently.

3.2 The tradeoff between accuracy and fairness

The tradeoff between accuracy and fairness can be thought of as an optimization problem—where the optimal solution is the one that maximizes some objective function, or a function that measures how well the model performs on a particular objective (Cooper and Abrams, 2021). In this section, I will discuss how in the machine learning literature accuracy and fairness require two separate objective functions that cannot be maximized simultaneously, thus resulting in the inherent tradeoff.

To understand how accuracy is defined, it is important to consider how machine learning algorithms are trained. Suppose a machine learning algorithm is being used to produce some evaluative score. In order to be represented by the data, the outcome of the algorithm must be specified in terms of an objective function to be maximized. The training phase involves the use of data in which the various input factors and objective function are known. The algorithm then picks up the numerous correlations that exist between the input factors and the objective function. The algorithm is then given input factors only and a prediction is made. The results of the prediction are then checked against the actual value of the objective function. Thus, accuracy measures how often the machine learning model correctly predicts or infers decision outcomes after training.

There are two more specific ways of defining accuracy within a machine learning algorithm. One way accuracy is often measured is through label alignment, or the percentage of correctly classified data points, where correctness is determined by whether the model's classification decision matches the known label (Cooper and Abrams, 2). In these types of machine learning algorithms, the algorithm makes a correct classification in relation to either an explicit or implicit set of classifying rules. For example, consider an algorithm that is meant to classify dogs versus cats and is trained on labeled images of dogs and cats. The algorithm may learn that "long snout" is a reliable rule in distinguishing cat and dog images. However, notice that things can go wrong when classifying subgroups of dogs that deviate from the majority. For example, this algorithm may mislabel Pomeranians as cats, since Pomeranians don't have long snouts like most other dogs (Cooper and Abrams, 3).

Another way of defining accuracy is relative to the algorithm's positive predictive value. In this case, the machine learning algorithm is concerned with making accurate predictions instead of simply accurate classifications. For example, consider a graduate school admissions algorithm where the aim is to maximize students' success in graduate school. Since 'success in graduate school' is an abstract concept that is difficult to measure, the modeler may choose to operationalize success by means of measurable outcomes like the number of publications or awards and recognitions achieved during graduate school. After deciding how to operationalize the objective or goal 'successful graduate student', the modeler then decides what data the model should be trained on. This data will include input factors that will serve as predictors for 'graduate student success' measured via publications and/or awards and recognitions. Plausible predictively reliable input factors for future publications and awards are place of undergraduate education, undergraduate GPA, letters of recommendation, etc. In the subsequent testing phase, the algorithm is given input information only, and a prediction is made for each person in relation to the objective function. The outputs are then compared with the actual data. So, for example, the algorithm may predict that a current graduate student that had a high bachelor's degree GPA from a prestigious institution will likely publish at least 1-2 papers during their time as a graduate student. The modeler would then check whether this publication record prediction is actually true of that graduate student—and if it is, the algorithm has made an accurate prediction.

The thing to notice here is that an algorithm's predictions are deemed accurate or inaccurate relative to a certain objective chosen by the modeler. Thus, modelers must make critical decisions when picking an objective to maximize and deciding how to operationalize this objective in the algorithm. In the prior example, 'success in graduate school' is the objective the admissions board wants to maximize and operationalizes this objective by means of something quantifiable 'number of publications, awards, and recognition'. However, notice, that the admissions board could have chosen to operationalize 'success in graduate school' differently, or they could have chosen a different objective to maximize altogether.

Algorithmic fairness has been an important topic of discussion in the machine-learning literature and as a result, there are numerous mathematical definitions of it. The definitions of algorithmic fairness similarly include some treatment of protected attributes, such as race, gender, sexual orientation, etc., and decision outcomes that can be evaluated for fair treatment (Huq, 2019). These various mathematical definitions of fairness can be grouped into the following categories: group fairness, fairness through blindness and individual fairness. Group fairness is fairness by “comparing the target variable outcome of a machine-learning process between two groups sorted along the sensitive variable” (Bent, 2019). For example, balancing the number of false positives or false negatives between two groups are methods of ensuring group fairness. Fairness through blindness is instead a fairness strategy in which information that encodes protected attributes like race, gender, etc., are removed (Bent, 2019). Finally, individual fairness looks for “disparities in treatment at the individual level for individuals with similar features” (Bent, 2019).

This leads us to the discussion of why accuracy and fairness are considered in inherent tension with one another. If fairness as a constraint limits the set of possible classification assignments to those that meet a fairness requirement, then fairness will prevent the algorithm from simply maximizing accuracy based on all the features that would otherwise be available to the algorithm (Bent, 2020). Thus, at first glance, the tradeoff is quite intuitive—if fairness constraints are just limiting the set of possible classification assignments to those that are collectively fair, then of course this will reduce accuracy, as “optimization over a subset of the data is a lower bound compared to optimization over the original set” (Wick, Panda, and Tristan, 2019). For example, suppose a car insurance company uses clients’ zip code as a variable that helps determine risk scores. Further suppose that zip code is a reliable indicator of how likely a person will get into future accidents, since those that live in areas that are densely populated are much more likely to get into a wreck than those that live in suburban areas. In this example, zip code also serves a proxy for race—because of housing discrimination and systemic injustice, Black people tend to live in the inner city and whites tend to live in the suburbs. So, although zip code is a reliable predictor for risk of future accidents, the fact that it serves as a proxy for race makes the algorithm unfair. However, removing the 'zip code' variable from the algorithm for reasons of fairness will diminish its predictive accuracy.

