Published as:
Lowe, Charles. "The Significance of Self-Fulfilling Science." *Philosophy of the Social Sciences* (online first): pp. 1-21.

http://journals.sagepub.com/doi/full/10.1177/0048393118767087

The Significance of Self-Fulfilling Science

Charles Lowe, Institute of Philosophy, Osnabrück University
lowech@uos.de

Abstract

Once lively debates concerning the philosophical significance of self-fulfilling science, or the causal contribution of science to bringing about the states of affairs it depicts, lapsed in the 1970s. Recent claims concerning the influence of economic theory on the behavior it predicts or explains seem poised to revitalize discussion, yet lack of clarity abounds concerning the key features of such cases and the philosophical issues to which they might be relevant. In this paper, I examine a paradigmatic case of self-fulfilling science, clarify its key features, and critically discuss two existing approaches to understanding such phenomena. Ultimately, I suggest a novel approach more well-suited to analyzing such cases and exploring their philosophical significance.

Keywords

self-fulfilling science, induced self-interest, performativity, reflexive prediction, values and science

1. Introduction

Storytellers and their audiences have long been fascinated by the idea that premonitions of future events may not be simply borne out but may themselves actually bring about the consequences they portend. Consider the tragedy of Oedipus, whose attempts to thwart the oracle's prophecy led, ironically, to its fulfillment in the form of patricide committed against his father, the king of Thebes. The key notion in this story, as in many similar ones from various times and cultures, is that actions undertaken in response to a claim that some future event would come to pass were, in fact, a crucial factor in the causal chain leading to that event's coming about. Following Merton's (1948) classic coining, such phenomena have come to be called *self-fulfilling prophecies*. Of course, for most, oracles and other traditional sources of divination no longer possess the same degree of epistemic authority they once seemed to have. This does not mean, however, that we have abandoned the idea of using knowledge about the likely course of future events to guide our actions in an attempt to avoid tragedy and obtain fortune. Where we once

turned to oracles and prophecies, we now consult scientific theories and predictions. Whether concerning the likely future development of our health, economic growth, or just about any other domain concerning matters of fact, up to and including the fate of the universe itself, science is now widely considered to be the best epistemic authority we have.

The great persuasive power of scientific pronouncements has led some to claim that scientific theories, models, or predictions may or even have themselves become self-fulfilling in a manner similar to that of the prophecy of Oedipus. From a literary perspective, this is certainly an intriguing suggestion. But what, if any, is the philosophical significance of what I shall term *self-fulfilling science*[1], specifically from the perspective of philosophy of science? Today, this question is not generally considered a serious subject of study and is likely to be viewed as a novelty at best and nonsense at worst. For a time, however, it seemed as though it might become a canonical, if minor, issue in mainstream philosophy of science.

At least two of the dominant figures in 20th century philosophy of science took such suggestions to be philosophically relevant. Karl Popper (1957, 13), in *The Poverty of Historicism*, considers whether "the Oedipus effect", or "the influence of the prediction on the predicted event (or, more generally, […] the influence of an item of information upon the situation to which the information refers)" might endanger the possibility of objective or precise scientific prediction about the social realm. Ernest Nagel (1961, 470), in *The Structure of Science*, claims that "the frequent occurrence of suicidal and self-fulfilling predictions concerning human affairs is undeniable", adding that "no theory adequate to the subject matter of the social sciences can ignore the fact that actions undertaken in the light of knowledge about some patterns of social behavior can often change those patterns." In addition to these isolated considerations, a spirited debate was carried out in the 1960s and '70s in the pages of *Philosophy of Science* concerning so-called *reflexive predictions*, which do not differ significantly from the phenomena discussed by Popper and Nagel except in name (Buck 1963; Romanos 1973; Vetterling 1976).

Despite this post-war burst of interest, however, philosophers of science have been almost entirely silent on the issue since the 1970s. Why is this so? In a recent paper, Kopec (2011) attributes this development, at least in the case of reflexive predictions, to an overly restrictive characterization of the phenomenon in question. Responding to flaws in the earlier account, he proposes a widened definition that not only allows consideration of a greater number of apparently relevant cases but also helps to highlight previously obscured methodological issues relevant to philosophy of science. Kopec's work shows that evaluating the significance of self-fulfilling science to philosophy of science is perhaps not as straightforward as it may initially

[1] Others have typically referred to self-fulfilling *predictions, theories*, or some other specific type of representation. I employ the term self-fulfilling *science* to capture the idea that many different types of scientific representation, broadly understood, may be involved in the phenomenon in question.

seem. To address the issue, we need answers to the following questions: How should we conceive of self-fulfilling science; that is, in virtue of which key feature or features should some instances of science be considered self-fulfilling? Do we have reason to believe that the phenomenon may actually exist or be prevalent enough to warrant scrutiny? And finally, (how) would the existence of science that displays these key features either generate new philosophical problems or contribute something novel to preexisting ones?

This paper contributes to reemerging discussion concerning self-fulfilling science and its significance by examining these questions in relation to a specific phenomenon, *induced self-interest*, which I introduce in the following section. Restricting focus to just one example is necessary given that evaluating the prevalence of self-fulfilling science more generally would require a level of engagement with data from multiple disciplines that goes far beyond the scope of this paper. The choice to focus on induced self-interest, in particular, is motived by the fact that this phenomenon has recently come to be viewed by many social scientists as a particularly strong, clear, and thus paradigmatic case of self-fulfilling science. I begin by presenting a schematic overview of the empirical evidence for induced self-interest and an influential analysis thereof by Ferraro, Pfeffer, and Sutton (2005). I then argue that two recently suggested characterizations of the phenomenon at the heart of induced self-interest, *reflexive prediction* and *performativity*, are ill-suited to the task of gauging the potential philosophical significance of such cases, because they misidentify and direct attention away from the key features. In doing so, the latter, more prominent, perspective in particular has led commentators to focus on the supposed relevance of self-fulfilling science to debates concerning scientific realism. In light of recent arguments to the effect that this purported relevance is largely illusory as well as the failings of the previous characterizations more generally, I ultimately suggest a novel approach that both better captures the key features of the paradigmatic case and encourages consideration of the significance of such cases to other topics of interest to philosophers of science.

