The Epistemic Impact of Theorizing: Generation Bias Implies Evaluation Bias

Finnur Dellsén

Accepted for publication in *Philosophical Studies*
(Please cite published version when available)

Abstract: It is often argued that while biases routinely influence the *generation* of scientific theories (in the 'context of discovery'), a subsequent *rational evaluation* of such theories (in the 'context of justification') will ensure that biases do not affect which theories are ultimately accepted. Against this line of thought, this paper shows that the existence of certain kinds of biases at the generation-stage *implies* the existence of biases at the evaluation-stage. The key argumentative move is to recognize that a scientist who comes up with a new theory about some phenomena has thereby gained an unusual type of evidence, viz. information about the space of theories that could be true of the phenomena. It follows that if there is bias in the generation of scientific theories in a given domain, then the rational evaluation of theories with reference to the total evidence in that domain will also be biased.

1. INTRODUCTION

It is hardly controversial at this point that scientists' own human interests, identities, and ideologies can influence the content of science, i.e. which theories are accepted as true within a particular science. To take a well-known example from evolutionary anthropology, it was once nearly uniformly accepted that the carved stones used as tools by our hominoid ancestors, and which are thought to have provided selection pressure for bipedalism and greater intelligence, were primarily used for hunting other animals. This 'man-the-hunter' model of human evolution was only seriously challenged with the influx of significant numbers of women into evolutionary anthropology in the 1970s. At that point, a 'woman-the-gatherer' model was proposed according to which the carved stones were primarily used to prepare edible vegetation. This episode exemplifies a general

phenomenon, widely discussed by feminist thinkers, of science being biased against theories that challenge dominant ideologies and power structures.[1]

But how, exactly, do the theories accepted in science become biased by the biases of those who practice it? Put differently, how do scientists' own social, political, and moral values – when biased – undermine the objectivity of scientific theories? Several influential accounts have been proposed to answer this and related questions, appealing to factors such as the role of background assumptions in scientific reasoning (Longino 1990; Intemann 2005), differing thresholds for inductive risk (Rudner 1953, Hempel 1965, Douglas 2000, 2009), and the ways in which scientific theory choice is based on 'cognitive' and 'non-cognitive' criteria (Kuhn 1977, Longino 1996). However, discussions of scientific objectivity have generally steered conspicuously clear of appealing to the effect of biases on theory generation, i.e. the process by which scientific theories are conceived of and formulated. For example, Longino (1994: 141-149) discusses five distinct ways in which gender and racial biases can influence scientific research, none of which concerns the generation of theories.[2] Similarly, Reiss and Sprenger's (2017: §3.1) list "four stages at which values [and thus biases] may affect sciences" without mentioning the stage at which scientists conceive of and formulate their theories.[3]

Apparently, then, the possibility of biases in theory generation has not generally been viewed as a significant threat to scientific objectivity. Indeed, even defenders of strong conceptions of scientific objectivity, such as the logical positivists, have seemed happy to acknowledge that biases influence which theories are conceived of and formulated. For example, as Reiss and Sprenger (2017: §3.2) point out, Reichenbach (1938: 6-7) is standardly interpreted as introducing the distinction between the 'context of discovery' and the 'context of justification' in order to argue that biases influence the former but not the latter.

[1] This particular episode is discussed at length by Longino (1990: 106- & 128-131); see also Longino and Doell (1983).

[2] Longino's elements are (i) research practices, (ii) research questions, (iii) research data, (iv) specific background assumptions, and (v) general background assumptions.

[3] Reiss and Sprenger's stages are (i) choosing a research problem, (ii) gathering evidence, (iii) accepting a theory, and (iv) the proliferation and application of results. Reiss and Sprenger are concerned with threats to scientific objectivity due to the influence of various moral, social and political *values*, but these values are more or less equivalent to what I define as a 'bias' below (see §2). Another example of discussions of objectivity and values that steers clear of theory generation is Elliot (2017).

On Reichenbach's view, even though scientists' biases "may influence the discovery, development and proliferation of a scientific theory," they are „irrelevant for justifying the acceptance of a theory, and for assessing how evidence bears on theory" (Reiss and Sprenger 2017: §3.2).

The idea here attributed to Reichenbach seems to be that even though theories may initially be generated in a biased way, a subsequent *rational evaluation* of scientific theories – e.g., by controlled experiments and systematic observation – will ensure that such biases do not affect which theories are ultimately accepted. Thus, although it is granted that the generation-stage of the scientific process is susceptible to various biases, the thought is that the evaluation-stage eliminates the effect of any such biases before scientific theories are accepted. This line of argument was perhaps most clearly expressed by Hempel:

> [...] scientific objectivity is safeguarded by the principle that while hypotheses and theories may be freely invented and *proposed* in science, they can be *accepted* into a body of scientific knowledge only if they pass critical scrutiny, which includes in particular the checking of suitable test implications by careful observation and experiment (Hempel 1966: 206).[4]

Since this argument involves trying to *confine* the influence of biases to the generation of scientific theories, I will refer to it as *the Confinement Defense* of the objectivity of science.

There are a number of ways of undermining this Confinement Defense. One obvious response is to reject or problematize the distinction between the contexts of discovery and justification, as Thomas Kuhn did so influentially (Kuhn 1962: 8; see also, e.g., Barnes 1972: 391; Knorr-Cetina 1981: 28-31; Kantorovich 1993: 101). Another type of response argues that even if Reichenbach's context distinction can be drawn, it does not make scientific theories immune to biases since biases also enter into the context of justification in aforementioned ways, e.g. through background assumptions, thresholds for inductive risks, or the application of cognitive/non-cognitive criteria in theory

[4] Similar arguments are often advocated by scientists themselves. Witness, for example, Carl Sagan in his popular television program *Cosmos*: "There are many hypotheses in science which are wrong. That's perfectly all right […] To be accepted, new ideas must survive the most rigorous standards of evidence and scrutiny" (Sagan 1990).

choice. Here I pursue a different line of response to the Confinement Defense, by developing an underappreciated argument proposed by Kathleen Okruhlik (1994).

