Uncertainty and Risk Surrounding the Application of Social Models

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.

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

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

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 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.

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) =

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

(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 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 $\varphi(x)$ that is worse than x. The idea behind assumption (b) is that the problem of identifying the best answer must be sufficiently

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

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

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

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.

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.

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.

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

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

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 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

[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.

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 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

[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.

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.

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

[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.

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.

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.