2. A Paradigmatic Case

In 1981, experimental psychologists Marwell and Ames set out to test 'the free rider hypothesis' as a predictor of behavior in collective action situations. In a series of experiments, they measured participants' willingness to contribute their own resources to the provision of a public good against the prediction, derived from a strong version of the free rider hypothesis, that absolutely no resources would be contributed. The results were clear:

> over and over again, in replication after replication, regardless of changes in a score of situational variables or subject characteristics, the strong version of the free rider hypothesis is contradicted by the evidence. People voluntarily contribute substantial

> portions of their resources - usually an average of between 40 and 60 percent - to the provision of a public good. (Marwell and Ames 1981, 307)

However, one of the variations in subject characteristics produced a surprising result: economics graduate students were much more likely to free ride than any other group tested, contributing on average only about 20 percent of their resources.

The finding that those with some formal training in economics are more likely than peers to engage in free riding or various other types of 'self-interested' behavior has since been replicated numerous times (e.g. Carter and Irons 1991; Frank, Gilovich, and Regan 1993; Frank and Schulze 2000; Baumann and Ross 2011). How best to account for this correlation, however, is still being debated. Some studies suggest a selection effect: those with a stronger tendency to act in a self-interested manner are also more likely to choose to study economics (Carter and Irons 1991; Frank and Schulze 2000). Others suggest what has been called a 'learning' or 'indoctrination' effect: willingness to engage in self-interested behavior is caused by or at least significantly strengthened by undergoing economic training (Frank, Gilovich, and Regan 1993; Baumann and Rose 2011). The suggestive evidence for such 'learning effects' has been taken by some social scientists to indicate that economics may be a self-fulfilling science.

Organizational scholars Ferraro, Pfeffer, and Sutton (2005) have been some of the most vocal proponents of this view. They examine such cases as part of their argument for the broader claim that "social science theories can become self-fulfilling by shaping institutional designs and management practices, as well as social norms and expectations about behavior, thereby creating the behavior they predict" (8). By identifying specific mechanisms by which theories may become self-fulfilling, the authors hope to 'operationalize' more general claims in the literature concerning self-fulfilling science. In particular, they argue that economics may lead to increased tendencies toward self-interested behavior through not only direct formal training, but other, subtler mechanisms as well. In order to facilitate discussion, I use the term *induced self-interest* to refer to the phenomenon at the heart of such cases. The authors interpret the above studies and others suggesting that economic training encourages belief in the appropriateness of self-interested behavior (Miller 1999) to argue that economics may cause induced self-interest, and thus become self-fulfilling, by establishing social norms and shaping institutional design:

> many of the experimental results of the tendency of economics students and economists to defect more, cooperate less, and, in general, behave more in accordance with the dictates of self-interest may be mediated by belief in the norm of self-interest and its prevalence. No tests of mediation in any of these studies are reported, but the argument and empirical implication are straightforward: one effect of economics training is to strengthen beliefs in the pervasiveness, appropriateness,

> and desirability of self-interested behavior, which, in turn, should lead to exhibiting more self-interested behavior. (Ferraro et al. 2005, 14)

Ferraro *et al*.'s claims are relevant to our purposes for a number of reasons. First, theirs is a widely influential[2] account of what has been called "the archetypical example of self-fulfillment" (Bergenholtz and Busch 2016). Second, they go farther than many other enthusiasts in presenting empirical evidence at least suggesting the actual existence of the phenomenon. Finally, several issues they see as arising from the existence of self-fulfilling science bear on well-established or emerging debates within philosophy of science. For these reasons, I treat the phenomenon of induced self-interest as described by Ferraro *et al.* and their account thereof as a paradigmatic case to which the question of the philosophical significance of self-fulfilling science may be put.

As noted above, our evaluation of the significance of self-fulfilling science depends upon and will be shaped by our conception of the phenomenon itself, or of the key features of some case in virtue of which we take it to be an example thereof. Despite their enthusiasm for 'self-fulfilling theories', however, Ferraro *et al.* provide no explicit definition or characterization of the phenomenon. They do favorably cite Merton's (1948, 195) famous characterization of the self-fulfilling prophecy as "a *false* definition of the situation evoking a new behavior which makes the originally false conception come *true.*" However, this serves merely as a starting point for their discussion and no indication is given that it is intended as a definition. The closest they come to making a clear statement is to refer readers to debates about *reflexive predictions* in philosophy of science and *performativity* in science studies. This is unsurprising, given that these are two of the only explicit notions available in the literature that seem, at first glance, to apply to the cases they discuss.[3] It may thus seem reasonable to assume that either one or both are capable of identifying the key features we are looking for. As I argue in the following section, however, both are ill-suited to the purpose.

3. Two Approaches to Understanding Induced Self-Interest

3.1. Reflexive Prediction

Mertonian self-fulfilling prophecy also served as the starting point for debate concerning the nature and significance of so-called *reflexive predictions* among philosophers of science during

[2] This claim is based on a very informal methodology: As of December 2017, Google Scholar reports over 1000 citations, and a random sampling of approximately 50 of these showed that the majority of authors citing the paper reported its conclusions without critical discussion.