Accordingly, the general view in the machine learning literature is that accuracy and fairness tradeoff on one another such that accuracy and fairness are modeled as two objective functions that cannot be simultaneously optimized (Chen, Johansson, and Sontag 2018; Menon and Williamson 2018; Bakker et al. 2019). However, the purpose of this paper is to present a different approach to addressing bias and discrimination that doesn't entail a tradeoff between the algorithm's accuracy and fairness. Those that claim that accuracy is inherently at odds with fairness fail to realize the fact that accuracy is, itself, socially shaped in two ways that impact how we should think about the tradeoff. First, accuracy is always understood with respect to some socio-politically influenced goal. In other words, there is no free standing, epistemically pure "accuracy". Second, in trying to reach that socially-chosen goal, accuracy must be operationalized in some specific way. Once we acknowledge these facts, the accuracy-fairness tradeoff must be understood differently since algorithmic predictions are oftentimes only 'accurate' with respect to standards that track structural inequalities. Consequently, we need to think more clearly about the way accuracy is defined within the model and whether the definition results in accurate predictions for all groups, including marginalized minorities. This will require us to think more carefully about the role sociopolitical values play in defining accuracy in the first place.

3.3 The role non-epistemic values play in defining accuracy

The push to utilize machine-learning algorithms in decision-making in both the public and private sector has been based on the assumption that automated classifications and predictions are objective and unbiased. In general, the common assumption is that if bias get into the algorithm at all, it does so by being already present in the data on which the algorithm is trained, and not by means of the explicit design decisions of the modeler (Johnson, 1). The purpose of this section is to push back on this general assumption.

More specifically, the discussion will draw upon the fact that 'accuracy' in the machine-learning context is not a purely epistemic norm, as defining what it means for a prediction to be accurate relies on a host of socio-political assumptions. When building and applying machine-learning algorithms, modelers must choose some objective or goal to maximize. For example, the objective of a recidivism model can be 'public safety'. Similarly, in the context of mortgage loans the objective to be maximized might be 'timely loan repayment'. Finally, for those using algorithms for college acceptance decisions, the objective to be maximized may be 'academic preparedness'. The point here is simply that machine-learning modelers must choose something to maximize in each case and that sociopolitical values can influence this decision.

To understand this point, let me make clear exactly what I mean by purely epistemic versus non-epistemic or sociopolitical values. Creating machine-learning algorithms requires a whole host of assumptions. Oftentimes these assumptions are guided by certain agreed upon epistemic values. For example, in machine learning, when writing proofs about an algorithm's proprieties, it is common to assume that the distribution is convex (Cooper and Abrams, 2021).

Assumptions like this enable modelers to guarantee certain conclusions about an algorithm's behavior, such as bounds on its convergence rate (ibid). Other examples more relevant to the topic at hand are the assumptions that statistical parity can model fairness or that label alignment is a proper accuracy metric. The point here is that these assumptions are similarly justified by their mathematical simplicity. Notice how simplicity is an epistemic value that modelers share along with "accuracy, consistency, scope, fruitfulness" to name a few others (Kuhn, 1977, p. 322). Therefore, epistemic values inform decisions relating to the production of true accounts, predictions, and theories in science, whereas, non-epistemic values aren't assumed to be entirely truth-tracking, as they involve social, moral, and political considerations (Rooney, 32).

While machine-learning researchers are accustomed to explicitly stating how epistemic values guide their assumptions, we have yet to see similar attention paid to the way non-epistemic values guide such assumptions. But consider how mathematical models require both the use of epistemic and non-epistemic values due to practical limitations. For instance, in creating climate models, modelers are faced with uncertainty that stems from an incomplete theoretical understanding of the climate system and from constraints placed by computing power (Winsberg 2010; 2012). Given this uncertainty, modelers must make critical choices about how to model a certain phenomenon and oftentimes, values guide these choices. What is important to note in relation to Winsberg's discussion of the roles values play in model and algorithm building is a distinction he makes between "epistemically forced" and "unforced" choices (2012, p. 124). An epistemically forced choice is one in which there are purely epistemic grounds for considering one model-building option over another. Otherwise, the choice is epistemically unforced, and thus requires social and political values in the decision process.

