Reflexive Measurement

James Michelson

Department of Philosophy, Carnegie Mellon University

September 27, 2023

Abstract

This essay offers a unified philosophical account of observer effects in social science. When agents are aware their behavior is subject to scientific inquiry they often act in ways that render measurements of them unreliable. This is the problem of *reflexive measurement*. In order to develop this novel account, I provide a general characterization of reflexivity which encompasses the full scope of scientific practice: theorizing, prediction, measurement, etc. The characterization captures the insights of contemporary philosophers of science working on reflexive prediction alongside observations by scientists grappling with observer-type effects in the course of their research. This account sheds new light on the use of reflexivity as a demarcation criterion between social and natural sciences and yields concrete proposals for how to overcome problems of reflexivity in applied science.

1 Introduction

Observer effects in the social sciences go by many names. It is not uncommon to speak of the 'Hawthorne effect' (see Landsberger 1958) or an 'experimenter effect' (Rosenthal 1966) when the scientist or their science has a causal effect on its target of study. The idea that a measurement can causally affect the phenomenon it investigates is so widespread it is even enshrined in the common adage known as 'Goodhart's Law'[1]. Observer-type effects turn out to be ubiquitous across the social sciences, recognized and explored

[1] 'When a measure becomes a target, it ceases to be a good measure' (Goodhart 1984).

by economists and psychologists alike (see, for example, Friedman 1953; Gergen 1973). Yet unified philosophical accounts of these diverse effects are rare despite their similar structure and widespread occurrence. For all that social scientists have labored long and hard to mitigate these problems in the course of their research, philosophers of science have neglected to put observer effects into broader philosophical perspective in a manner that can aid the practice of science.

In philosophy of science, the related idea of a 'self-fulfilling prophecy' is an ancient one, going back as far as the story of Oedipus, a mythical ancient Greek king whose best efforts to thwart an oracle's prophecy led to its tragic fulfillment. In its more modern guise philosophers of science have called the idea *reflexive prediction* and its counterpart notion in the social sciences is commonly traced back to sociologist Robert K. Merton's (1948) discussion of "self-fulfilling science". Like observer effects, this concept also captures the idea of the causal impact of science on what it studies. A canonical example of this phenomenon is a bank run: announcing an impending bank run may subsequently incite one. Contemporary philosophers of science have stressed the challenge reflexive prediction poses for theory development and testing in the social sciences (see Kopec 2011; Lowe 2018). Despite a number of highly distinguished accounts of the idea of reflexive prediction over the past century, the animating idea of the causal effect of science on what it studies and the challenge that it poses is rarely married to a discussion of measurement.

This essay offers a novel account of the concept of *reflexive measurement*. This account captures the salient features of diverse observer-type effects and recovers the intuition that a measurement is reflexive when agents are aware they are subjected to it. This is similar to the well-known idea of 'measurement as intervention' in the philosophy of economics (e.g., Morgan 2001). However, unlike existing versions of this account of measurement, a sensitivity to the different ways measurements causally affect their target of study necessitates revising the naive version of the measurement-as-intervention story. Measurements can sometimes fundamentally alter the phenomenon they investigate and other times only affect the data collected, leaving the underlying phenomenon unchanged. For example, in the context of survey research, it is common for respondents to lie about their preferences and opinions but the survey—the measurement instrument—does not causally affect the underlying phenomenon. The different ways measurements can causally affect the phenomenon they investigate are illustrated in more detail in a pair of examples below (examples 3 and 4) drawn from contemporary social science.

Behind the account of reflexive measurement presented here is a reconceptualization of reflexivity in science writ large. Existing philosophical accounts of reflexive prediction would be of considerable help in better understanding the causal effects of measurements on their targets of study were it not for the fact that these insights are specifically tied to an understanding of prediction and rarely given in any form of generality. Thus, it is necessary to step back from specific scientific practices (e.g., prediction, measurement, theorizing) and ascertain how science—broadly understood as a sociological phenomenon—interacts with what it studies. A clear pattern then emerges. The causal effects of science occur only when agents are aware of the science that investigates their behavior. This, in a nutshell, is reflexivity[2]. This more general philosophical characterization of reflexivity is supported by scientific accounts from empirical psychology and recent developments in theoretical computer science.

The engagement with scientific literature beyond the traditional domains of economics and sociology paves the way for a potential solution to the problem of lying, misreporting, and withholding data—a common form of reflexive measurement across scientific domains. Drawing on insights from the mathematical and experimental psychologists (Luce 1995; Gergen 1973), designing a measurement to ensure truth-telling is beneficial for those studied yields measurements that are more reliable evidence for the underlying phenomenon. In some sense, in the face of reflexivity, measurements need to be made incentive-compatible or, as I refer to it below, measurements need to be *reflexively optimal*. This idea also complements recent developments in theoretical computer science in the field of incentive-compatible learning and performative prediction (Hardt et al. 2016; Perdomo et al. 2020; Cai, Daskalakis, and Papadimitriou 2015). The result of this engagement with psychology and computer science is that existing mathematical approaches to dealing with measurement are found to be inadequate. The commonly used framework of de-biasing systematic measurement error can be shown to be insufficient for addressing the more fundamental problem of reflexive measurement.