[3] Ian Hacking's (1995) well-known notion of *looping effects* might be considered another candidate. However, one of Hacking's main contentions is that the effects of scientific categorization on objects of study vary so widely from case to case that there is 'no general story' to be told about the phenomenon. To the extent that accounts of self-fulfilling science relate to Hacking's notion, they must be viewed as an attempt to tell a kind of general story concerning one specific type of looping effect.

the 1960s and '70s. Buck (1963) introduced the topic by explicitly citing Merton's formulation, but settled on a more complex definition, around which further discussion revolved. Until recently, the consensus was that Romanos (1973, 106) supplied the definitive account, according to which a prediction is reflexive if and only if its so-called "formulation/dissemination style" is "a causal factor relative to the prediction's coming out true or false." Kopec (2011) has recently attempted to revive this debate by criticizing Romanos' definition and suggesting a revision. Despite their differences, however, none of the proposed definitions stray far from Merton's original notion that a change in the literal truth or falsity of a prediction is an essential part of self-fulfillment. It is this feature, I argue, that makes the notion of reflexive prediction inadequate to capture what is at stake in the paradigmatic case.

To see why, let us first turn to Kopec's critique of Romanos' definition. He begins by analyzing the potentially ambiguous notion of a prediction's dissemination 'being a causal factor relative to the prediction's coming out true or false'. Kopec claims this is most plausibly read as stating that a prediction is reflexive if and only if its dissemination in a certain manner is *sufficient* to make it to come out true when it otherwise, i.e. in absence of dissemination, would have come out false.[4] He then points out that this sufficiency condition rules out any prediction for which the actual outcome depends, even minimally, upon chance. "This definition", he says, "is therefore unlikely to apply to any predictions made in the social sciences", before concluding that "it is no wonder that the excitement over such predictions ended back in the 1970s" (Kopec 2011, 1253).

In response, he suggests that the *strongly reflexive predictions* picked out by Romanos' definition are a subset of the class of *weakly reflexive predictions*, defined as those whose "mode of dissemination is sufficient to *change the probability* of the predicted event occurring from what it would be if not disseminated" (2011, 1253, emphasis added). Kopec claims several advantages for his definition. First, it captures a broader range of apparently relevant cases. Second, it allows us to distinguish between reflexive predictions of varying strength within this range. While both count as weakly reflexive, predictions that make the occurrence of the predicted event significantly more probable are stronger than those that make it only marginally so. Finally, by shifting focus to probability, Kopec's definition highlights certain methodological issues that escaped his predecessors. In particular, such cases might cause serious problems for those who employ Bayesian or likelihoodist confirmation-theoretic frameworks, given that dissemination of reflexive predictions can change the evidential import of observing their obtainment. Kopec's account thus serves as a good example of how clarifying the key features of cases we intuitively view as self-fulfilling science is an essential step in evaluating their philosophical significance.

[4] Although reflexive prediction is usually defined to include both self-fulfilling and self-defeating predictions, I discuss only the former here.

However, although the evidential issues he raises might also apply to the paradigmatic case, there are at least two ways in which both his and Romanos' definitions of reflexive predictions fail to capture its key features.

First, in stark contrast to the view of many social scientists that induced self-interest is a particularly strong and clear, and thus paradigmatic, example of self-fulfilling science, these definitions either fail to characterize such cases as self-fulfilling science at all or recognize them only as an exceedingly weak form thereof. To see this, consider again the experiments of Marwell and Ames. For now, let us imagine that we are dealing with a single experiment and a one-shot prediction rather than the decidedly messier situation we actually face. A strong version of the free rider hypothesis predicts that participants will contribute no portion (0%) of their resources to the provision of the public good. This prediction is communicated to only one subset of participants. Those informed of the prediction contribute on average only 20% of their resources, while the others contribute around 50%. Evidence that this discrepancy is most reasonably attributable to dissemination of the prediction rather than to some other factor is interpreted as suggesting the prediction was in some sense self-fulfilling. But in what sense? Clearly, the prediction was not strongly reflexive, as it did not come out true. But was it weakly reflexive, that is, did the prediction's dissemination make it more likely that it would come out true than would otherwise have been the case?

If the predicted outcome of 0% total contributions were impossible for any reason, then we must answer no, because we could not then meaningfully speak of an increase in the probability of such an occurrence. This would be the case, for example, if there existed some yet-unidentified psychological law that caused each participant to contribute at least a single resource. Of course, we have no reason to believe such a law actually exists, nor that anything else stands in the way of the in-principle possibility of the predicted occurrence. Indeed, taking the strong free rider hypothesis seriously requires considering this to be a possible, however unlikely, outcome. This being the case, it can be reasonably claimed that the reduction of contributions caused by dissemination in our example increased the probability of the predicted outcome and, thus, that the prediction was indeed weakly reflexive.