In short, this argument aims to show that *if* biases affect the generation of scientific theories (in the 'context of discovery'), *then* such biases will also affect the rational evaluation of theories (in the 'context of justification'). Since the Confinement Defense accepts the antecedent of this conditional but rejects its consequent, this in effect shows that the Confinement Defense is *incoherent*. Dialectically at least, the argument explored in this paper is thus more powerful than either of the two types of responses mentioned above. As we shall see, the argument presented in this paper is also more powerful than Okruhlik's own version of the argument, since it will not be assumed here that we must adopt any specific model of rational theory evaluation in science. The key to this argument is to recognize that a scientist who comes up with a new theory about some phenomena has thereby gained an unusual type of evidence, viz. information about the space of theories that could be true of the phenomena. It follows, I argue, that if there is bias in the generation of scientific theories in a given domain, then the rational evaluation of theories with reference to the total evidence in that domain will also be biased.

2. THE IMPLICATION THESIS

Let us start by defining the type of bias we will be concerned with in this paper. As I will be using the term, an agent or process has a *(theoretical) bias* if she/it privileges theories (hypotheses, conjectures, models) in one class over corresponding theories in a relevant contrast class. For example, someone who has an androcentric bias privileges theories that support or emphasize masculinity and male points of view over corresponding theories that support or emphasize femininity and female points of view. A *generation-bias* is a more specific kind of (theoretical) bias that is exhibited by agents or processes that privilege the generation of one class of theories over another class of theories, in the sense of being more likely to generate theories in the former than in the latter. Similarly, an agent or process has an *evaluation-bias* if she/it privileges the rational evaluation of theories in one class over theories in another, in the sense

of being more likely to positively evaluate (e.g., by accepting or assigning a higher probability to) theories in the former class than theories in the latter.[5]

It is worth noting that, so far, it is not built into these definitions that theoretical biases are pernicious in any way. Indeed, note that these definitions allow one to be biased in favor of truths as against falsehoods, or in favor what the evidence supports as against what it undermines. With that said, I will be interested in the kind of biases that favor one set of theories over another on the basis of considerations that (arguably) have nothing to do with truth or evidential support, such as gender and racial biases. These biases correspond roughly to what Longino (1990: 4-6) calls 'contextual values', i.e. the personal, social, and cultural values that belong to the broader context in which science is done (as opposed to the 'constitutive values' that determine what counts as acceptable scientific practice). In this paper, we will concentrate on this narrower set of biases grounded in contextual values.[6]

Also worth noting is that I will mostly be interested in biases that operate at the level of groups or communities of scientists, rather than at the level of individual scientists. To motivate this focus, consider the possibility that the individual scientists that comprise a community could be heavily biased in 'opposite' ways, so that the net effect of individual biases is an unbiased scientific community. For example, if roughly half of a scientific community has a strong androcentric evaluation-bias, while the other half has a correspondingly strong gynocentric evaluation-bias, then the overall effect might be that the community is no more likely to positively evaluate theories that support masculine or male points of view than those that support feminine or female point of view.[7] The

[5] Note that my definition of 'bias' here differs from Antony's 'empiricist' definition of bias as "possession of belief or interest prior to investigation" (Antony 1993: 188). For the purposes of this paper, a more 'operational' definition is appropriate, i.e. one on which bias can be identified in terms of the agent's dispositions to behave in certain ways rather than her belief or interests (which may or may not be manifested in the agent's behavior). Furthermore, it is not at all clear whether, or how, Antony's definition could be made to subsume implicit biases, which are usually taken to be non-doxastic states (i.e. not beliefs) and which clearly need not line up with the agent's interests.

[6] Thus, in what follows, the term 'bias' should always be taken to refer to biases that are grounded in contextual values.

[7] Indeed, a number of feminist thinkers have argued that the most promising way to make science as a whole more objective or unbiased is to ensure that scientists have complementary biases in roughly this way (see, e.g., Longino 1990, 2002; Antony 1993; Solomon 2001).

more worrying phenomenon is when a scientific community as a whole exhibits a bias, e.g. because while some of the scientists are relatively unbiased, a large enough subgroup is heavily biased in the roughly the same way. It is this type of *community-level bias* that I will primarily be concerned with here.

Given these stipulations of the kinds of biases I will be concerned with, the thesis for which I will be arguing in this paper is, roughly, that *if* there is generation-bias in some theoretical domain, *then* there is also evaluation-bias within that same domain. In slogan form, *generation-bias implies evaluation-bias*; I will refer to this as *the Implication Thesis*. However, let me immediately flag that I will later qualify this by locating a specific subcategory of generation-bias – what I will call *competitor-generation-bias* – and argue that this specific type of generation-bias implies evaluation-bias. Although I will thus be arguing for a qualified Implication Thesis, the upshot is much the same for defenses of scientific objectivity such as the Confinement Defense (more on this in section 5).

Let me end this section by contrasting the claim for which I will be arguing with a more innocuous sense in which biases in the generation of scientific theories effect how such theories are rationally evaluated. Clearly, scientists can only evaluate theories that have already been formulated, so if the set of theories that have been generated is biased, then so too is the set of theories that could be evaluated as confirmed (or, indeed, as disconfirmed) by the available evidence. In this way, positive theory evaluation, and thus the potential acceptance of theories, is necessarily constrained by the (possibly biased) process of generating theories.[8] This point, although of course correct, is significantly weaker than the thesis for which I will argue in the present paper.[9] My point will not merely be

[8] Many thanks to an anonymous reviewer for suggesting that I contrast the Implication Thesis with this more innocuous point.

[9] Indeed, it seems to me that this point would not threaten the Confinement Defense of scientific objectivity at all, at least not on a plausible construal thereof. After all, a proponent of the Confinement Defense could respond that, even granting this point, each individual theory that has in fact been generated and evaluated (positively or negatively) would be evaluated in just the same way regardless of whether it and its competitors were generated in a biased way. In particular, such a proponent could argue that a theory that is sufficiently positively evaluated to be accepted would still be evaluated in just the same way regardless of whether its generation was biased or unbiased. Thus the fact that only theories that have been generated could be (positively) evaluated would not threaten the central contention of the Confinement Defense that biases in theory generation do not undermine our reasons for accepting the theories that we do in fact accept.

that scientists can only evaluate theories that have been generated, but that some forms of theory generation effect what the result of the evaluations will be, i.e. whether and the extent to which such an evaluation is positive or negative. Put differently, my contention here is not simply that biases in theory generation effect *whether* some theory is evaluated – and thus potentially accepted – but also *how* (positively or negatively) the theory is evaluated.