The essay is structured as follows. In section 2, I review the major conceptual innovations on the topic

[2] Note, the concepts of *performativity* and *reactivity* are very similar and specific differences are discussed where appropriate. Since the point of departure for this work on measurement is the philosophical concept of reflexive prediction this account draws most heavily on those sources. A full discussion of the differences between these related concepts is beyond the scope of this paper.

of reflexive prediction by philosophers of science over the past century. Additionally, this review is supplemented by considerations of observer effects by psychologists and recent developments in theoretical computer science which are particularly germane to the topic of reflexive measurement. Section 3 collects these accounts and provides a general characterization of reflexivity in science. The central insight of this section is that agents' awareness of their position in a scientific study is the key causal pathway for reflexive effects. This account sheds new light on an old, established demarcation between social and natural sciences, which is further explored in section 4. With a more general account of reflexivity in hand, the topic of measurement is then explored in section 5. Reflexive measurement is best captured by the measurement-as-intervention view, however, the naive understanding of this position requires modification in light of the distinction between data and phenomena. When the causal effects of a measurement only affect the data collected but leave the underlying phenomenon intact, then by addressing the incentives that lead to the collection of unreliable data, scientists can mitigate reflexive effects by designing reflexively optimal measurements. This proposal is discussed in section 6, which precedes the concluding section (7).

2 Literature Review

The idea that science causally affects its target of investigation has been long discussed in both science and philosophy. In philosophy of science alone, it is known as 'reflexivity', 'performativity', and sometimes 'reactivity'. However, this notion is mostly commonly discussed in the context of scientific theories and predictions. In this section, I provide a stylized overview of the conceptual development of this idea over the past century in order to subsequently develop an account of how measurements can casually affect what they measure. I draw from the well-developed concept of *reflexive prediction* in philosophy of science and supplement this understanding with developments in contemporary science which explicitly concern tackling the problem of the causal effect of science on what it studies. For a more even treatment of the development of reflexivity, albeit with less focus on contemporary science, see the historical overview of (Mackinnon 2006)[3].

[3] Since (Mackinnon 2006) primarily addresses the developments of reflexivity in economics and sociology it suffers from a lack of consideration of reflexivity concerns in empirical psychology, which I

Mid-twentieth century philosophical accounts of reflexivity following the sociologist Robert K. Merton's (1948) seminal account of 'self-fulfilling science' are isolated. Karl Popper (1953) briefly discussed a general formulation of the idea in the context of historicism in philosophy of social science and Ernest Nagel (1961) also noted the challenge this posed for theory construction in the social sciences. Subsequent accounts in the 1960s and 1970s focused more narrowly on the causal role predictions in the social sciences have on shaping their own truth-conditions (e.g., Buck 1963; Romanos 1973). These accounts focus heavily on the formulation and dissemination style (*FD-style*) of the prediction: whether the prediction was published in a newspaper or discussed on cable news, whether the prediction was public or private, etc. To different but ultimately similar degrees, these authors acknowledge that a single prediction will not, by itself, have a reflexive effect independent of how it comes to be known by those it makes predictions about.

These accounts from the 1960s and 1970s understood predictions in science as having a definite true/false truth value. A significant recent contribution by Kopec (2011) challenged this view, articulating a conception of probabilistic reflexive predictions. Here, a prediction is reflexive if it changes the probability of the event it predicts[4]. The common use of applied statistics in developing predictions in the social sciences is better accommodated by the account of (Kopec 2011). For example, social scientists have investigated whether public opinion polls indicating a favorite candidate in an upcoming election can increase that candidate's probability of winning (Rothschild and Malhotra 2014) and also whether the effects of election forecasts may depress voter turnout (Westwood, Messing, and Lelkes 2020). A narrow focus on the ultimate truth condition of the prediction misses the ways in which a prediction can nonetheless change individual behavior while leaving the result aggregate phenomena unchanged. Thus, even if a number of voters vote differently in light of public predictions, the result of an election may nonetheless remain

explicitly address below. Additionally, the related issue of performativity in computer science is missing from this account.

[4] Kopec distinguishes between strong and weak predictions, where 'strong' reflexive predictions are predictions that "switch the truth-value of the prediction" and 'weak' reflexive predictions merely "change the probability of the predicted event" (Kopec 2011, p1252-3). Since the strong reflexive predictions are a subset of weak reflexive predictions, and the latter offers a substantive conceptual innovation over previous accounts, I consider only the latter here.

unchanged.

A further criticism of existing philosophical work on reflexive prediction is that it fails to account of the idea that certain predictions may be more or less reflexive (Lowe 2018; Cejka 2022). The "Mertonian-derived, truth-centric notion of reflexive prediction" (Lowe 2018, p10) fails to capture the idea that, in some cases, the effect of a reflexive prediction on (even the probability of) an event is "marginal at best" (Lowe 2018, p8). If an impending bank run is announced on the front page of the newspaper of record it has a vastly different effect than on the front page of a local student newspaper. The shift in focus to degrees of reflexivity represents a welcome change in our philosophical understanding of the many ways scientific predictions can interact with the social world.

Scientists have also developed their own accounts of reflexivity which are notably different from the philosophical views considered above. Contemporary work by economists on reflexivity has "situate[d] the concept in recent thinking on complex adaptive systems" (Beinhocker 2013, p331). This turn towards characterizing reflexivity in terms of systems-type thinking is associated with financier and investor George Soros, who has claimed that understanding the concept of reflexivity has enabled him to profit from his investments in financial markets (see, for example, Soros 2013). On this account, reflexivity is a property of systems. The systems-account emphasizes the interactions between agents and their environment, as well as explicitly conceptualizing agents' goals and cognitive abilities. An upshot of this account is that it arranges different systems along a 'spectrum of complexity' (Beinhocker 2013, p337) and enables the comparison of physical, human, and artificial systems in terms of reflexivity and complexity—an unusual feature of accounts of reflexivity.