However, the degree to which probability is increased in such a case is marginal at best. Although we can't exclude the possibility in principle, we feel quite justified in assigning a vanishingly small probability to an outcome in which at least those in the uninformed group will fail to contribute any resources. Such a result would fly in the face both of the actual results of Marwell and Ames' experiments and our experiences with human subjects more generally. Because the probability of the predicted outcome of 0% total contributions depends strongly on the probability that the uninformed group will contribute nothing, the fact that the latter is vanishingly small means the former is as well. This is the case no matter how strongly

dissemination influences the behavior of the informed group. Even if the effect were so strong as to *guarantee* a complete lack of contributions from the informed group, the probability of the predicted event itself would remain vanishingly small, even if marginally increased by dissemination. Because Kopec's notion measures the strength of reflexive predictions according to the degree of increased probability their dissemination causes, it treats cases such as the above as exemplifying only an exceedingly weak form of self-fulfilling science. This result conflicts directly with the statements and intuitions of social scientists who view the kinds of effects observed in Marwell and Ames' experiments as a particularly strong form of induced self-interest and a paradigmatic example of self-fulfilling science. Thus, although it fairs better than strongly reflexive prediction by encompassing the paradigmatic case, Kopec's notion of weakly reflexive prediction fails to adequately gauge the strength of the effect.

The second way in which such accounts fail becomes apparent when we more carefully consider what entity or representation it is, exactly, that is supposedly self-fulfilling in the paradigmatic case. Arguably, the simplified re-description discussed above distorts things: presumably, a one-shot prediction derived from the strong version of the free rider hypothesis is not what is supposedly self-fulfilling in the paradigmatic case. Ferraro *et al.*'s general claim is that social scientific *theories* may become self-fulfilling by influencing institutional designs, social norms, and expectations. In describing induced self-interest in particular, they state that "the core economic *assumption* of self-interest is a *prediction* about how people will behave, but also serves as a norm that regulates behavior" (2005, 14, emphasis added). It seems there are at least three kinds of representation one might claim to be self-fulfilling in this case: a theory, an assumption, or a prediction. Even if we settle upon 'prediction', however, this cannot be understood as meaning a one-shot prediction concerning some particular future event. If it is to act as a core economic assumption, one that "forms the foundation for other fundamental premises in economics" (2005, 11), then this 'prediction' must be understood rather as a general description of behavioral patterns and/or the motivations that account for them.[5]

So, what is the relevant representation in the paradigmatic case? The authors make no explicit reference to any *specific* theory that is supposedly self-fulfilling, but rather refer generically to the effects of "economic theory". Nor do they further specify *the* assumption or general description of self-interested behavior they take to be at the core of economic theory. In fact, it is difficult to identify any single specific theory or assumption to which they might be referring, given that the economic training claimed to induce self-interested behavior consists of various

[5] To apply the notion of reflexive prediction to this case we must grant that it can be adapted to types of representation other than one-shot predictions. As I do not find this suggestion to be particularly problematic and wish to focus on a different reason why the account is ill-suited in this case, I will not address this issue further.

theories, models, and assumptions of differing content. What's more, the 'assumption' of self-interest employed in such models and theories is often understood to be an idealization or even an explicitly prescriptive norm that may not even be truth-apt in the first place.

In order to apply the notion of reflexive prediction to the paradigmatic case, we must identify a specific scientific representation the dissemination of which is sufficient to bring about (strongly reflexive) or make more probable (weakly reflexive) its truth in a strict and literal sense. This presupposes that the representation in question be both truth-apt and have a clearly delineable content. However, Ferraro *et al*. neither give us any indication as to what this might be, nor does such explicit identification even seem to be required to adequately capture the spirit of their claims regarding induced self-interest.

The problems outlined above show that the Mertonian-derived, truth-centric notion of reflexive prediction fails to capture several key features of induced self-interest. To the extent that this case is paradigmatic of self-fulfilling science, reflexive prediction is ill-suited as a general account of the phenomenon.[6] It seems we need a more flexible notion, one that does not require us to make a claim concerning effects on the actual or probable truth of a single, clearly delineable representation. In fact, the second notion mentioned by Ferraro *et al*., so-called *performativity*, seems to meet this criterion. Despite this, however, I argue that adopting this perspective comes at the cost of inviting a host of problematic assumptions and associations that distract from rather than clarify the key features of the paradigmatic case.

3.2. Performativity

Performativity is a term, thesis, or approach used by sociologists of economics and other practitioners of science studies to describe ways in which (economic) theory shapes, constitutes, enacts, or 'performs' (economic) reality. The term itself is derived from J. L. Austin's notion of linguistic performatives, utterances that 'do something' rather than simply describe or report something about the world. The utterances "I apologize for…" or "I promise that…", for example, do not simply state or describe some preexisting state of affairs, but are rather constitutive of the act of apologizing or promising when issued in the proper contexts. By analogy, economic theories, models, and assumptions are said to sometimes be performative in the sense that they do not simply describe a preexisting economic reality, but are rather themselves part of what brings about that reality. Performativity researchers are thus generally engaged in detailed historical and sociological research on various developments in economic theory and reality with an eye to how to the former might help bring about the latter.

[6] Of course, this does not preclude the possibility that it may be apt for describing a specific type of the general phenomenon of self-fulfilling science or other closely related phenomena. Nor does it bear directly on the possible philosophical relevance of the problems concerning evidence evaluation highlighted in Kopec's account.

Despite this empirical work and the growing popularity of the approach in recent years, the central concept of *performativity* itself has remained relatively unclear and contested. Of those who have offered explicit characterizations, MacKenzie (2006) has provided one of the clearest and most widely cited. For this reason, as well as for the fact that Ferraro *et al*. cite MacKenzie's work explicitly, I restrict my focus to his account here. At the most general level, says MacKenzie, the *performativity thesis* states that "the academic discipline of economics does not always stand outside the economy, analyzing it as an external thing; sometimes it is an intrinsic part of economic processes" (2006, 16). More specifically, he distinguishes between three types:

> "Generic" performativity: An aspect of economics (a theory, model, concept, procedure, data set, etc.) is used by participants in economic processes, regulators, etc.
>
> "Effective" performativity: The practical use of an aspect of economics has an effect on economic processes.
>
> "Barnesian" performativity: Practical use of an aspect of economics makes economic processes more like their depiction by economics.