3. OKRUHLIK'S ARGUMENT

The argument that I will give for the Implication Thesis (or a qualified version thereof) is inspired by, and can be viewed as a development of, an underappreciated argument given by Kathleen Okruhlik.[10] Before I spell out my own version of the argument, I will briefly consider Okruhlik's original argument and what I consider to be an important limitation of the argument.

In her "Gender and the Biological Sciences" (1994), Okruhlik argues that what I am calling the Confinement Defense "makes no sense at all" if we accept what she refers to as an "irreducibly comparative" model of scientific rationality (Okruhlik 1994: 33). According to Okruhlik, non-comparative models of theory evaluation in science have become obsolete: "we now recognize that one does not actually compare the test hypothesis to nature directly in the hope of getting a 'yes' or 'no' ('true' or 'false') answer; nor does one compare it to all logically possible rival hypotheses" (Okruhlik 1994: 33). Rather, says Okruhlik, one always compares a hypothesis with other available hypotheses that have been articulated and developed to the point of being testable.[11] From this Okruhlik

[10] Another argument that comes close to the one I will make below is sketched briefly by Elliot and McKaughan (2009: 607-608), who argue that proposing new theories "can transform what appeared to be irrelevant facts into crucial pieces of evidence" (Elliot and McKaughan 2009: 608). However, it is not clear from Elliot and McKaughan's brief discussion what it is for evidence to be 'transformed' in their sense, especially since they appear to deny that this type of transformation "alter[s] the evidential relationship between the available theories and data" (Elliot and McKaughan 2009: 608). By contrast, I argue below that proposing new theories can make it rational to evaluate old theories less favorably (e.g., by it becoming rational to assign a lower probability to them), even when there is no change in the relevant empirical data. In this sense, *pace* Elliot and McKaughan, I maintain that proposing new theories can alter the evidential relationship between available theories and data.

[11] This idea is reminiscent of some early conceptions of Inference to the Best Explanation (IBE), where the evaluative step merely renders the comparative verdict that one theory provides a *better* explanation than available alternatives (e.g., Harman 1965; Thagard

infers that evaluations of theories, even when perfectly rational, cannot determine whether a theory is likely to be true or false in an absolute sense, but "only that it is epistemically superior to the other actually available contenders" (Okruhlik 1994: 34).

As Okruhlik points out, this would mean that contra the Confinement Defense, "nothing in the appraisal machinery will completely 'purify' the successful theory" (Okruhlik 1994: 34). In short, this is because a theory may be epistemically superior to the theories in one class of available alternatives but not another. So, if the set of available theories – i.e. the set of theories that have been generated in a given domain – is disproportionally populated by theories which conform to some particular bias, then the theory that ends up being evaluated as the 'best' of these available theories will presumably be more likely to conform to that bias as well. In the extreme case, every single one of the available theories would conform to the bias in question, in which case the 'best' of them would inevitably do so as well.

It is worth emphasizing that Okruhlik's argument relies on her assumption that scientific methodology is capable only of delivering comparative evaluations of scientific theories – i.e. her 'irreducibly comparative' model of scientific rationality. The idea here is not merely that comparisons between available theories is an important part of the scientific process, in that such comparisons will figure as part of the input, or part of the process itself, of rational theory evaluation. That much is undeniable and uncontroversial – theories are clearly not evaluated in isolation from other competing theories. Indeed, as we shall see, even models of scientific rationality that are non-comparative in Okruhlik's sense – i.e., in the sense of rendering science capable of delivering absolute verdicts regarding its theories – can accommodate this rather straightforward point about the importance of comparisons in theory evaluation.

However, Okruhlik also makes the stronger claim that scientific methodology is incapable of delivering *verdicts* that are stronger than the comparative claim that one theory is 'epistemically superior' to its extant rivals. This is a claim that concerns the output, rather than the input or the process itself, of scientific theory evaluations. In particular, Okruhlik claims that one cannot

1978). However, as I explain below, Okruhlik's model is much more radically comparative than standard conceptions of IBE.

determine whether a claim is true or false, or indeed probably true or false, since this goes beyond the comparative claim that one hypothesis is superior to another.[12] In order to flag this specific and stronger sense in which Okruhlik suggests that scientific rationality is 'irreducibly comparative', I will refer to Okruhlik's model of scientific rationality as *irreducibly verdict-comparative*.

Okruhlik's commitment to the idea that scientific rationality is irreducibly verdict-comparative is essential to her argument against the Confinement Defense; Okruhlik's argument does not go through without it. To see why, note that if scientific methodology were capable of delivering absolute as opposed to merely comparative evaluations, the Confinement Defense will simply claim that each theory can be evaluated as (probably) true or false regardless of which theories have been generated at a particular point in time. It doesn't matter whether, in the process of making this kind of absolute evaluation of theories, scientists often (or even always) compare one theory to another. After all, Okruhlik is not arguing that evaluative comparisons between theories are themselves subject to biases; rather, her argument is explicitly meant to establish that biased theory-generation would lead to biased theory-acceptance even if it is granted for the sake of the argument that comparisons between theories are unbiased (Okruhlik 1994: 33).