Despite the common focus on reflexivity in economics by philosophers of science issues of reflexivity are found across many other sciences. It is helpful to also draw from the discipline of experimental psychology that has confronted reflexivity as a practical challenge. In doing so it becomes possible to develop a clearer overall picture of the causal effect of science on what it investigates. Phenomena like the 'experimenter effect' (Rosenthal 1966) and 'demand characteristics' (Orne 1962) have been well-known for decades and demonstrate the problems that come with either revealing information (e.g., a theoretical premise or expected result) to study participants during the course of research or participants 'guessing' the aims and objectives of the study and then adjusting their behavior accordingly[5]. The implications

[5] Experimenter effects also include the presence of implicit cues which can unconsciously influence

of these findings constitute "a fundamental difference" (Gergen 1973, p313) between natural and social science.

Writing about the dangers of theories of psychology that are falsifiable "at will" by knowing study participants, mathematical psychologist R. Duncan Luce proposed the 'non-oxymoron criterion' (Luce 1995, p3) for theory-testing: scientists should be confident that their experimental design allows the theory to be tested despite the subject's knowledge of the theory. In other words, psychological hypotheses should not be able to be confirmed (disconfirmed) by the study participant at will. Psychologists have differed in their recommendations for how to avoid this. On the one hand, considering only "naive subjects" ensures that study participants are uninformed and therefore theories can be tested in "an uncontaminated way" (Gergen 1973, p313). On the other hand, we can directly address the self-interest of the subjects such that it "behooves the subjects to reveal their true preferences" (Luce 1995, p9). Truth-telling is to be made, in some sense, incentive-compatible. These views complement existing philosophical accounts by highlighting characteristics of study participants (e.g. how informed they are, their goals and desires) which contribute to the reflexivity of science.

I will return to Luce's comments about study participants "reveal[ing] their true preferences" in more detail below. It is worth noting, however, that recent developments in theoretical computer science specifically address reflexivity concerns at the intersection of measurement and prediction by explicitly modeling self-interest when designing algorithms that learn from data. This literature on *incentive-compatible learning*[6] concerns eliciting accurate data when the source of the data has knowledge of the structure of the algorithm and its subsequent use. Examples of this kind of work include eliciting truthful information when people can strategically withhold data (Krishnaswamy et al. 2021) and obtaining high-quality data when data collection is costly (Cai, Daskalakis, and Papadimitriou 2015). I will cover in more detail the simpler case of (Caragiannis, Procaccia, and Shah 2016), who consider the classic problem of estimating the population mean of an unknown single-dimensional distribution where samples are supplied by strategic agents who wish to pull the estimate as close as possible to their own value. The approach of this line of work addresses reflexivity directly and attempts to mitigate its effects by explicitly modeling study participants' behavior. I discuss this subtlety in more detail in Section 5.

[6] This is also sometimes called *strategic classification* (Hardt et al. 2016) or *incentive compatible estimation* (Cai, Daskalakis, and Papadimitriou 2015) or *performative prediction* (Perdomo et al. 2020).

self-interest such it is beneficial to those studied to reveal their true preferences.

In conclusion, it is important to bear in mind that in addition to philosophy of science and economics, disciplines as diverse as psychology (Luce 1995), political science (Rothschild and Malhotra 2014; Westwood, Messing, and Lelkes 2020), complex systems (Beinhocker 2013), and even theoretical computer science (Hardt et al. 2016; Perdomo et al. 2020) have all grappled with issues of reflexivity in different forms. As I will detail below, observer effects are ubiquitous across scientific domains and only a broader understanding of these effects can do justice to the complexity of reflexivity in science. This literature review aims to surface some of the common concerns across these disciplines alongside the development of reflexivity as a key idea in contemporary philosophy of science. The next section will tie together these concerns into a general characterization of the concept of reflexivity.

3 Characterizing Reflexivity

Since almost all explicit definitions of reflexivity are inextricably tied to prediction[7] in this section I propose a characterization of reflexivity which applies to all scientific practices. As such, it will necessarily be broader in scope and include more of scientific practice than is common on other accounts. The proposed account is closer in spirit to earlier attempts to understand reflexivity which attempted to grapple with the "complicated interaction between observer and observed" (Popper 1953, p14) at a high level of generality. Drawing from the approach of Grunberg (1986), the account proposed here sheds light on the causal pathway by which reflexive effects manifest themselves. This serves to fix ideas for the discussion of measurement in subsequent sections. Additionally, this account extends to non-social scientific domains; a feature of reflexivity that has been widely under-appreciated by those who insist the concept uniquely applies to the study of humans and human behavior.

The motivating question for this more general account is: which scientific practices might be reflexive? All conceptions of reflexivity considered in the preceding section implicitly rely on a view of science that encompasses the social interaction between 'observer and observed' (Popper 1953) or 'scientist and study participant'(Gergen 1973; Luce 1995). A broad view requires us to consider science as a sociological

[7] The exception being systems accounts (e.g., Beinhocker 2013; Soros 2013), however, these have their own limitations, as discussed below.

phenomenon and allow our characterization of reflexivity to include facts concerning how the science in question interacts with its target of study. Who the scientist is or what institution they work at can have an outsized impact on the results of a scientific study. This is, effectively, a more general formulation of the 'formulation/dissemination-style' (*FD-style*) of reflexive predictions (Romanos 1973) which naturally extends to other scientific practices like measurement.