It is the last of these that interests us. MacKenzie (2006, 19-20) notes that one might read "Barnesian"[7] performativity as simply another term for Mertonian self-fulfilling prophecy, but offers several reasons for "preferring the terminology" he uses. Interestingly, he does not seem to realize that there is a substantial, rather than merely terminological, difference between the two. While self-fulfilling prophecy involves the bringing about of a state of affairs that makes what would otherwise have been a false prediction come out true, "Barnesian" performativity requires only increasing similarity between "processes" and "their depiction" via "practical use" of an aspect of science.

If we are willing to characterize the behaviors witnessed in cases of induced self-interest as "economic processes", the theories, models, assumptions, and other representations and/or idealizations that make up the economic training in question as "depictions" thereof, and this training itself as a "practical use" of economics, then "Barnesian" performativity seems to meet the criterion developed in the previous section. This formulation seemingly does not require us, as in the case of reflexive predictions, to identify a single truth-apt representation with a clearly delineable content whose truth must be either brought about or made more likely in order to describe such a case as self-fulfilling. It is enough to say that the widespread practical use of

[7] The term "Barnesian" is derived from the surname of sociologist Barry Barnes, whose work on the role of self-validating or 'bootstrapped' inference in the constitution of stable social institutions has served as inspiration to many performativity researchers.

economic depictions of self-interested behavior in economic training has, on the whole, made some actual economic behavioral patterns *more like* the depictions themselves. Also, the notion of 'practical use' seems to be an improvement over 'dissemination' in capturing Ferraro *et al*.'s mechanisms of institutional design and norm establishment. However, although "Barnesian" performativity seems to avoid these specific failings of reflexive prediction, there are other good reasons to rethink the recent trend of framing debates concerning self-fulfilling science and its possible philosophical significance primarily in such terms.

If the performativity approach provides a notion that initially seems apt to describe the paradigmatic case, it does so at cost of introducing a good deal of conceptual ambiguity and distracting theoretical baggage. One issue is that, although the notion intuitively seems a good fit, key terms it employs, like *depiction* and *making processes more like* their depictions, are not sufficiently spelled out. Perhaps the most problematic conceptual issue, however, is a failure among many performativity scholars to clearly distinguish between causal and constitutive understandings of their claims. Mäki (2013) argues that by co-opting the language of performativity from Austin, these authors imply that the relationship between economic theory and reality is a constitutive one. Just as uttering "I promise to…" does not simply cause a promise to come about, but rather establishes its existence constitutively, a constitutive relationship in economics "would require that uttering or writing down an economic model for an audience (that understands the model and perceives the uttering as genuine and done in appropriate circumstances) establishes the model world as part of the real world" (447). Because the connection between economic theory and reality envisioned by performativity scholars requires "practical use" that goes far beyond simply uttering, Mäki argues, their claims must in fact be understood to refer to causal rather than constitutive processes. Despite this, the performativity thesis is often conceived of as concerning the constitution of economic reality rather than causal effects within that reality. This, in turn, has led many to view it as being or entailing a strong form of scientific antirealism about economics and the objects it depicts.

The perceived association between performativity and antirealism has tended to direct what little philosophical attention has been paid to the paradigmatic case towards its possible significance to debates concerning scientific realism. The following section shows this by examining a dispute between Ferraro *et al*. and Felin and Foss (2009a; 2009b), and subsequent commentary by Bergenholtz and Busch (2016), about the potential consequences self-fulfilling science. Unfortunately, I argue, this narrow focus has resulted in the overshadowing of other legitimate, and possibly more fruitful, perspectives.

4. The Significance of Self-Fulfilling Science

Because the primary aim of Ferraro *et al*.'s (2005) original paper was to 'operationalize' more general claims about self-fulfilling theories by identifying specific mechanisms, their initial discussion of the implications of their findings was quite limited. Critics Felin and Foss (2009a), on the other hand, in addition to calling many of Ferraro *et al*.'s findings themselves into question, also stress the potentially dire consequences of claims concerning self-fulfilling science. As they see it, "the strong forms of the self-fulfilling nature of theories and language are sobering because, if true, they threaten the fundamental definition of science and theory as an attempt to understand and predict objective reality" (655).

This critical response led Ferraro *et al*. (2009, 673-674) to reflect further on the theoretical and practical consequences they see as arising from their descriptive claims. They highlight three issues in particular. First, they suggest that self-fulfilling science generates a special kind of responsibility for researchers in disciplines where it may occur. This goes beyond what is advocated by standard research and business ethics approaches by requiring researchers to consider the "ethical consequences of theory"; it "focuses on the ethical and moral consequences of what we teach and how we do our research." Second, they identify possible consequences for the evaluation of evidence. Testing potentially self-fulfilling theories requires "more subtlety and more attention to the mechanisms that may make them appear true even if they are not." Finally, they consider that, if multiple potentially self-fulfilling theories contend with one another, "multiple futures and realities are possible". From this, they draw the conclusion that, in such cases, "we have the opportunity to both envision and create a different and maybe even better, more humane, and just world." Summing up, Ferraro *et al*. identify three avenues for further research concerning the possible theoretical and practical consequences of self-fulfilling science: issues of *moral responsibility, evidence evaluation,* and *multiple possible futures*.