More recently, Gabriel Johnson has argued for that distinguishing between the role epistemic and non-epistemic values play in generating and applying machine-learning algorithms is often untenable (2020). Drawing insight from Helen Longino (1995), Johnson argues that all values, including epistemic ones, are formed with regards to a certain socio-political context. As a result, even epistemic values necessarily reflect the sociopolitical features of the environment to which they are applied. To make this point clear, Johnson uses the example of clinical drug trial for the common sleep-aid Ambien. The prescription drug Ambien was approved as a sleep aid by the FDA in 1992. Due to concerns related to simplicity, pharmacologists took the male metabolic system to be the paradigm case, and generalized their findings based on the male metabolic system to women. This resulted in similar recommended doses between men and women. However, this had dire effects for women taking the drug, as it was found later that women were taking nearly twice the amount they should have been, based on their body mass and metabolic rates. In this case, scientists' commitment to simplicity required that they posit the fewest kinds of entities in this context, resulting in the male body to be taken as the essential model of human physiology. But notice how the epistemic value of simplicity is infected by the sociopolitical context in which it is formed—the simple model of human physiology is one based off the male body, as men have more power and privilege in our sociopolitical context. Thus, adherence to simplicity will take in the very sociopolitical values on which the power dynamics in society are formed (Johnson, 2020).

Johnson's discussion is helpful for understanding the accuracy-fairness tradeoff in machine learning and importantly relates to the point about accuracy addressed in this paper. Since accuracy is always defined with relation to some socially-chosen goal, the epistemic value of accuracy always involves non-epistemic considerations. For example, consider how a college admissions algorithm can maximize one of the following two objectives—'preparedness for school' or 'innate promise' and how the latter objective can result in fairer outcomes for marginalized minorities. An objective function such as academic preparedness can be operationalized by means of any characteristic that correlates with a student's preexisting skills and knowledge. So, for instance, the measure 'academic ranking of high school attended' can serve as a measurable stand in for 'academic preparedness'. Notice how if the college board actually focused on this objective function and its operationalization, the outcomes would be worse for minority students, as white students attend high-ranking schools at higher rates than marginalized minorities. In this example, although the algorithm is unfair since it downgrades racialized minority applicants, it is still accurate—it is true that 'academic preparedness' can be reliably maximized by means of the measure 'academic ranking of high school attended'.

This example reflects how our standards for accuracy are often shaped by the social environment that reflects pervasive, systemic inequalities for unprivileged groups. This also demonstrates some of the challenges of using trade-off and optimization tools in algorithmic fairness research. For instance, if the accuracy metric is condition on past unfairness, what is the trade-off between fairness and accuracy actually measuring? Notice how if accuracy metrics encode past unfairness for marginalized groups, the fairness-accuracy tradeoff is effectively positioning fairness in a tradeoff with a form of unfairness.

Thus, there are two lessons to be drawn from this example. First, modelers discussing the accuracy-fairness tradeoff fail to realize that their choice to privilege accuracy over fairness, and the various metrics they use to define accuracy, are themselves value-laden judgements. Second, the example shows that sometimes making an algorithm fairer requires we change our objective or goal to be maximized altogether. Instead of maximizing 'academic preparedness' in college admissions, we could instead try to maximize something like ‘innate promise’ that can be operationalized by means of something like ‘high school GPA’. This objective would be result in fairer outcomes for marginalized minorities and so arguably should be our goal in implementing college admissions algorithms. However, we tend to see college admissions focus on the former objective instead and then later account for the unfair outcomes by means of some fairness metric—like balancing false positives between white and Black students, or false negative, or some other parity notion of fairness that is in tension with the algorithm’s accuracy. What I am suggesting here is that considering fairness as a different objective to be optimized *after* the predictively accurate algorithm has been built is the wrong approach. Modelers should be thinking of fairness at the start, like when deciding on a particular objective the algorithm does well in relation to.

Thinking more carefully about what groups the algorithm is making accurate predictions for versus not, and for what reasons, helps us realize that we can make an algorithm fairer and more accurate at the same time. In principle, we can choose an objective to maximize that correlates with fairness, such that, the accuracy-fairness tradeoff is eliminated. This shows that the tradeoff isn’t inherent, unlike what is claimed in the machine-learning literature, and instead arises from the ways we choose what we want from our algorithms.

But, of course, sometimes we want our machine learning algorithms to categorize people or make predictions according to goals that are notably distinct from fairness. In such cases, is the tradeoff between accuracy and fairness inevitable? Unless the objective for the algorithm is explicitly biased, the answer to this question is ‘no’. This is because, in cases where we want to make predictions or categorizations according to goals besides fairness, the tradeoff can be greatly mitigated through how we operationalize accuracy. However, because we live in a biased system, the ways we operationalized accuracy tends to reflect systemic injustices. In the next section, I will argue that the way we operationalize an algorithm's objective affects its fairness. In order to support this argument, I will consider recidivism and bail setting algorithms.

3.4 The role non-epistemic values play in operationalizing accuracy

One particular case that is popular in the discussion of the accuracy-fairness tradeoff concerns criminal justice risk-assessment algorithms such as COMPAS or the Correctional Offender Management Profiling for Alternative Sanctions. COMPAS, developed and owned by Northpointe, is a program used by judges across the United States to produce recidivism risk scores. COMPAS works by collecting data about defendants awaiting trial and produces a risk score for the defendant based on statistical analysis. This risk score is then used by judges to make decisions about setting bail, establishing the need for pretrial detention, sentencing, or parole, among other things. Despite its promise of objectivity, in 2016 ProPublica revealed that in an analysis of over 7,000 COMPAS uses, the program was twice as likely to falsely label Black defendants as future criminals than white defendants (O'Neil 2016; Johnson 2020). Along these lines, white defendants were mislabeled as low risk at a much higher rate than Black defendants (ibid).