In my view, Okruhlik's argument gets at something important and is underappreciated in the current literature on scientific objectivity, bias and values. However, its reliance on the idea that scientific rationality is irreducibly verdict-comparative is a significant weakness of the argument in its current form. This is so for two related reasons. First, appealing to an irreducibly verdict-comparative model of scientific rationality is *dialectically weak*, since most if not all proponents of the Confinement Defense will reject such a model for independent reasons. Contrary to what Okruhlik seems to suggest, philosophers of science do not generally reject contrary models of scientific rationality, i.e. what we may call *verdict-absolutist* models. Indeed, the model of scientific rationality that has the strongest claim to being the current orthodoxy among

[12] Thus Okruhlik would have to deny, for example, that the best explanation of one's evidence is probably and/or approximately true. On her view, assuming she accepts some form of IBE (see previous footnote), we could at most assert that it is *likelier to the true*, or perhaps *more approximately true*, than available rival explanations.

philosophers of science is *Bayesian Confirmation Theory* (BCT),[13] which dictates that each scientific theory under consideration be assigned an *absolute* numerical probability value. This clearly goes beyond a mere comparative verdict that one theory is epistemically superior to another. It is of course true that, given the absolute probabilities of two theories T_1 and T_2, we can also (trivially) compare their probabilities within BCT. But the point here is that BCT is not *irreducibly verdict-comparative*, since such a comparison is based on – and *reduces to* – a comparison of absolute probabilities.

Indeed, the same is true of the other model of scientific rationality that enjoys widespread popularity among philosophers of science, *Inference to the Best Explanation* (IBE). For although an instance of IBE certainly involves a comparative evaluation of one theory as providing a 'better' explanation than competing theories, it also involves inferring that the relevant theory is (probably and/or approximately) true, as opposed to merely that the theory is epistemically superior to alternatives (Douven 2017a: §2). The comparison involved in IBE with reference to theories' explanatory virtues is a step in the process of making the inference; it is not the inference's conclusion or verdict. Indeed, the fact that the conclusion of IBE is absolute while the explanatory comparison involved in it isn't gives rise to a well-known problem for IBE, viz. that the best explanation might merely be the best of a bad lot (van Fraassen 1989: 142-3).[14] So even IBE, which is comparative in an important sense, involves arriving at the kind of absolute ('yes' or 'no') verdicts that Okruhlik's conception of scientific rationality explicitly does not allow for.

Second, Okruhlik's contention that theory evaluation is irreducibly verdict-comparative is *implausible* as a description of actual scientific practice (or indeed as a prescription for what science ought to be like), because scientists *can*

[13] The dominance of BCT among contemporary philosophers of science is acknowledged by its proponents (e.g., Earman 1993: 2; Strevens 2017: 5) as well as its critics (e.g., Godfrey-Smith 2003: 202; Norton 2018: 3).

[14] Douven (2017b) refers to this problem as the *asymmetry problem*. On Douven's description of the problem, the issue is that most formulations of IBE "license an inference to an absolute verdict—that a given hypothesis is true—from what will typically only be a relative judgment, namely, that the hypothesis is the best explanation among those on the table" (Douven 2017b: 9). It is perhaps worth noting that some influential conceptions of IBE propose to avoid this problem by including an 'absolutist' requirement on the conditions for IBE to the effect that the inferred explanation should not merely be the best, but also "satisfactory" (Musgrave 1988) or "good enough" (Lipton 2004); see also Dellsén (2017, 2018) for a different approach to the problem.

and do reach non-comparative verdicts about many scientific theories. Consider, for instance, the theory of natural selection, the atomic theory of matter, the double-helix structure of DNA, the kinetic theory of heat, and the theory that human activity is a significant cause of increased global temperatures (anthropogenic climate change). According to Okruhlik's model of scientific rationality, the most we can say about these theories is that they are 'epistemically superior' to their extant rivals. However, if these theories could not be evaluated absolutely, i.e. as probably true or false, scientists would not be justified in relying on them for predictions, explanations, and public engagement in the way that they often do. For example, when IPCC scientists announced that "[i]t is *very likely* [defined as probability 90-100%] that human influence has contributed to the observed global scale changes in the frequency and intensity of daily temperature extremes since the mid-20th century" (Intergovernmental Panel on Climate Change 2014: 7), they were explicitly reporting an absolute estimation. They were not, by contrast, merely reporting that the relevant claim is epistemically superior to currently available alternatives.

None of this is to deny that the availability of competing theories plays an important role in scientific reasoning. Indeed, my argument below for a qualified Implication Thesis is partly based on an analysis of how the availability of competing theories influences rational evaluation of scientific theories. However, Okruhlik's contention that it is impossible to reach non-comparative verdicts in scientific theory evaluation greatly overstates the extent to which competing theories dictate this process, and thus opens up Okruhlik's argument to the charge that its key assumption – that scientific rationality is irreducibly verdict-comparative – does not square either with scientific practice or with widely accepted models of scientific rationality, such as Bayesian Confirmation Theory and Inference to the Best Explanation.

4. THEORIES AS EVIDENCE (OF A SORT)

During the decade following the publication of Einstein's special theory of relativity in 1905, European physicists became increasingly confident that Einstein's new theory was true, and that its previously-accepted alternatives, such as Lorentz's ether theory, were false. And yet, as Earman (1992: 196-7) points out, little new empirical evidence pertaining to these alternative theories was recorded during the period. Indeed, Einstein's own paper (Einstein 1905)

famously did not report any new observations or experiments; rather, the paper simply appealed to some well-known physical anomalies, such as the fact that no ether drift had ever been observed, and formulated a new theory that seemed to explain these empirical phenomena better than any of the previously available theories.

The important point here is that physicists came to significantly re-evaluate previously available theories (such as Lorentz's ether theory) over a period when they gained next to no new empirical evidence. But this presents us with an apparent difficulty: On the one hand, if one gains no new evidence between one point in time and another, it seems that rationality would require that one's evaluation of any theory should remain the same (at least if one's initial evaluation of the theory was itself rational). On the other hand, common sense and scientific practice both suggest that it was rational for physicists to significantly change their evaluation of Lorentz's ether theory, for example, when learning about Einstein's new theory. How can these two claims be reconciled?