Each of these issues can be seen as relating to established or emerging topics of discussion in philosophy of science. Douglas (2010) argues that consideration of the moral responsibility of scientists was once considered a proper and natural task for philosophy of science, and urges a return to a more socially engaged understanding of the field. Almost mirroring Ferraro *et al*.'s suggestion, she even places special emphasis on the need to go beyond the standard approaches of research ethics by considering the consequences of scientists' theories and actions in order to achieve a full mapping of the 'moral terrain of science' (Douglas 2014). Evidence evaluation has, of course, long been a canonical topic in the field. As already noted, Kopec and earlier discussants of reflexive prediction have examined issues of this sort specific to self-fulfilling science. Finally, the notion that scientists might help bring about a better world by choosing to promote potentially self-fulfilling theories with preferable outcomes can be related to recent discussions

concerning the balancing of epistemic and non-epistemic interests and values in science (cf. Elliott and McKaughan 2014).

Despite the variety of issues suggested by Ferraro *et al*., however, commentators have primarily focused on the perceived threat of performativity-based accounts of self-fulfilling theories to scientific realism. This can be seen, first of all, in the way Felin and Foss (2009b) filter their responses to Ferraro *et al*.'s suggestions through the specter of the antirealism they see as inherent to the performativity approach. Based on their reading of several classics in the field, they claim that "according to the performativity perspective, then, we cannot even meaningfully speak of the ex ante 'truth' or 'reality' of theories, because theories themselves participate in defining and creating what is truthful and what is real" (Felin and Foss 2009b, 676). But if this is true, what sense does it make to speak of testing such theories in an attempt to assess their truth or falsity? "If not truth," they ask, "what then should be the basis for choosing one theory over another?" By denying the notion of 'ex ante truth', they say, performativity implies a radical reinterpretation and repurposing of science itself from advancing knowledge about the world to attempting to bring about "the best of all possible worlds". In a final blow, they charge Ferraro *et al*. with (possibly unintentionally) taking a strong stand on the side of social constructionists in the so-called 'science wars' through their adoption of the performativity account.

This tendency to focus on the possible significance of induced self-interest to issues of scientific realism can also be observed in Bergenholtz and Busch (2016), the only extended treatment of the debate between Ferraro *et al*. and Felin and Foss published in a dedicated mainstream philosophy of science journal of which I am aware. Bergenholtz and Busch's stated aim is to "cool the fire" in this debate by examining two "threats" that self-fulfilling theories may pose to social science: the challenge to realism and the suggestion of a special kind of moral responsibility for individual scientists. Echoing Felin and Foss, they begin by stating that "the overall terminology and structure of this debate is, in part, derived from the classical debate between (social) constructionism at one end of the spectrum and realism on the other" (Bergenholtz and Busch 2016, 25). After introducing the notion of performativity and briefly laying out the paradigmatic case, they then characterize what they take to be the crux of the debate thusly:

> This example highlights the main issue at stake; not only will the behavior of individuals change to some degree due to the adoption of a scientific theory, but some underlying empirical regularities (Felin and Foss 2009a) will also change, which might turn the theories from false to true. Self-fulfillment is, thus, threatening to undermine objectivity in general and a fundamental premise of scientific realism in particular: that there are some sort of theoretical mechanisms […] and regularities

> out there in the world not directly affected by our theorizing about them. (Bergenholtz and Busch 2016, 29)[8]

The bulk of their paper is then devoted to developing a number of arguments meant to show that this purported threat to realism is largely illusory. In comparison, their treatment of the threat of a special moral responsibility for individual researchers comes across as an afterthought, both in terms of dedicated space and argumentative rigor. Clearly, Bergenholtz and Busch's attention is focused firmly on dismantling the threat to realism rather than examining the issues suggested by Ferraro *et al*.

Bergenholtz and Busch (2016) are not alone in thinking this purported threat has more bark than bite. Mäki (2012) argues that the causal effects described by performativity scholars do not conflict with any reasonable account of realism concerning economics or the social sciences more generally:

> It is no threat to scientific realism about economics to acknowledge the possibility of causal economics-dependence of some items in the real-world economy. After all, economics as an academic discipline is itself social activity exercised within society, so such connections are a natural feature of social reality. Good social science will investigate such connections together with other causal connections in society at large. (Mäki 2012, 21)

Evaluating the purported threat to realism and the arguments against it ultimately goes beyond the scope of this paper. My argument against framing discussion of induced self-interest and related cases in terms of performativity rests instead upon the claim that the conceptual ambiguities and theoretical baggage of this approach conspire to encourage an overly narrow conception of the potential philosophical significance of self-fulfilling science. If Mäki and Bergenholtz and Busch are correct that the threat to realism is illusory, then the worries of Felin and Foss appear misguided, and discussion thereof seems to offer little of interest to philosophers of science. Even if their arguments prove to be flawed, however, the fact that the perceived association between performativity and antirealism has obscured and diverted attention away from other promising avenues of research, such as those suggested by Ferraro *et al*., already gives us a very good reason to consider alternatives. In the following section, I develop a characterization of self-fulfilling science that both better captures the key features of the

[8] The final statement in this quote is actually much stronger than what those claiming the existence of self-fulfilling science are committed to. The existence of phenomena affected by our theorizing about them in no way entails that there are no phenomena not thusly affected, nor do Ferraro *et al*. claim anything of the sort.

paradigmatic case than the two notions just discussed and clears the way for consideration of other avenues of further research.