In light of ProPublica's analysis that exposed bias in the COMPAS algorithm, Northpoint published a validation study in 2009 where it was found that the risk of recidivism score had an accuracy rate of 68% for a sample of 2,238 people (ProPublica, 2016). The study also showed that the score was slightly less predictive for Black men than white men—67% versus 69%. Tim Brennan, the founder of Northpoint, argued that these findings indicate the reliability of the COMPAS algorithm. When asked about the differences in accuracy rates between white and Black inmates, Brennan highlighted how it is difficult to construct a score that doesn't include items that can be correlated with race—like poverty, joblessness, and social marginalization (ProPublica, 2016). According to Brennan, “if those [predictors] are omitted from your risk assessment, accuracy goes down” (ProPublica, 2016). In other words, the algorithm’s (relatively) high accuracy rate of 68% is partly due to the fact that it uses variables that correlate with race. As a result, making the algorithm fairer by removing these variables will decrease the algorithm’s accuracy.

ProPublica’s analysis of the COMPAS model sparked a heated and captivating debate in the ethics of algorithms literature. Due to the discrepancy between recidivism predictions between Black and white inmates, many scholars proposed ways of making recidivism algorithms fairer through some form of statistical parity. These methods similarly aimed to equalize metrics between individuals or groups by, for instance, requiring equal rates of accurate and inaccurate predictions between one group and another (Hellman 2020, Bent 2019, Johnson 2020). The underlying assumption in this literature is that, given the COMPAS algorithm’s high predictive accuracy of 68%, perhaps the best way to mitigate issues of injustice is to apply a technical formalization of fairness to the COMPAS algorithm via statistical parity (ibid).

However, when determining how to make the COMPAS model fairer, we must also consider the fact that an algorithm’s predictive accuracy is always measured relative to some objective to be maximized. The COMPAS algorithm attempts to ‘maximize public safety’ by accurately predicting the chances of recidivism. Since ‘chances of recidivism’ is an abstract concept that is difficult to measure, Northpoint chose to operationalize recidivism by means of the measurable outcome “a new arrest within two years” or more specifically, “a new misdemeanor or felony offense within two years of the COMPAS administration date” (Brennan et al., 2009).

Thus, the COMPAS algorithm's unfair and discriminatory outcomes depends on the way recidivism is operationalized in the model. If reoffending is defined as arrests within two years after release, it is evident how the algorithm may make accurate predictions at the expense of fairness, as racialized minorities have a higher chance of being arrested than their white counterparts, not because they are more criminal or dangerous, but instead because their communities are overpoliced. In fact, a 2013 study by the New York Civil Liberties Union found that while Black and Latinx males between the ages of fourteen and twenty-four made up only 4.7 percent of the city's population, they accounted for 40.6 percent of the stop-and-frisk checks by police (O'Neil, 2016, 58). Thus, operationalizing a defendant's chances of recidivism as 'new arrests within two years' will systematically label racialized minorities as a higher risk than their white counterparts, simply because they are more prone to have run-ins with police due to racialized policing practices.

If the COMPAS algorithm's predictions were instead accurate relative to simply chances of committing a crime in the future, regardless of whether the defendant gets arrested, racialized minorities subject to over-policing would not be deemed riskier than their white counterparts. To support this claim, consider how although Black and white Americans use cannabis at relatively similar rates (10.7% for Black Americans and 8.4% for white Americas) and that whites make up a much larger proportion of the American population, Black Americans are still four times more likely than whites to be arrested for marijuana possession (National Survey on Drug Use and Health, 2016; Washington Post, 2020). What this example shows is that the chances of committing a future crime via marijuana possession is similar across white and Black defendants since the marijuana consumption rates between white and Black Americans is comparable. However, when it comes to the chances of committing a future crime and getting caught or arrested, the chances are far higher for Black inmates. Since the accuracy of the COMPAS algorithm is predicated on the propensity to be arrested instead of simply the propensity to commit crimes in the future, Black inmates will be deemed riskier to no fault of their own.

Let me make this same point in relation to a different example. Recall, Northpoint chose to operationalize recidivism by means of the measurable outcome “a new arrest within two years” and notice how broad this definition of recidivism is (Brennan et al., 2009). A full range of crimes are taken into account, including small infractions and misdemeanors like driving with an expired license. As mentioned earlier, relative to this way of operationalizing recidivism, the COMPAS algorithm has a predictive accuracy of 68%. However, if we focused on just predicting violent crime, such that, recidivism in the model instead means “a new violent arrest within two years”, notice how the COMPAS algorithm’s predictive accuracy substantially diminishes. As ProPublica’s findings demonstrated, only 20% of the people predicted to commit a violent crime actually went on to do so. What this shows is that the COMPAS algorithm’s predictive accuracy when it comes to violent crimes is actually quite low.