The answer is simple: Einstein's discovery of the special theory of relativity does constitute a type of evidence after all, viz. additional information about the space of theories that could explain the physical phenomena in question. Whereas it was previously thought that any plausible theory of the mechanics of moving bodies would have to posit an absolute reference frame in which the physical laws held true, Einstein's discovery of special relativity revealed (among other things) that this assumption was not necessary. Thus, Einstein's theory effectively showed that a region of logical space that was previously thought to be ruled out by experiment does indeed contain plausible contenders to then-dominant theories such as Lorentz's ether theory.[15]

Of course, this piece of theoretical information is clearly not of the usual *empirical* kind that we tend to associate with the term 'evidence'. But it is also clear that it still counts as information of the sort that a rational agent should take into account when evaluating a given scientific theory. If we want to reserve the term 'evidence' for *empirical* evidence, such as observations and experimental

[15] The type of situation described here is in a sense the converse of the current situation in particle physics, in which repeated failed attempts to come up with a plausible alternative to string theory has arguably contributed to scientists becoming quite confident that no such alternative exists (Dawid 2013; Dawid, Hartmann, and Sprenger 2015).

results, then we could say that Einstein's discovery is *evidence in an extended sense*. Which label we choose for Einstein's discovery is not important for our purposes; what's important is that it is possible to gain a type of purely theoretical information about the space of theories that could explain a given set of data, and that this type of information can alter the rational evaluation of previously available explanations for that data.

It is worth noting that the point I am making here has been implicitly and explicitly acknowledged by both proponents and critics of Bayesian Confirmation Theory (BCT).[16] Thus sympathetic critics of BCT, such as Chihara (1987: 556-60) and Earman (1992: 195-8) argue that discoveries of new alternative theories of precisely this sort present a special difficulty for BCT, because Bayesian conditionalization cannot explain the rationality of assigning probabilities to entirely new theories and changing one's probabilities in the old theories. This is known as *the problem of new theories* (related, but not identical, to *the problem of old evidence* – see Glymour 1980). In response, Bayesians such as Maher (1995) and Wenmeckers and Romeijn (2013) argue that conservative extensions of orthodox Bayesianism *can* provide a rule for assigning probabilities to new theories and for modifying the probabilities assigned to extant theories.

My concern here is not with determining which, if any, of these Bayesian responses to the problem of new theories is correct. Rather, I mention this debate in order to highlight that the very notion that new theories present a problem for BCT presupposes that discovering new theories can have the epistemic impact that I have described – i.e. that it counts as 'evidence' in the extended sense identified above. To see this clearly, consider how the Bayesian will describe the evidential situation before and after discovering a new theory T_{new}. *Before* discovering T_{new}, the Bayesian agent will assign subjective probabilities to a set of already-conceived competing theories $T_1,\ldots,T_k$ (not including T_{new}), along with a 'catch-all hypothesis' C which effectively asserts that $T_1,\ldots,T_k$ are all false. The axioms of the probability calculus demand that these probabilities, $P_{before}(T_1),\ldots,P_{before}(T_k)$, and $P_{before}(C)$ sum to unity. *After* discovering T_{new}, the

[16] I choose to focus on BCT in what follows in part because it is by far the most widely endorsed framework for rational theory evaluation among philosophers of science (see footnote 13); in part because BCT clearly provides the means to evaluate theories in an absolute – i.e. not merely comparative – manner (in contrast to Okruhlik's 'irreducably comparative' model); and in part because BCT has well-known *prima facie* difficulties in handling the epistemic impact of new theories.

Bayesian agent will assign probabilities to the original theories $T_1,\ldots,T_k$; the new theory T_{new}; and a new catch-all hypothesis C_{new} which asserts that $T_1,\ldots,T_k$ *and* T_{new} are all false. As before, the probability axioms demand that these new probability assignments, $P_{after}(T_1),\ldots,P_{after}(T_k)$, $P_{after}(T_{new})$, and $P_{after}(C_{new})$, sum to unity. Thus, unless our Bayesian agent had already somehow anticipated that she would come to discover a plausible new theory, she will assign a higher probability to the disjunction of T_{new} and C_{new}, i.e. to the claim that the new theory is true or that some yet-to-be-conceived theory is true, than she assigned to the original catch-all C.[17] But then it follows that, in order to satisfy the demands of the probability axioms, the Bayesian agent must lower her probability assignments regarding at least some of the theories $T_1,\ldots,T_k$ – in particular, she must adjust them such that $P_{before}(T_1) + \ldots + P_{before}(T_k) > P_{after}(T_1) + \ldots + P_{after}(T_k)$.[18]

Of course, this does not by itself solve the Bayesian problem of new evidence, for it does not say what probability to assign to T_{new} after discovering it, or indeed how exactly to adjust the probability of $T_1,\ldots,T_k$.[19] What it does show is that any general solution to the problem of new evidence must allow for the probabilities for $T_1,\ldots,T_k$ to change in light of the discovery of the new theory T_{new} and subsequent probability assignments to T_{new} and the new catch-all C_{new}. Thus the discovery of T_{new} is 'evidence' in this extended sense of being a piece of information that should, rationally, lead one to alter one's subjective probabilities – in this case, by lowering the probability assignments to competing theories $T_1,\ldots,T_k$. This is of course exactly what happened in the case discussed above,

---

[17] Any general Bayesian solution must at least allow for this possibility. Indeed, this is exactly the sort of situation that is described when the problem of new evidence is described (see, e.g. Earman 1992: 196-7).

[18] To see this, note first that since C_{new} is (by construction) incompatible with T_{new}, the probability of their disjunction is equal to the sum of their individual probabilities: $P_{after}(T_{new}$ or $C_{new}) = P_{after}(T_{new}) + P_{after}(C_{new})$. So the situation we are focusing on is one where:

(1) $P_{before}(C) > P_{after}(T_{new}) + P_{after}(C_{new})$

Now, as noted, the probability axioms demand that the probabilities before and after both sum to unity, i.e. that:

(2) $P_{before}(T_1) + \ldots + P_{before}(T_k) + P_{before}(C) = P_{after}(T_1) + \ldots + P_{after}(T_k) + P_{after}(T_{new}) + P_{after}(C_{new}) = 1$

(1) and (2) jointly entail that $P_{before}(T_1) + \ldots + P_{before}(T_k) > P_{after}(T_1) + \ldots + P_{after}(T_k)$, as desired.

[19] For that, I refer the reader to Maher 1995 and Wenmeckers and Romeijn 2013.

where Einstein's special theory of relativity plays the role of the new theory T_{new} and Lorentz's ether theory as one of $T_1,\ldots,T_k$.