However, adopting broader sociological considerations is no small requirement. This means that irrespective of whether a specific scientific practice is known, the mere knowledge of the institution that carries it out can be sufficient to elicit a reflexive effect[8]. Consider that when Google dropped its "don't be evil" motto (Basu 2015) users may have felt the need to change their behavior when interacting with Google's products. Even without knowing the specific scientific practices Google was carrying out to investigate their users' behaviors, this might—in the broad sense of being a social interaction between observer and observed—constitute an instance of reflexivity for Google's study of its own users.

This further entails that the private/public distinction that animates so much of the reflexive prediction literature is no longer helpful (e.g., Buck 1963; Romanos 1973; Grunberg 1986). To see this consider the following example (adapted from Grunberg 1986):

Example 1 (Sumerian Economic Forecasts) *The current Chairman of US Federal Reserve Jerome Powell delivers the Federal Reserve's annual economic forecasts on national television in ancient Sumerian with a presentation in cuneiform characters.*

Since, effectively, no one understands ancient Sumerian the forecast would be considered private. Setting aside the issue of market overreactions (e.g., Bondt and Thaler 1985), this would be an unprecedented action for a Chairman of the US Federal Reserve and may undermine investors' faith in the competence of major US financial institutions. This should constitute an instance of reflexivity in the same way that Google changing its motto should: the broader sociological context of a scientific practice can have enormous causal impacts on what it investigates.

The causal effect that science has on its target of study is clearly at the heart of all conceptions of reflexivity. On an overly simplistic view, reflexivity can even be understood as: the explicans causally

[8] There are even collateral effects from neighboring institutions or scientific practices. These are explored in the case of measurement in example 4.

affects the explicandum. Some kind of causal effect is clearly a necessary condition for the occurrence of reflexivity—on this, all philosophical accounts agree. But accounts of reflexivity differ in how they approach this. On one view, reflexivity is understood as a causal effect with a counterfactual component (Romanos 1973; Buck 1963). Another view emphasizes the causal effect of reflexivity as a stochastic phenomenon. A reflexive prediction, for example, changes the probability of an event occurring (Kopec 2011) and can even be ascertained by a test of statistical significance (Cejka 2022). A different kind of account gives definitions of reflexivity which omit causal language altogether in favor of clearly specifying the pathways along which the causal effects of reflexive science play out. Thus, for example, a reflexive prediction is "an utterance... made public in a language in terms that can be understood by the agents to whose behavior it refers and who therefore can by their actions either falsify or fulfill it" (Grunberg 1986, p476)[9]. A distinct advantage of this final approach is that it subsumes the causal effects of science on what it studies by specifying the mechanism by which agents might come to frustrate or fulfill a scientific prediction.

Before offering my own version of this type of account of reflexivity, it is important to be clear about the nature of "agents" that constitute part of the phenomenon investigated by scientists. In my view, the systems account of reflexivity (e.g., Beinhocker 2013; Soros 2013) correctly captures the important features of agency, including agents' goals, cognitive capacities, and actions within the scope of a definition of reflexive system[10]. It is important to consider why this is particularly helpful. Firstly, note that reflexivity may characterize sciences that investigate collections of humans: organizations, governments, firms, etc., which act with a singular purpose. These can be modeled as agents. Secondly, it would be philosoph-

[9] Alternatively, "in order to be reflexive it is sufficient for a public prediction to be partially believed" (Grunberg 1986, p484).

[10] However, I do not believe the most promising path towards characterizing reflexivity is to "situate the concept in recent thinking on complex adaptive systems" (Beinhocker 2013, p331). Although there are undoubtedly good reasons to think about reflexivity in this manner for large-scale, complex phenomena like financial markets, the 'systems' approach is heavy-handed for the kind of small-scale scientific investigations like laboratory studies (Luce 1995) and individual medical diagnoses (Hacking 1995). Especially since some systems accounts of reflexivity (e.g., Beinhocker 2013, p332) require that all reflexive systems be complex systems.

ically underwhelming to propose an account of reflexivity which rules out interesting cases like missiles (Grünbaum 1963) or thermostats (Beinhocker 2013) simply because the only agents to which the concept of reflexivity applies are human beings[11]. It is desirable to simultaneously capture the intuition that there is something particularly philosophically interesting about the problems faced by social science but also that we should be open to discovering these problems in other scientific domains. Although there will certainly be disagreement over what constitutes agency, this ambiguity is a deliberate feature of the account presented here and is explored in more detail in section 4.

The account of reflexivity proposed here requires that the causal effect of observation or measurement or prediction—any form of scientific practice—on the target of inquiry be mediated through the awareness of the agents that constitute part of the phenomenon under investigation. Here, I am trying to generalize to all scientific practices the idea that the mechanism for a prediction to be reflexive is for it to be "partially believed" (Grunberg 1986, above). It is meaningless to speak of "belief" in the context of measurement. Some minimal degree of awareness of being observed is the relevant necessary condition for reflexivity. This requirement widens the scope of what is to be considered reflexive, as did the move to include the broad sociological context of scientific practice beyond, for example, individual public predictions. What ultimately matters for reflexivity is not how a given prediction, theory, or measurement was published or disseminated (i.e., its *FD-style*) but instead that the agents came to learn it.