A plausible reason for this discrepancy is the following. Predictors that proxy for race like poverty, joblessness, and social marginalization contribute to the accuracy of the COMPAS model if we what we are interested in is “arrests within two years” as these factors may result in a former inmate to “drive with an expired license”, shoplift, or some other petty crime (ProPublica, 2016). However, if we instead focused on “a new violent arrest within two years” it is unclear whether and to what extent poverty, joblessness and social marginalization would predict reoccurring violent behavior. What we do know is that as of now, for the COMPAS algorithm, these metrics are quite unreliable in predicting violent re-offenses.

Thus, if Northpoint instead sought to maximize public safety by accurately predicting chances of recommitting violent offences within two years of release, perhaps the algorithm would concurrently be made fairer and more accurate. Recall, that the variables in the COMPAS algorithm are quite poor at predicting violent crime. Furthermore, many of the Black inmates with no record of past violent crimes nevertheless received a high-risk score, whereas, white inmates with a prior history of violence received lower risk scores (ProPublica, 2016). Thus, the predictors used in the COMPAS model that often proxy for race helped with generating reliable predictions concerning future arrests broadly construed but are nevertheless very unreliable in predicting future arrests based on violent crimes.

Moreover, it is important to note that operationalizing ‘chances of recidivism’ in this way has dire long-term effects on racialized minorities. According to the COMPAS algorithm, a person who scores ‘high risk’ is likely to come from a neighborhood where run-ins with police is a common occurrence (O’Neil, 2016). Due to their high-risk score, the defendant gets a longer sentence—which makes later rehabilitation more difficult. When they are finally released into the same overpoliced neighborhood, they have a criminal record which makes finding a job more difficult, thus making it more likely that they will commit another crime. If they do in fact commit another crime, the COMPAS algorithm can claim that it had made an accurate prediction, when in fact the result was created by the effects of using the algorithm that operationalizes recidivism as new arrest within two years. As a result, if modelers aren’t careful when deciding how to operationalize a certain objective, the predictive accuracy of an algorithm can be predicated on a pernicious feedback loop like the one outlined here.

The purpose of this paper is to present a different approach to addressing bias and discrimination in the use of the COMPAS algorithm that doesn’t entail a tradeoff between the algorithm’s accuracy and fairness. As previously mentioned, accuracy measures how well an algorithm performs on a particular objective that is selected by the modeler. Often there is more than one objective to satisfy simultaneously in the algorithm and those objectives can be in tension with one another. In such instances, it is common to pose this problem as optimizing a trade-off (Cooper and Abrams, 2021). In the context of criminal justice, bail and recidivism, the accuracy of decisions has been framed as how best to “maximize public safety” while still satisfying some formal fairness constraints that aim to reduce racial disparities in decision outcomes (Huq et al., 2017). My suggestion has been the following: recidivism algorithms can maximize public safety while remaining fair by focusing on an objective to maximize that doesn’t tradeoff with fairness. As argued, sometimes making an algorithm fairer requires that we change our objective function altogether. In other cases, like the one at hand, the objective to be maximized need not be changed and instead, the way we operationalize this objective in the model must instead be changed for fairer outcomes.

On this picture, we should worry less about the inherence of the accuracy-fairness tradeoff, and more about what goes into defining and operationalizing ‘accuracy’ within the machine-learning algorithm in the first place. More specifically, we should ask ourselves: ‘how do we pick objectives for our algorithms, and operationalize these objectives, such that the trade-off is less severe and worrisome?’ By thinking more carefully about who the algorithm is making accurate predictions for versus not, and for what reasons, helps us realize that we can make an algorithm fairer and more accurate at the same time. This will require modelers to be rigorous and clear about the role non-epistemic values play in guiding how they define and operationalize accuracy in the algorithm.

I am not suggesting that making non-epistemic values explicit is a sufficient solution to the problem of inequity and bias in algorithmic decision-making. Nevertheless, it would still help facilitate greater scrutiny about the appropriateness of proposed algorithmic fairness solutions. For instance, my analysis would allow for considering that fairness and accuracy could in fact be in accord. Moreover, the argument presented here requires machine-learning researchers to be more introspective about how their particular sociopolitical values might inform their modeling choices and assumptions. Clarifying these implicit non-epistemic values and their role in building machine-learning algorithms facilitates rethinking how we measure and understand accuracy and at the very least, presents a novel solution to the problem of algorithmic bias that is worth exploring.

3.5 Conclusion

To conclude, I would like to reiterate how machine learning algorithms require a host of both epistemic and non-epistemic assumptions. Epistemic assumptions come naturally to modelers like, for example, modeling some real-world problem as an optimization problem where the best solution is one that either minimizes some cost function or maximizes some objective function. What modelers often fail to realize is that these epistemic modeling assumptions also involve non-epistemic norms. The arguments presented in this paper support the conclusion that modelers should take the time to make explicit the non-epistemic assumptions that underly their work, as being rigorous and clear about non-epistemic values allows them to be reviewed and vetted with the same rigor as the purely mathematical assumptions. This is especially important when we consider how difficult, and oftentimes untenable it is to separate the modeling assumptions that are purely epistemic from those that non-epistemic.