5. Self-fulfilling Science as Increased Conformation

Although they turned out to be ill-suited to our task, we can still learn a few things from the two approaches just discussed, especially when their relative strengths and weakness are contrasted. Reflexive prediction excels by making explicit the counterfactual form of claims concerning self-fulfilling science. In the cases that intuitively interest us, we feel that, had 'dissemination' of a scientific representation not taken place, then some state of affairs standing in the right kind of relation to this representation would have failed to obtain. The account fails to develop an adequate criterion by which we are to gauge the presence or strength of self-fulfilling science by tying the relevant kind of relation too closely to the truth of a clearly delineable, necessarily truth-apt representation. "Barnesian" performativity avoids this issue by requiring only that some part of reality be made more like its 'depiction' by science through 'practical use' of some aspect of that science. Identifying the relevant kind of relation involved in terms of increased similarity seems to capture the key features of the paradigmatic case in a more intuitive manner than reflexive prediction. However, it does so at the cost of failing to further elucidate the key terms involved and encouraging an overly narrow conception of the possible philosophical relevance of such cases.

We need an account that builds on the respective strengths of these previous attempts by spelling out in more detail the nature of the various components involved without thereby leading us into the problems outlined above. To achieve this goal, it will be useful to draw on a further concept from the philosophy of science literature that, while developed for other purposes, can help get a grasp on the key relation we seek to understand. This is Longino's (2002) notion of *conformation*, which she introduces as a "general term for epistemological success of content" that encompasses other, more specific, senses such as truth and similarity, among others:

> I am proposing to treat conformation as a general term for a family of epistemological success concepts including truth, but also isomorphism, homomorphism, similarity, fit, alignment, and other such notions. Classical truth is a limiting concept in a category of evaluation that in general admits of degree and requires the specification of respects. Truth is where degree and respects fall away. This approach avoids the crudity of a binary evaluation, and hence avoids one of the problems attributed to true or false. (Longino 2002, 117)

The concept also has a number of other features clearly relevant to our previous attempts to characterize the key features of the paradigmatic case:

> It can apply to laws and to statistical claims that are not literally true, but that capture the relations in which we are interested. […] Idealizations like laws are not, strictly speaking, true because there is not a particular situation that they accurately and precisely represent, but they conform to the range of phenomena over which they are idealizations in the way a map conforms to its terrain. […] Conformation is also more suitable than true or false for expressing the ways in which complex content, such as a theory or model, is successful representation. […] We often want to say of a whole complex–for example, the theory of optics, the theory of special relativity, or the synthetic theory of evolution–that it constitutes knowledge. It conforms, even though its components conform in different respects and to different degrees. Its components are not all, strictly speaking, true, but as long as the whole conforms to its object in this sense, it constitutes knowledge. (Longino 2002, 115-118)

Employing this notion to make sense of the insights gleaned from our discussions of reflexive predictions and performativity, I suggest the following characterization of Ferraro *et al*.'s general claim concerning the paradigmatic case:

Induced self-interest is a case of self-fulfilling science in virtue of the fact that the degree of conformation between the relevant complex of economic representations and the entities to which it relates is greater than it would have been, had not some practical use of the former contributed causally to changes in the latter. The notions of 'representations' and their 'relation' to entities, as well as the 'degree of conformation' between them, should be understood in the broad and disjunctive senses developed in Longino's quotes above. The representations in question can vary in kind and specificity, including general theoretical statements, precise predictions, non-truth-apt idealizations, and others. The kinds of relation in which these representations may stand to various entities in the world vary between truth, similarity, fit, and others. Finally, an increased degree of conformation may exist with regards to the relevant complex of economic representations even when the kinds and degrees of conformation with regards to more specific representations within this complex vary considerably. All that is required by the claim, understood in the general sense of Ferraro *et al*., is an increase of conformation *on the whole* caused by changes induced in some entities through practical use of the relevant representations.

Understanding self-fulfilling science in terms of increased conformation also makes better sense of more specific claims, such as those concerning the effect observed in Marwell and Ames' experiment. The strong free rider prediction discussed in my simplified re-description of the experiment comes out as self-fulfilling because the outcome caused by its dissemination, in which a significant subset of participants contributed significantly fewer resources, clearly conforms to a greater degree with the predicted outcome of 0% total contributions than any plausible

outcome in absence of dissemination. This holds despite the fact that dissemination was neither sufficient to guarantee the predicted outcome nor to make its actual occurrence significantly more probable. In fact, because this account measures the strength of such effects according to the relative increase in conformation they cause rather than the degree to which they make the actual occurrence of the predicted event more probable, its assessment aligns with many social scientists' in viewing the observed effect a strong case of self-fulfillment. Finally, by being more explicit about the various types of representation and "epistemological success of content" that may be involved in what the performativity approach simply refers to as entities becoming 'more like' their 'depictions', the conformation-based approach offers resources that can help make sense of claims of self-fulfilling science in messier cases, in which it is difficult to identify which *specific* representations are to be considered relevant in the first place.

Ultimately, the above characterization leaves too many questions unanswered to serve as a complete account of self-fulfilling science. Some questions relate to theoretical issues. Clearly, Longino's account of conformation and my use of it require further elaboration.[9] We must also ask what exactly constitutes 'practical use', or if there are any restrictions on the kind of causal influences we are willing to consider as contributing to such effects. Other questions relate to issues of application. For example, is there anything to be said in general about how we should go about determining which representations are relevant in an apparent case of self-fulfilling science, about how to 'measure' and 'weigh' degrees of conformation, or how much of an increase is required before we can legitimately speak of self-fulfillment?