The implications of this novel understanding can be fully seen in the following example.

Example 2 (Stoplight Example) *A researcher aims to measure traffic patterns at an intersection. They stand on the side of the road noting the presence and absence of cars waiting at a stoplight. Inadvertently, however, they keep stepping on the cable that powers the stoplight, affecting the frequency with which the stoplight changes color.*

In this example, although the scientist causally intervenes in the target system they are seeking to study, the effect of this causal intervention is not a function of the agents' awareness of the science or scientist that studies their behavior. Thus, if the drivers of the cars (i.e., the agents) are unaware of the science that investigates their behavior, then the scientific study is not reflexive. This extends to the institution that the scientist is working for: if the agents are unaware not only of the scientist at the stoplight but of the

[11] To assert the only kinds of agents which manifest reflexive effects are humans is question-begging with regards to claiming that reflexivity only applies in the social sciences.

broader sociological context in which they conduct their science, only then is science truly non-reflexive. As noted above, large corporations like Google and intelligence agencies like the CIA, which are often jokingly considered omniscient, will elicit reflexive effects even if the specific scientific investigations they carry out are unknown to those they study.

It is tempting to argue that awareness is also sufficient for reflexivity, however, I believe this kind of precise definition leads to the consideration of unhelpful counterexamples[12]. This is a feature of the inclusion of collateral reflexive effects (from, say, neighboring scientific institutions with bad reputations as covered in example 4 below) in the proposed characterization of reflexivity. When coupled with awareness as the appropriate causal pathway for mediating reflexivity renders the range of cases to which the designation of reflexivity applies very broad. This is partly by design: the goal of this account is to extend existing ideas and intuitions about reflexivity to (potentially) cover all scientific practices. Even Ian Hacking's (1995) seminal account of the causal effects of scientific theories themselves cannot be included in a discussion of reflexive predictions without significantly changing the scope of the argument (and all the relevant definitions of reflexive prediction). Treating awareness as a sufficient condition for reflexive entails that a science can be reflexive without causally affecting its target of study, a conclusion completely at odds with the animating idea of the characterization given here.

The characterization of reflexivity proposed here entails that almost all[13] social scientific practice is

[12] Assume only one species of aliens exists and consider that their alien social science which investigates human behavior on earth is entirely undetectable by us (i.e., has no causal impact we can discern). Despite this, some members of the public believe that aliens are real. Perhaps they have they have filled their imaginations with stories of Area 51 or watched too many *X-Files* episodes. Thus, they adapt and change their behavior in ways they think might frustrate alien social science. Since the account here argues in favor of considering collateral effects of institutions on earth (e.g., the FBI, Hollywood, etc.) a key part of reflexivity, then the alien social science is reflexive despite the lack of *any* causal effect on the phenomena it investigates.

[13] There are two exceptions here. First, when the agents that comprise the target of investigation are entirely unaware of the science that investigates them (as covered above). Second, the study of historical phenomena since there is no causal effect of contemporary science on past events (I will return to this below).

reflexive. This feature of my characterization of reflexivity might strike the reader as unwelcome. Yet narrowly defining reflexivity in terms of "causal factors" (Romanos 1973; Buck 1963) or "changes in probability" (Kopec 2011) to pick out particular instances of reflexivity lands philosophers of science in the awkward position of assuming the role of scientists: determining what is and isn't reflexive in virtue of measurable effects. Moreover, recent developments in how to think about reflexivity emphasizing that reflexivity is a matter of degree (Lowe 2018) lend support to the idea that even minimal reflexive effects are still worthy of inclusion in the definition of reflexivity.

Ultimately, if reflexivity is defined by its effects it lands us with an arbitrary delineation of the term. Consider that a definition using the language of "causal factors" and "changes in probability" entails that the same prediction uttered in two almost identical circumstances might nonetheless result in one being reflexive while the other is not. These differing circumstances could be different days of the week, neighboring geographic regions, or even just differ in as much as a single study participant. Even more problematic is the fact that the absence of a discernible reflexive effect does not indicate the absence of reflexivity. A public prediction might result in exactly the same pattern of behavior (or probability of its occurrence) but the motivations for carrying out the behavior may have completely changed as a result of the prediction. A focus on the causal pathway by which reflexivity manifests allows philosophers of science to sidestep issues with ascertaining whether there is an appropriately reflexive causal effect for a given scientific practice.

Thankfully, social scientists are increasingly aware of the reflexive effects of their science. Contrary to earlier philosophical accounts of reflexivity which could only find a "a great deal of anecdotal evidence" (Grunberg 1986, p487) for the existence of reflexive predictions, a serious effort has been made to investigate the effects of public predictions in areas like election forecasting. Well-known election forecasts in the United States like Nate Silver's 538 website[14] which get national press coverage are now being investigated for their effects in depressing voter turnout (Westwood, Messing, and Lelkes 2020). Additionally, opinion polls indicating a favorite candidate in an upcoming election can increase the probability of that candidate winning (Rothschild and Malhotra 2014). The advantage of taking seriously the recommendation to treat reflexive effects as varying by degree (Lowe 2018) is that it leaves open the possibility of

[14] `https://fivethirtyeight.com`

acknowledging that almost all social science is reflexive, though much of it might have barely any effect at all. Philosophers can offer a clear account of reflexivity and let scientists determine where it is appropriate to worry about it.