This brief excursion into Bayesian territory may have given the impression that the type of non-empirical evidence I am concerned with can only be accounted for in a Bayesian framework, BCT. Not so.[20] As the example of Einstein's discovery of special relativity illustrates, scientists can, due to their theoretical discovery of a previously unconceived alternative, come to rationally reevaluate theories even in the absence of empirical evidence for or against those theories. This is a *datum* – a fact of scientific life – that any model of scientific rationality worth its salt will have to reckon with in one way or another. Thus a model of scientific rationality that fails to account for the possibility of rational reevaluations of this type is *ipso facto* inadequate. While I have used BCT to illustrate how such a reevaluation could be manifested in that particular model, the same phenomenon will therefore necessarily resurface in any adequate account of scientific rationality that could serve as an alternative to BCT.

5. THE IMPLICATION THESIS REVISITED

For our purposes, the crucial upshot of these considerations is that the extent to which a given scientific theory is positively evaluated depends in part on whether (and the extent to which) plausible alternative theories have been conceived and formulated. To see what this has to do with the Implication Thesis, note that with respect to an extant theory T, scientists can be more and less likely to develop *alternatives* to T – i.e., competing theories of the same set of phenomena. Thus it may happen that due to some contextual value, scientists are more (or less) likely to develop alternatives to T than they would otherwise be, e.g. if T challenges a prevalent gender stereotype. This would be a kind of generation-bias in favor of the class of alternatives to T; let us call it *competitor-generation-bias.*

Let me illustrate this type of bias with a couple of examples from the history of science. First consider R. A. Fischer's opposition to the causal link between smoking and lung cancer (Fischer 1959). After a distinguished career as a statistician and geneticist, Fischer retired from his position at Cambridge in

[20] Indeed, as noted in footnote 16, I have chosen to discuss how to model this type of evidence within BCT in part because it is initially not at all obvious that BCT could accommodate evidence of this type at all.

1957 and shortly afterward began to publicly question the notion that smoking causes lung cancer, which was increasingly becoming widely accepted by medical researchers at the time. One of Fischer's main lines of opposition consisted in developing an alternative hypothesis to explain the well-documented statistical correlation between smoking and lung cancer. In particular, Fischer proposed that lung cancer causes smoking rather than the other way around, via an unconscious irritation or pain that is caused by lung cancer and that causes (increased) smoking. As Fischer puts it, "anyone suffering from a chronic inflammation in part of the body (something that does not give rise to conscious pain) is not unlikely to be associated with smoking more frequently, or smoking rather than not smoking" (1959: 22).

I leave it to the reader to decide whether Fischer's explanation was plausible, even relative to the empirical evidence available at the time. The important point here is that Fischer's proposal of this hypothesis was fairly clearly influenced by what we would now refer to as his contextual values. Not only was Fischer himself a smoker of cigarettes and pipes; he was also a political conservative who was skeptical of taxation and government regulation of private industry, such as that which was being proposed to reduce smoking; furthermore, Fischer also received a fee from the tobacco industry (although to be fair Stolley (1991: 425) estimates that the fee was "probably not large"). In sum, although it is certainly possible for Fischer's interest in proposing alternative explanations of the correlation between smoking and lung cancer to have been motivated by non-contextual factors, it is at least plausible that this episode illustrates the influence of contextual bias at the stage of generating theories.

The other example that I propose as plausibly exemplifying competitor-generation-bias concerns evolutionary explanations of female orgasm. On one way of carving up logical space, there are two possible types of evolutionary explanations of the fact that female humans have orgasms in sexual intercourse: this is either an *adaptation* – i.e. a trait that has been selected for in natural selection – or a *spandrel* – i.e. a trait that has evolved as a byproduct of some other trait or evolutionary process. Each type of explanation will have to be fleshed out so as to answer its own distinctive types of questions. For example, the first type of explanation will have to spell out what selection pressures gave rise to female orgasm, while the second type will have to say what the female orgasm is a

byproduct of. So there is considerable room for theorizing within each of the two explanation-types.

Nevertheless, as Lloyd's (2005; see also 1993) documents, the theoretical landscape is dominated by adaptation-based explanations. Of the 21 explanations Lloyd reviews, all but one assume that the female orgasm is an adaptation rather than a spandrel. This apparent preference for adaptation-based explanations is not justified by the available evidence, which is at best equivocal and at worst favors spandrel-based explanations over its adaptation-based counterparts.[21] Lloyd attributes this surprising situation in part to a general bias in evolutionary biology for adaptationist explanations, but also – and more significantly for our purposes – to androcentrism, including the implicit assumption that female sexuality is like male sexuality; and to a focus on procreation as the only type of evolutionarily significant sexual intercourse (Lloyd 2005: 229-235). The latter are clearly based on contextual values, and thus count as (contextual) biases in the relevant sense. Specifically, they are a form of competitor-generation-bias, since they influence the generation of explanations for female orgasm which could serve as alternatives to the various adaptation-based explanations that currently dominate the field.

Now, what has this to do with the Implication Thesis and the Confinement Defense of scientific objectivity? Well, as we have seen, the availability of plausible competitors to a given theory undermines the epistemic status of that theory in a rational evaluation (as the availability of special relativity undermined the epistemic status of Lorentz's ether theory). Thus if there is a competitor-generation-bias in favor of generating alternatives to T, then the epistemic status of T will be more likely to be undermined by the presence of plausible alternatives than it would otherwise be. It follows that, all other things being equal, scientists will be less likely to have a positive rational evaluation of T than they would otherwise have, due to nothing other than the fact that there is a competitor-generation-bias in favor of generating alternatives to T. This establishes the Implication Thesis in a suitably qualified form: *competitor-generation-bias implies evaluation-bias*.

Importantly, this argument does not presuppose any specific conception of scientific rationality or theory evaluation. Thus, while Okruhlik's argument

[21] Lloyd herself (2005: 107-148) argues that the spandrel-based explanation, due to Symons (1979), is most plausible.

relied on an irreducibly verdict-comparative model of scientific rationality, the current argument goes through even if it is assumed that scientific theories can be, and are, rationally evaluated in an absolute or non-comparative manner – i.e., as (approximately and/or probably) true or false. Specifically, we have seen how the discovery of new competing theories can constitute a kind of evidence in an extended sense, in that it leads rational agents to revise their evaluations of previously available theories. This holds even on verdict-absolutist models of scientific reasoning such as Bayesian Confirmation Theory, since even Bayesians acknowledge that discovering new theories can and do have this type of epistemic impact.