Furthermore, as the arguments suggest, clarifying the implicit role non-epistemic values play in the mathematical assumptions facilitates rethinking how we can measure accuracy. To the best of my knowledge, no preexisting algorithmic fairness scholarship has considered making algorithms fairer by changing the way accuracy is defined and operationalized in the model. This is likely because operationalizing fairness and accuracy as epistemic and mathematical metrics helps situate algorithms as purely value-free tools. But once we examine more carefully what objective function an algorithm is accurate with respect to, we notice that there is a social component involved in the measurement of predictive accuracy. As a result, we can maximize accuracy and fairness at the same time by defining accurate prediction relative to an objective function, or goal, that actually tracks the phenomenon in question for all groups of people, including those that are marginalized.

CONCLUDING REMARKS

This dissertation project was motivated by the debate over the ideal of value-free science. The worry scientists have long had is that non-epistemic values undermine the objectivity of science by contaminating the search for truth with social, political and ethical priorities and motives. As displayed in this dissertation, there are many ways in which science is responsible to society, as the fruits of science are often used in value-laden setting. When one considers this more seriously, a clear separation between science and social concerns becomes untenable.

In light of this fact, this dissertation connects the philosophical literature with existing scientific techniques in the hope of illuminating the ethical obligations scientists have to society at large. How should the social implications of a study effect our epistemic standards? What is the correct basis for trust in science by the broader society and what role should values play in science given this basis? Finally, what explicit role are social and political values playing in the use of mathematical models?

This final question is extremely important if we consider how mathematical models and more specifically, algorithms associated with AI, big data, and machine learning play a central role in a wide range of public and private practices. This dissertation contributes to debates over equity in modeling and algorithmic contexts, with particular attention to the ways mathematical models can perpetuating social inequalities. Ultimately, this dissertation is suggestive of ways that understanding and evaluating the use of mathematical models requires new norms and new theoretical tools. My hope is that the arguments presented in this dissertation will generate a more nuanced and refined understanding of the way mathematical models are undeniably value-laden.

BIBLIOGRAPHY

Ajunwa, Ifeoma (2019). "The Paradox of Automation as Anti-Bias Intervention." *Cardozo Law Review* 41: 1671.

Bakker, Michiel A, et al. (2019). “On Fairness in Budget-Constrained Decision Making.” *KDD Workshop of Explainable Artificial Intelligence*

Balko, Radley (2020) “There’s overwhelming evidence that the criminal justice system is racist. Here’s the proof.” *Washington Post*

Bent, Jason R (2019). "Is Algorithmic Affirmative Action Legal." *Georgetown Law Journal* 108: 803.

Bolsen, Toby, and Druckman, James N, (2015). “Counteracting the Politicization of Science” *Journal of Communication.* 745-769.

Boudet, H, Clarke, C, Bugden, D, Maibach, M, Roser-Renouf, C, & Leiserowitz, A (2014). ““Fracking” controversy and communication: Using national survey data to understand public perceptions of hydraulic fracturing.” *Energy Policy.* 65, 57–67. doi:10.1016/ j.enpol.2013.10.017.

Bright, Liam K, (2017). Decision Theoretic Model of the Productivity Gap. Erkenntnis, 82(2):421–442.

Calcott, Brett (2014). The Creation and Reuse of Information in Gene Regulatory Networks. *Philosophy of Science* 81:1-12.

Chen, Irene Y, Johansson Federik D, Sontag, David (2018). “Why is My Classifier Discriminatory?” *Proceedings of the 32nd International Conference on Neural Information Processing Systems* :3543-3554.

Cooper, A. Feder and Abrams, Ellen (2021). “Emergent Unfairness in Algorithmic Fairness-Accuracy Trade-Off Research.” *AIES*

Cutts, Felicity T, Henderson, Roberts H, Clements, John C, et al. (1991). “Principles of measles control, Bull.” *World Health Organization.* 69(1): 1-7.

Dietz, T (2013). “Bringing values and deliberation to science communication.” *Proceedings of the National Academy of Sciences of the United States of America.* 110(3), 14081–14087. doi:10.1073/pnas.1212740110.

Douglas, Heather (2000). Inductive risk and values in science. *Philosophy of Science,* 559-579.

Douglas, Heather (2009). *Science, policy, and the value-free ideal.* Pittsburg PA: University of Pittsburg Press.

Douglas, Heather (2018). Values in Science. *The Oxford Handbook in the Philosophy of Science. Forthcoming.*

Dutta, Sanghamitra, Wei Dennis, Yueksel Hazar, Chen Pin-Yu, Liu Sijia, and Varshney Kush, (2020). “Is there a Trade-Off Between Fairness and Accuracy? A Perspective Using Mismatched Hypothesis Testing.” *Proceedings of the 37th International Conference on Machine Learning*:119.

Elliott, Kevin (2017). *A Tapestry of Values: An Introduction to Values in Science.* New York: Oxford University Press.

Finkelstein, Martin, Conley Valerie M, Schuster Jack H (2016). Taking the measure of faculty diversity. *Advanced Higher Education.*

Fisher v. University of Texas, Austin (2016). Brief for Lt. Gen. Julius W. Becton, Jr., et al., as Amici Curiae in Support of respondents. (no. 14-981).