In summary: I provided a sociological picture of scientific practice, whereby a reflexive scientific practice can causally affect its target of inquiry when the cause is mediated through the agents' awareness. Although this condition is necessary for reflexivity, and, indeed, it is often sufficient, a definition should be avoided: it adds little to our scientific and philosophical understanding of a wide-ranging, complex phenomenon and only serves to distract us with far-fetched counterexamples. Furthermore, it is important not to define reflexivity by its effects. Today's election forecasts are reflexive, as are tomorrow's—irrespective of whether one elicits a causal effect and the other does not. What matters is whether the agents that comprise the phenomenon under investigation are aware of the prediction (and so it goes for measurement, etc.). Philosophers of science should let scientists determine the effects of reflexivity; our role is to clarify the phenomenon of reflexivity as one that does or does not apply to various scientific practices and domains. However, before instantiating this account in the novel context of measurement, it is important to consider the implications of this view for what kinds of science are reflexive.

4 Reflexivity & Social Science

An upshot of the preceding section is that it engenders a reconsideration of the claim that reflexivity is exclusively a property of the social sciences and therefore can be used as a demarcation criterion between scientific domains. It has been commonly asserted by philosophers (e.g., Buck 1963; Popper 1953) that reflexivity (in some form) "is a phenomenon proper to the social sciences" (Grunberg 1986, p484). Against this position, some have argued that feedback systems like missiles should be considered reflexive (Grünbaum 1956; Grünbaum 1963). Contemporary research has eschewed debates about demarcation (e.g., Lowe 2018; Kopec 2011; Cejka 2022), however, the use of reflexivity as a demarcation criterion between scientific domains remains undisputed. I do not wish to challenge the intuition that the study of living human beings and their behavior presents challenges not found when studying stars or atoms. This intuition strikes me as entirely correct. However, the characterization of the preceding section offers a means of reconciling this intuition with a more nuanced understanding of where reflexivity does and does not apply, ultimately entailing a rejection of it as a phenomenon “proper” to the social sciences.

It is never entirely clarified what is meant by “social science” throughout these arguments. Clearly, disciplines like economics, political science, and sociology are all included. On the side of “natural science”, presumably, we have subjects like physics, chemistry, and biology. Through these (and a few other) examples, this division of scientific domains effectively takes place at the level of academic departments. This is regrettable for two reasons. First, reflexivity does not apply uniformly within even a single social science and, secondly, interdisciplinary subjects like cognitive science are left out entirely.

Before considering the possibility of reflexive natural science, it is helpful to briefly dwell on the nature of agency alluded to in the preceding section. The picture of agency I’m relying on in my account of reflexivity is drawn from those who approach the study of reflexivity using the framework of complex systems (e.g., Beinhocker 2013; Soros 2013). These accounts rely on a system-level description of phenomena (like financial markets) which includes agents with goals, available actions, variable cognitive functions, and even “internal models” of how their actions yield consequences (Beinhocker 2013, p331). My account stressed agents’ awareness but eschewed the stronger criterion of “understanding” is used in some accounts of reflexive prediction (e.g. Grunberg 1986). My aim is to capture the idea of an observer effect in its broadest possible formulation. Reflexivity also requires agents—in some form or other—to be part of the phenomenon under investigation. I think it uncontroversial to say the study of stars and atoms can never be reflexive: these phenomena are simply not agents in any relevant sense of the word. Thus, the proposed account of reflexivity in some sense ‘lines up’ with the intuition that much of the social sciences are reflexive whereas much of the natural sciences are not.

However, we must not be too quick to jump to conclusions about the natural sciences; without a closer consideration of entities that meet the standards for agency, we should not conclude that natural sciences are entirely non-reflexive. Most notably, I think one scientific domain that most certainly meets the criteria for reflexivity is ecology: in particular the study of primates. Primates are most certainly aware of being observed. There is evidence of observer effects when studying Capuchin monkeys (Metcalfe, Yaicurima, and Papworth 2022) and there is even evidence it can be mitigated by habituation (Crofoot et al. 2010). The study of primates presents an excellent case for reflexivity: they have different goals, levels of awareness, and cognitive faculties than humans, and thus reflexivity entails entirely different behaviors than, say, humans reacting to election forecasts. Thus, it is possible to see the study of the natural world and the social world along a continuum where it is the differences in the types of agents studied and the causal interactions between scientists and their target of study that make for differences in the reflexivity of science.

As alluded to above, fields like cognitive science and artificial intelligence might include reflexive scientific practices. Ultimately, the inclusion of fields like these in the category of reflexive sciences hinges on the appropriate definition of agency and awareness. Counterexamples like missiles (Grünbaum 1956; Grünbaum 1963) and thermostats (Beinhocker 2013) to definitions of reflexivity which entail that only social scientific domains are reflexive prompt interesting questions about the nature of agency and its relation to awareness of scientific practices. These counterexamples are clear instances of causal feedback loops: they are self-correcting in the sense that once a goal is specified, these systems will take actions to change their environment to satisfy the goal[15]. Doing justice to all potential counterarguments involves giving a comprehensive account of agency (and also awareness), which is beyond the scope of this paper. However, my goal here is to bring to the reader's attention that the broad formulation of reflexivity in terms of agents' awareness encompasses scientific domains of study beyond the social sciences.