Despite the need for further work, however, the arguments presented in this paper show that the conformation-based approach fares better than either reflexive prediction or performativity both in capturing the key features of the paradigmatic case and encouraging reflection on previously ignored facets of such cases and their potential significance to philosophical issues other than scientific realism. To mention just one example, examining the apparent relevance of Ferraro *et al.*'s claims concerning *moral responsibility* and *multiple possible futures* to issues currently being discussed by scholars interested in the role of non-epistemic values and goals in science (e.g. Douglas 2014; Elliott and McKaughan 2014) seems to offer an especially promising opportunity for further research of this kind. Whether or not this particular avenue ultimately proves fruitful, it is my hope that the account developed here will contribute more generally to

[9] In fact, it may be objected that Longino's account introduces no less and no less problematic or distracting theoretical baggage than performativity approaches. My impression is that the associations evoked by the approach may actually help highlight legitimate issues, such as those related to the topic of values and science. Discussion of this claim, however, goes beyond the scope of this paper. For now, it is enough to note that the account developed here does not depend essentially on the notion of conformation or Longino's account thereof. Those who take them to be more trouble than they are worth are free to substitute less theoretically charged alternatives that do more or less the same work.

an enhanced understanding of and increased engagement with the reemerging and fascinating topic of self-fulfilling science.

Acknowledgements

My thanks go out to participants at the 6th conference of the European Network for the Philosophy of the Social Sciences (Cracow University of Economics, Sept. 20-22, 2017), attendees of the most recent MüBiOs meeting at Osnabrück University (Feb. 2, 2018), and two anonymous reviewers for helpful comments on previous versions of this work.

References

Bauman, Yoram, and Elaina Rose. 2011. "Selection or Indoctrination: Why Do Economics Students Donate Less than the Rest?" *Journal of Economic Behavior & Organization* 79: 318-327.

Bergenholtz, Carsten, and Jacob Busch. 2016. "Self-Fulfillment of Social Science Theories: Cooling the Fire." *Philosophy of the Social Sciences*. 46 (1): 24-43.

Buck, Roger C. 1963. "Reflexive Predictions." *Philosophy of Science* 30: 359-69.

Carter, John R., and Michael D. Irons. 1991. "Are Economists Different, and If So, Why?" *The Journal of Economic Perspectives* 5 (2): 171-177.

Douglas, Heather. 2010. "Engagement for Progress: Applied Philosophy of Science in Context." *Synthese* 177: 317-335.

Douglas, Heather. 2014. "The Moral Terrain of Science." *Erkenntnis* 79: 961-979.

Elliott, Kevin C., and Daniel J. McKaughan. 2014. "Nonepistemic Values and the Multiple Goals of Science." *Philosophy of Science* 81: 1-21.

Felin, Teppo, and Nicolai J. Foss. 2009a. "Performativity of Theory, Arbitrary Conventions, and Possible Worlds: A Reality Check." Organization Science 20 (3): 676-678.

Felin, Teppo, and Nicolai J. Foss. 2009b. "Social Reality, the Boundaries of Self-Fulfilling Prophecy, and Economics." *Organization Science* 20 (3): 654-668.

Ferraro, Fabrizio, Jeffrey Pfeffer, and Robert I. Sutton. 2005. "Economics Language and Assumptions: How Theories Can Become Self-Fulfilling." *Academy of Management Review* 30 (1): 8-24.

Ferraro, Fabrizio, Jeffrey Pfeffer, and Robert I. Sutton. 2009. "How and Why Theories Matter: A Comment on Felin and Foss (2009)." *Organization Science* 20 (3): 669-675.

Frank, Björn, and Günther G. Schulze. 2000. "Does Economics Make Citizens Corrupt?" *Journal of Economic Behavior and Organization*, 43: 101-113.

Frank, Robert H., Thomas Gilovich, and Dennis T. Regan. 1993. "Does Studying Economics Inhibit Cooperation?" *The Journal of Economic Perspectives* 7 (2): 159-171.

Hacking, Ian. 1995. "The Looping Effects of Human Kinds." in *Causal Cognition: A Multidisciplinary Debate*, edited by Dan Sperber, Davis Premack, and Ann James Premack, 351-383. Oxford University Press.

Kopec, Matthew. 2011. "A More Fulfilling (and Frustrating) Take on Reflexive Predictions." *Philosophy of Science* 78: 1249-1259.

Longino, Helen E. 2002. *The Fate of Knowledge*. Princeton University Press.

MacKenzie, Donald. 2006. *An Engine, Not a Camera: How Financial Models Shape Markets*. Cambridge, MA: MIT Press.

Mäki, Uskali. 2012. "Realism and Antirealism About Economics." In *Handbook of the Philosophy of Science. Volume 13: Philosophy of Economics*, edited by Uskali Mäki, 3-24. Elsevier.

Mäki, Uskali. 2013. "Performativity: Saving Austin from MacKenzie." In *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, edited by Vassilios Karakostas and Dennis Dieks, 443-453. Springer.

Marwell, Gerald, and Ruth E. Ames. 1981. "Economists Free Ride, Does Anyone Else?" *Journal of Public Economics* 15: 295-310.

Merton, Robert K. 1948. "The Self-Fulfilling Prophecy." *Antioch Review* 8: 193-210.

Miller, Dale T. 1999. "The Norm of Self-Interest." *American Psychologist* 54 (12): 1053-1060.

Nagel, Ernest. 1961. *The Structure of Science: Problems in the Logic of Scientific Explanation*. London: Routledge.

Popper, Karl R. 1957. *The Poverty of Historicism*. London: Routledge.

Romanos, George D. 1973. "Reflexive Predictions." *Philosophy of Science* 40: 97-109.

Vetterling, Mary K. 1976. "More on Reflexive Predictions." *Philosophy of Science* 43: 278-82.