Funk, Cary, Hefferson, Meg, Kennedy, Brian, et al. (2019). “Trust and Mistrust in Americans’ Views of Scientific Experts.” *Pew Research Center.*

Gelbspan, Ross (1997). The heat is on: The high stakes battle over earth’s threatened climate. Reading, MA: Addison-Wesley.

Gelbspan, Ross (2004). Boiling point: How politicians, big oil and coal, journalists, and activists are fueling the climate crisis—and what we can do to avert disaster. New York: Basic Books.

Godfrey-Smith, Peter (2014). “Sender-Receiver Systems within and between Organisms.” *Philosophy of Science* 81:866-78.

Gramlich, John (2020). “5 facts about Fox News.” *Pew Research Center.*

Harms, William (2004). *Information & Meaning in Evolutionary Processes.* Cambridge: Cambridge University Press.

Hellman, Deborah (2020). "Measuring Algorithmic Fairness." *Virginia Law Review* 106: 811.

Hempel, Carl G (1965). Science and Human Values. *Aspects of Scientific Explanation.* 81-96. New York, NY: The Free Press.

Hicks, Daniel J (2017). “Scientific Controversies as Proxy Politics.” *Issues in Science and Technology.* Vol: XXXIII.

Hicks, Daniel J (2018). “Inductive Risk and Regulatory Toxicology: A Comment on de Melo-Martín and Intemann.” *Philosophy of Science* 85 (1): 164–74.

Hilleman, M.R., Weibel, R.E., Buynak, E.B., et al. (1967). “Live, Attenuated Mumps Virus Vaccine Protective Efficacy as Measured Field Evaluation.” *New England Journal of Medicine*. 276: 252-258.

Hong, Lu and Page, Scott E, (1988). Diversity and Optimality. *Research in Economics*. 98:8-77.

Hong, Lu and Page, Scottt E, (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America*, 101(46):16385–16389.

Huttegger, S.M. (2007). Evolution and the explanation of meaning. *Philosophy of Science*, 74, 1–27.

Huttegger, Simon, & Zollman, Kevin J S, (2011). Language, games, and evolution. In Signaling games: Dynamics of evolution and learning, (pp. 160–176). Berlin Heidelberg: Springer.

Huq, Aziz Z (2019). "Racial Equity in Algorithmic Criminal Justice." *Duke Law Journal* 68, no. 6: 1043-1134.

Intemann, Kristen and Inmaculada de Melo-Martin (2016). *The Fight Against Doubt: How to bridge the gap between scientist and the public.* Oxford University Press.

John, Stephen (2018). "Epistemic trust and the ethics of science communication: Against transparency, openness, sincerity, and honesty," *Social Epistemology* 32: 72-87.

Johnson, Gabbrielle M, (2020). Algorithmic Bias: on the Implicit Biases of Social Technology. *Synthese* https://doi-org.proxy.lib.uwaterloo.ca/10.1007/s11229-020-02696-y

Jhun, Jen, Patricia Palacios, and James O. Weatherall (2017). "Market Crashes as Critical Phenomena? Explanation, Idealization, and Universality in Econophysics." *Synthese*.

Kitcher, Phillip (2011). *Science in a Democratic Society*. London: Prometheus books.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R. Sunstein (2018). "Discrimination in the Age of Algorithms." *Journal of Legal Analysis* 10.

Kovaka, Karen (2019). "Climate change denial and beliefs about science," *Synthese.* Available at: https://doi.org/10.1007/s11229-019-02210-z.

Kuhn, Thomas (1977). "Objectivity, Value, and Theory Choice." *Thomas Kuhn The Essential Tension: Selected Studies in Scientific Tradition and Change.* Chicago: The University of Chicago Press, 320-339.

Lakatos, Imre (1978). *The Methodology of Scientific Research Programmes: Philosophical Papers. Vol. 1*. Cambridge: Cambridge University Press.

Landemore, Helene (2012). Deliberation, cognitive diversity, and democratic inclusiveness: an epistemic argument for the random selection of representatives. *Synthese.* 190:1209.

Lewis, David K (1969). *Convention.* Cambridge, MA: Harvard University Press

Longino, Helen E (1990). *Science as social knowledge: Values and objectivity in scientific inquiry.* Princeton, NJ: Princeton University Press.

Longino, Helen E (1995). “Gender, politics, and the theoretical virtues.” *Synthese* 104(3), 383-397.

Longino, Helen E (1996). “Cognitive and non-cognitive values in science: Rethinking the dichotomy.” *In Nelson, LH. & Nelson, H. Feminism, science, and the philosophy of science* (pp. 39-58). Dordrecht: Kluwer.

Martini, Carlo (2014). Experts in Science: A View From the Trenches. Synthese, 191(1):3–15.

McAllister, James W (2012). “Climate Science Controversies and the Demand for Access to Empirical Data.” *Philosophy of Science* 79 (5): 871–880.

Menon, Aditya Krishna and Williamson, Robert C (2018). “The Cost of Fairness in Binary Classification.” *Proceedings of the 1st Conference on Fairness, Accountability and Transparency: Proceedings of Machine Learning Research*:81.