The goal of this section was to argue that the traditional conception of reflexivity as only applicable to the social sciences is incorrect and ultimately misguided. If the age-old division between "social" and "natural" science is based on the reflexivity of one (but not the other) then this understanding of the division reflects anachronistic thinking. Natural scientific domains like ecology are clearly reflexive. Furthermore, as the previous examples of thermostats and missiles show, reflexivity hinges on a definition of agency and what it means for agents to be "aware" of scientific practices. Ultimately, the framework developed here to better understand reflexivity allows us to extend considerations of the causal feedback effects of science to entirely new domains and forms of scientific practice.

[15] Furthermore, these kinds of systems may soon represent some form of scientific practice: the growth of automation in science means that in the not-too-distant future, it is entirely plausible that major policy decisions may be taken autonomously (e.g., Zheng et al. 2022)

5 Reflexive Measurement

Despite the enormous amounts of ink spilled by scientists lamenting the challenges of collecting accurate data from study participants in laboratories and surveys, this aspect of scientific practice has been mostly overlooked by philosophers of science working on reflexivity. In this section, I instantiate the concept of reflexivity in the context of scientific measurement. Note, however, no new philosophical account of measurement is given in this section[16]. Instead, the discussion of measurement sits closer to how a scientist encounters it. The focus of this section is on data collection since no measurement is possible without it. The heuristics to which scientists avail themselves to understand observer effects are exactly the level at which this account is pitched: it is an attempt to unify these solutions under a single philosophical perspective.

It is worth beginning with what is known about observer effects by scientists working across different fields[17]. In its most general formulation across the social sciences, this is known as the 'Hawthorne effect' (see Landsberger 1958) or 'experimenter effect'[18] (Rosenthal 1966), where humans react to being

[16] I make the minimal assumption that measurement requires data collection and focus on this aspect of measurement throughout this section. This aspect of measurement is common to all philosophical accounts of measurement I have been able to find. Additionally, this understanding of measurement remains neutral with respect to questions on operationalism and conventionalism, realism and measurement, model-based accounts of measurement, etc (see Tal 2020 for an extended discussion of philosophical accounts of measurement in science).

[17] Note, whether or not the following observer-type effects are general and to what extent they replicate is an active area of research in every scientific domain that discovers any hint of these effects. However, consideration of these effects is considered standard practice in 'good study design' across scientific fields, so much so that they are commonly found in textbooks across the social sciences (see, for example, Groves et al. 2011; Stantcheva 2022; Goodwin 2009).

[18] Though experimenter effects occur when study participants are not explicitly aware of them (e.g., through implicit cues that are registered subconsciously) the focus of this section is the reflexive effects of measurement. As per the characterization in the preceding section, these are only the effects that participants are aware of.

observed and change their behavior in light of this observation. This general effect has been given a myriad of more specific formulations in different circumstances. To name a few of the most common found in scientific experiments: 'demand characteristics' are a phenomenon where study participants in an experiment act in ways they think the scientist desires (Orne 1962); the 'Pygmalion effect' is a psychological phenomenon whereby high expectations lead to improved performance (Rosenthal and Jacobson 1968); the 'John Henry effect' concerns the actions that study participants take on learning they are placed in a control group (as opposed to a treatment group) to overcome the disadvantage of being an experimental control (Colman 2008, p399). Outside of experiments, observer effects are commonly found in applied survey research: 'priming' occurs when a survey asks leading questions which can skew survey responses (Stantcheva 2022, §6.2); additionally, 'social desirability bias' is the phenomenon of surveys respondent lying or not sharing sensitive opinions (Krumpal 2013). Even 'Goodhart's Law' has its origins in the challenges faced by economists measuring economic indicators to set monetary policy (Goodhart 1984).

These disparate concerns all have the same root: the causal effects of scientists and their science on the target of study. Crucially, the kind of effects cataloged above are all mediated through the awareness of the agents studied. Study participants in laboratory experiments and respondents taking surveys are all fully aware of their role in scientific studies. Indeed, this is required for ethics approval. Thus, reflexive measurement can best be understood as the idea of 'measurement as intervention', which has long been known in philosophy of economics[19]. Writing about the role of measurement instruments in economics, Mary Morgan shrewdly writes:

> "The ways in which the economic body is investigated and data are collected, categorized, analyzed, reduced, and reassembled amount to a set of experimental interventions—not in the economic process itself, but rather in the information collected from that process." (Morgan 2001, p237)

However, Morgan contends that the interventions do not causally affect the "economic process itself"

[19] Note the parallels with the contemporary theoretical computer science literature on strategic classification where the estimate of the target variable is called a "treatment" (Miller, Milli, and Hardt 2020) in an explicitly causal sense.

and instead affect the “information collected from that process”. In the context of much of contemporary economics, this seems apt, however, more broadly this account fails in other settings[20]. This insightful intuition about measurement instruments can be better appreciated by further considering the difference between ‘data‘ and ‘phenomena’.

A helpful philosophical account of the difference between data and phenomenon was developed by Jim Woodward (1989). Phenomena are “relatively stable and general features of the world which are potential objects of explanation and prediction by general theory”, whereas data “by contrast, play the role of evidence for claims about phenomena” (Woodward 1989, p393-4). What matters in any scientific description or analysis of a phenomenon is that “the data should be *reliable evidence* for the phenomena in question” (Woodward 1989, p398, emphasis original). Furthermore,

> “Scientific investigation is typically carried on in a noisy environment; an environment in which the data we confront reflect the operation of many different causal factors, a number of which are due to the local, idiosyncratic features of the instruments we employ (including our senses) or the particular background situation in which we find ourselves.” (Woodward 1989, p398)

In the context of the account of reflexive measurement introduced above (i.e., ‘measurement as intervention’), data reflect the operation of measurement instruments and the broader sociological context in which scientists administer their measurement. Crucially, however, the causal effect of the measurement on what is being measured might affect the underlying phenomena itself and/or the data collected about it[21]. To see this more clearly, I now consider two concrete examples.