Where does this leave us with regard to the Confinement Defense of scientific objectivity? Recall that the Implication Thesis, if true, would make the Confinement Defense incoherent, since the latter explicitly grants the prevalence of biases in theory-generation but denies that they play any role in rational theory-evaluation. Of course, proponents of the Confinement Defense could avoid incoherence by retreating from the first claim, i.e. by denying the existence of generation-bias in science. However, this response suffers from the sheer implausibility of claiming that biases based on contextual values cannot play any role in the process of identifying and formulating scientific theories. The initial appeal of the Confinement Defense was that it seemed to offer a way of defending the scientific process as fundamentally rational even while admitting that one of its constituent parts, viz. theory-generation, would never be immune to bias. If the Implication Thesis is true, this is a hopeless task.

Although I have argued for a *qualified* version of the Implication Thesis – i.e. that competitor-generation-bias implies evaluation-bias – the upshot for the Confinement Defense is much the same. If the qualified Implication Thesis is true, proponents of the Confinement Defense can only avoid incoherence by denying the existence of competitor-generation-biases, i.e. biases in favor of generating alternatives to some theories rather than others. But this too is exceedingly implausible in many cases, as is illustrated by Fischer's development of the hypothesis that lung cancer causes smoking and the conspicuous dearth of spandrel-based explanations of female orgasm in comparison to (arguably androcentric and procreation-focused) adaptation-based explanations. More generally, it would be nothing short of a miracle if scientists' ideologies, political beliefs, social commitments, etc. – in short, their contextual values – did not

regularly lead them to focus their attention on conceiving and formulating alternatives to some theories at the expense of others. This is especially so for theories with significant social or political implications, where scientists may have strong incentives to identify and develop alternative theories that accord with their contextual values – or the contextual values of those who fund or influence them.

One might still insist that, ideally, scientists should generate *all* possible theories of a given phenomenon – or at least all those possible theories that would be worth taking seriously – before they evaluate and potentially accept any one of these. This would effectively eliminate the possibility of generation-bias at the time of theory evaluation, since it would make it impossible for the scientific community to be more likely to have generated theories in one class than in another. (All theories would be equally likely to have been generated, viz. 100% or maximally likely.)[22] The obvious problem with this suggestion is that in actual scientific practice it is rarely, if ever, feasible to generate all or even most (serious) theories of any interesting phenomenon before any of them is evaluated. Science is done in real time, and this requires scientists to judge the plausibility of theories long before they could become confident that all or most possible theories have been generated. Of course, things would be different if scientists or humans generally were theoretically omniscient. But since they are not, and arguably never will be, this particular counterfactual is irrelevant to an analysis of how biases influence actual scientific practice.

7. CONCLUDING REMARKS

The Confinement Defense of the objectivity of science relies on the idea that the process of rationally evaluating scientific theories is not subject to bias even though the process of generating theories undeniably is. Developing an argument proposed by Okruhlik (1994), I have argued that this position is unstable, because the existence of one type of bias in the generation of scientific theories *implies* that the rational evaluation of theories will also be biased. In closing, I wish to draw out two broader implications of this argument for

[22] Recall the definition of 'generation-bias' at the beginning of section 2.

philosophy of science and the practical issue of how to organize scientific communities.

First, the argument shows that philosophical discussions of scientific objectivity should not ignore biases that operate at the stage of theory-generation (i.e., in the 'context of discovery', in one sense of that term). All too often, what I have called generation-bias is ignored or treated as irrelevant in discussions of scientific objectivity and biases, presumably because it has been assumed that such biases do not ultimately affect which theories are accepted in science (Longino 1994: 141-149; Reiss and Sprenger 2017: §3.1; see also Elliot 2017: 10). The argument of the present paper shows that discussions of this kind are at best incomplete; at worst, they may falsely lead one to conclude that science will be objective or unbiased once the influence of (contextual) biases have been eliminated from the factors they do consider.

Second, what are the practical implications of the above argument for how to counteract pernicious biases in science? Perhaps in contrast to some other kinds of biases that operate in science, it is hard to see how the issue of generation-bias could feasibly be addressed at the individual level, e.g. by reeducating individual scientists or incentivizing individual behaviors. After all, one cannot prevent scientists from conceiving of and proposing scientific theories that accord with their biases without instituting some form of active censuring or thought-policing. Accordingly, the more feasible solution may be to make sure that the scientific community at a given time exhibits diversity with regard to which kinds of theories each scientist is likely to generate.[23] Even if each scientist within such a community is biased in their own way, diversity may ensure that the scientific community as a whole is relatively unbiased, since the bias of each scientist would be complemented with another scientist's opposite bias.[24] The upshot may thus be that the best way to promote the relevant kind of scientific

[23] Or at a minimum to try to minimize the effects of forces that cause scientific communities to become more homogenous in this regard, such as what Holman and Bruner (2017) call "industrial selection".

[24] As I have noted above (footnote 7), similar solutions have been proposed to counteract other sorts of biases in science, e.g. by Longino (1990, 2002), Antony (1993) and Solomon (2001).

objectivity involves the independently desirable aim of diversifying scientific communities, e.g. with regard to gender and racial identities.[25]

REFERENCES

Antony, L. 1993. Quine as a Feminist: The Radical Import of Naturalized Epistemology. In A Mind of One's Own, ed. L. Antony and C. Witt. Boulder: Westview, pp. 185-225.

Barnes, B. 1972. Sociological explanation and natural science: A Kuhnian reappraisal. *Archives Européens de Sociologie* 13: 373-393.

Chihara, C.S. 1987. Some Problems for Bayesian Confirmation Theory. *The British Journal for the Philosophy of Science* 38: 551-560.

Dawid, R. 2013. *String Theory and the Scientific Method*. Cambridge: Cambridge University Press.