Mitchell, Sandra D (2004). “The Prescribed and Proscribed Values in Science Policy.” *Science, Values, and Objectivity.* Pittsburgh, PA: University of Pittsburgh Press, 245–55.

Nisbett, Richard and Ross, Lee (1980). *Human Inference: Strategies and Shortcomings of Social Judgment.* Prentice–Hall, Englewood Cliffs, NJ.

Noble, Safiya Umoja (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

Northpointe, I. (2015). *Practitioner’s Guide to COMPAS Core*.

Nuccitelli, Dana (2019). “Fox News made the US a hotbed of climate denial. Kids are the cure.” *Bulletin of the Atomic Scientists.*

Nyhan, B, Reifler, J, Richey, S, and Freed, GL, (2014). “Effective messages in vaccine promotion: a randomized trial.” *Pediatrics.* Apr; 133(4):e835-42.

O’Connor, Cailin (2014). “Evolving Perceptual Categories.” *Philosophy of Science* 81:840-51.

O’Connor, Cailin (2015). Ambiguity Is Kinda Good Sometimes. *Philosophy of Science.* 82:101-121.

O’Connor, Cailin (2017). Modeling Minimal Conditions for Inequity. *Forthcoming.* 1231.

O’Connor, Cailin, and Weatherall, James (2019). *The Misinformation Age: How False Beliefs Spread.* Yale University Press: New Haven.

O'Neil, Cathy (2016). *Weapons of Math destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Oreskes, Naomi and Conway, Erik M (2010). *Merchants of Doubt.* New York: Bloomsbury.

Page, Scott (2007). Diversity trumps ability theorem. *In The difference: How the power of diversity creates better groups, firms, schools, and societies* (pp. 131–174). Princeton, NJ: Princeton University Press.

Papineau, David (2012). “Correlations and Causes.” *Philosophical Devices*. Oxford: OUP.

Parker, Wendy S and Winsberg, Eric (2018). Values and evidence: how models make a difference. European Journal for Philosophy of Science, 8(1):125–142.

Pearce, Fred (2010). “How the ‘climategate’ scandal is bogus and based on climate skeptics’ lies.” *The Guardian*. https://www.theguardian.com/environment/2010/feb/01/climate-emails-sceptics

Pessach, Dana, and Erez Shmueli (2020). "Algorithmic Fairness." *arXiv preprint arXiv:2001.09784*.

Prince, Anya and Daniel Schwarcz (2020). "Proxy Discrimination in the Age of Artificial Intelligence and Big Data." *Iowa Law Review* 105, no. 3: 1257.

ProPublica (2016). “Machine Bias”

Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy (2020). "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices." In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 469-481.

Robbins, Stephen (1994). *Organizational Behavior*. Prentice–Hall, Saddle River, NJ.

Rooney, Phillis (1992). On Values in Science: Is the Epistemic/Non-Epistemic Distinction Useful? PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, 1992(1):13–22.

Rudner, Richard (1953). The Scientists Qua Scientist Makes Value Judgements. *Philosophy of Science.* 20:1-6.

Singer, Daniel J (2018). Diversity, Not Randomness, Trumps Ability. *Philosophy of Science. Forthcoming.*

Skyrms, Brian (2010). *Signals: Evolution, Learning and Information.* New York: Oxford University Press.

Stanev, Roger (2017). “Inductive risk and outcomes in composite outcome measures,” in K. Elliott and T. Richards (eds.), *Exploring Inductive Risk: Case Studies of Values in Science.* New York: Oxford University Press, 171-192.

Stegenga, Jacob (2016). Three Criteria for Consensus Conferences. Foundations of Science, 21(1):35–49.

Substance Abuse and Mental Services Administration (2016). *National Survey on Drug Use and Health*. Rockville, MD.

Thomas, D. A. & Ely, R. J. (1996) *Harvard Bus. Rev*. September–October 1996, no. 96510.

UCLA (2014). Proposed Diversity Requirement. https://ccle.ucla.edu/pluginfile.php/743624/mod_resource/content/6/082014%20REVISED%20DiversityReqProposal.pdf.

Union of Concerned Scientists (2007). *Smoke, Mirror and Hot Air: How ExxonMobil Uses Big Tobacco's Tactics to Manufacture Uncertainty on Climate Science.* Two Brattle Square Cambridge, MA: 02238-9105.

Wakefield, Andrew, et al., (1998). "Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children." *Lancet. 351(9103):637-41.*

Washington, H., and Cook, J. 2011. *Climate Change Denial: Heads in the Sand.* Earthscan. New York, NY.

Weibel, R.E., Buynak, E.B., Stokes, J., et al. (1967). "Evaluation Of Live Attenuated Mumps Virus Vaccine, Strain Jeryl Lynn, First International Conference on Vaccines Against Viral and Rickettsial Diseases of Man." *World Health Organization,* No. 147.

Weisberg, Michael (2013). *Simulation and Similarity: Using Models to Understand the World.* Oxford University Press.

Williams, Bernard (2002). *Truth and Truthfulness.* Princeton: Princeton University Press.

Winsberg, Eric (2012). "Values and uncertainties in the predictions of global climate models." *Kennedy Institute of Ethics Journal*, 22(2), 111-137.