Dawid, R., S. Hartmann, and J. Sprenger. 2015. The No Alternatives Argument. *The British Journal for Philosophy of Science* 66: 213-234.

Dellsén, F. 2017. Abductively robust inference. *Analysis* 77: 20-29.

Dellsén. F. 2018. The heuristic conception of inference to the best explanation. *Philosophical Studies* 175: 1745-1766.

Douglas, H. 2000. Inductive Risk and Values in Science. *Philosophy of Science* 67: 559-79.

Douglas, H. 2009. *Science, Policy, and the Value-Free Ideal*. Pittsburgh: University of Pittsburgh Press.

Earman, J. 1992. *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge: MIT Press.

Einstein, A. 1905. Zur Elektrodynamik bewegter Körper. *Annalen der Physik* 17: 891-921.

Elliot, K. C. and D. J. McKaughan. 2009. How Values in Scientific Discovery and Pursuit Alter Theory Appraisal. *Philosophy of Science* 76: 598-611.

Elliot, K.C. 2017. *A Tapestry of Values*. Oxford: Oxford University Press.

Fischer, R. A. 1959. *Smoking – the Cancer Controversy; Some Attempts to Assess the*

[25] For helpful discussions and feedback on this paper, I am grateful to Juan Colomina-Almiñana, Whitney Lilly, the Cottage Reading Group, and audiences at the Central APA in Chicago (2018), the Wintergames at Inland Norway University of Applied Sciences (2017), and the conference Who's Got the Power at the University of Iceland (2017).

Evidence. Edinburgh: Oliver and Boyd.

Glymour, C. 1980. *Theory and Evidence*. Princeton: Princeton University Press.

Godfrey-Smith, P. 2003. Theory and Reality: An Introduction to the Philosophy of Science. Chicago: University of Chicago Press.

Harman, G. 1965. The inference to the best explanation. *The Philosophical Review* 74: 88-95.

Hempel, C. G. 1965. Science and Human Values. In *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science.* New York: The Free Press, pp. 81-96.

Hempel, C. G. 1966. *Philosophy of Natural Science*. Englewood Cliffs: Prentice-Hall.

Holman, B. and J. Bruner. 2017. Experimentation by Industrial Selection. *Philosophy of Science* 84: 1008-1019.

Intemann, K. 2005. Feminism, underdetermination, and values in science. *Philosophy of Science* 72: 1001-1012.

Intergovernmental Panel on Climate Change 2014. Climate Change 2014 Synthesis Report: Summary for Policymakers. Available at: <http://www.ipcc.ch/pdf/assessment-report/ar5/syr/AR5_SYR_FINAL_SPM.pdf>.

Kantorovich, A. 1993. *Scientific Discovery – Logic and Tinkering*. New York: State University of New York Press.

Knorr-Cetina, K. 1981. *The Manufacture of Knowledge*. Oxford: Pergamon Press.

Kuhn, T. S. .1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Kuhn, T. S. 1977. Objectivity, Value Judgment, and Theory Choice. In *The Essential Tension*. Chicago: University of Chicago Press, pp. 320-39.

Lipton, P. 2004. *Inference to the Best Explanation*. Second edition. London and New York: Routledge.

Lloyd, E. A. 1993. Pre-theoretical Assumptions in Evolutionary Explanations of Female Sexuality. *Philosophical Studies* 69: 139-153.

Lloyd, E. A. 2005. *The Case of the Female Orgasm: Bias in the Science of Evolution*. Cambridge, MA: Harvard University Press.

Longino, H. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton: Princeton University Press.

Longino, H. 1994. Gender and Racial Biases in Scientific Research. In K. Shrader-Frechette, *Ethics of Scientific Research*. Lanham: Rowman and Littlefield,

pp. 139-151.

Longino, H. 1996. Cognitive and Non-Cognitive Values in Science: Rethinking the Dichotomy. In L. H. Nelson and J. Nelson (eds.), *Feminism, Science, and the Philosophy of Science*. Dordrecht: Springer, pp. 39-58.

Longino, H. 2002. *The Fate of Knowledge*. Princeton, NJ: Princeton University Press.

Longino, H. and R. Doell. 1983. Body, Bias, and Behavior: A Comparative Analysis of Reasoning in Two Areas of Biological Science. *Signs: Journal of Women in Culture and Society* 9: 206-227.

Maher, P. 1995. Probabilities for new theories. *Philosophical Studies* 77: 103-115.

Musgrave, A. 1988. The ultimate argument for scientific realism. In *Relativism and Realism in Science*, ed. R. Nola, 229–52. Dordrecht and Boston: Kluwer.

Norton, J. 2018. *The Material Theory of Induction*. Unpublished manuscript. Available at: < https://www.pitt.edu/~jdnorton/homepage/research/ind_material.html>.

Okruhlik, K. 1994. Gender and the biological sciences. *Canadian Journal of Philosophy* 24: 21-42.

Reichenbach, H. 1938. *Experience and Prediction*. Chicago: University of Chicago Press.

Reiss, J. and J. Sprenger. 2017. Scientific Objectivity. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Winter 2017 Edition)*. Available at <https://plato.stanford.edu/archives/win2017/entries/scientific-objectivity/>.

Rudner, R. 1953. The Scientist qua Scientist Makes Value Judgments. *Philosophy of Science* 20: 1-6.

Solomon, M. 2001. *Social Empiricism*. Cambridge, MA: MIT Press.

Strevens, M. 2017. Notes on Bayesian Confirmation Theory. Unpublished manuscript. Available at: <http://www.nyu.edu/classes/strevens/BCT/BCT.pdf>.

Stolley, P. D. 1991. When Genius Errs: R. A. Fischer and the Lung Cancer Controversy. *American Journal of Epidemiology* 133: 416-425.

Symons, D. 1979. The Evolution of Human Sexuality. New York: Oxford University Press.

Thagard, P. 1978. The best explanation: Criteria for theory choice. *Journal of*

Philosophy 75: 76-92.

Weisberg, J. 2009. Locating IBE in the Bayesian framework. *Synthese* 167: 125-143.

Wenmeckers, S., and Romeijn, J.W. 2016. New theory about old evidence. *Synthese* 193: 1225-1250.