To make vivid how an act of measurement may cause the phenomena under investigation to change, consider the following well-known experiment by psychologist Philip Zimbardo:

[20] See Example 3. I explore this in more detail below.

[21] Contemporary philosophical work on reactivity in similar contexts of measurement (e.g., Runhardt 2023) draws on a conception of reactivity developed by (Golembiewski, Billingsley, and Yeager 1976) which is an exemplar of this first type of reflexive measurement: a measurement which causally affects the underlying phenomenon.

Example 3 (Stanford Prison Experiment) *In 1971 a psychologist recruited participants for a "psychological study of prison life", which was a planned one-to-two week experiment that simulated prison life (see* Stanford Prison Experiment *2023, §2 for details). The goal was to assess the psychological effects of becoming a prisoner or prison guard.*

A full account of the Stanford Prison Experiment can be found in (Zimbardo 2008). It was prematurely ended after only 5 days since "prisoners were withdrawing and behaving in pathological ways, and... some of the guards were behaving sadistically" (*Stanford Prison Experiment* 2023, §8). The ethics of this kind of experiment have been questioned: participants who simulated prisoners were deliberately made to feel humiliated (*Stanford Prison Experiment* 2023, §3). Zimbardo's method of investigating the effects of simulated prisoner-guard has also been criticized as poor scientific practice (Texier 2019). Ultimately, it is abundantly clear that if the effects of the experiments are so pronounced as to induce a study participant to "[break] down and began to cry hysterically" (*Stanford Prison Experiment* 2023, §8) then the measurement is reflexive in the sense of causally affecting the underlying phenomenon.

In contrast, a reflexive measurement can causally affect the data collected by a scientist without altering the underlying phenomenon under investigation. This is common in almost all forms of survey research where participants have the opportunity to lie or misrepresent their opinions. Here, a sensitive issue like the approval of a controversial political figure may be unaffected by a reflexive measurement but the data collected may be influenced by the study participants' reluctance to truthfully report their views. I believe this is the most promising way to realize Mary Morgan's insight that measurements of the economy are interventions "in the information collected from that process".

The next example provides a concrete case where the social context of the scientific practice can create exactly this kind of effect.

Example 4 (2020 US Census) *In the run-up to the 2020 US Census, then-President of the United States Donald Trump made repeated remarks about the possibility of adding a citizenship question to the census (see Blake 2022). Subsequently, the Hispanic response rate was more than three times lower on the 2020 census than on the 2010 census*[22]*.*

[22] See Appendix 1 for calculation of this figure.

This example highlights an important and often overlooked facet of scientific practice: *who* the scientist is or *what* institution they represent can have direct consequences on their ability to investigate phenomena in the face of reflexivity. Here, explicit condemnation of undocumented immigrants by former president Donald Trump likely had a causal role in more than tripling the number of those of Hispanic origin in the US who didn't complete the 2020 census compared to 2010. The underlying phenomena of interest investigated by the census (e.g., respondent's age, gender, income, etc.) don't change but the data collected are influenced by a powerful leader with the ability to use census data to create policies that leave undocumented immigrants worse off by deporting them.

The sociological picture of science presented in the previous section which focuses on scientific practice is central to the view of reflexive measurement proposed here. Institutions, people, expectations, etc all matter in changing the calculus of costs and benefits that study participants carry out when providing data. Though measurements can alter an underlying phenomenon (as in the Stanford Prison Experiment example above) this is often considered poor research design and scientists typically seek to eliminate these effects insofar as they are able to. Mitigating reflexive measurement effects on the data is a more difficult proposition that requires reasoning about agents' goals and cognitive abilities. Given the challenges associated with overcoming this latter kind of reflexive measurement issue, I now turn to concrete proposals from across scientific domains developed specifically to collect data that are more "reliable evidence" for the phenomenon in question.

6 A Concrete Proposal

Measurement without care on behalf of scientists to mitigate observer-type effects on data collection—even if the underlying phenomenon is unchanged—will result in data that are not reliable evidence for the phenomenon in question. The resulting inferences and predictions will be artifacts of the measurement rather than accurately represent the phenomenon under investigation. I believe the solution to problems of reflexive measurement lies in developing a scientific understanding of how our measurements affect the incentive structures of the agents we collect data about. The rest of this section collects and synthesizes disparate observations from scientists dealing with observer effects to motivate a concrete approach for modeling the reflexive causal effects of measurement instruments.

We can be clearer about the particular phenomenon of reflexive measurement where a measurement casually affects the data collected but leaves the phenomenon unchanged. The change in the distribution of the sample resulting from a reflexive measurement can be thought of as a kind of *distribution shift*[23]. The distribution shift in the sample represents a departure from the population-level data model. The measurement itself is an intervention that, for example, affects study participants' willingness to lie or conceal information in the face of a prying scientist. Notice this intervention only affects the sample since it is only the sample that is subjected to the measurement. Thus, this kind of reflexive measurement can be thought of as a kind of distribution shift where the sample distribution no longer represents the population distribution.

By way of contrast to this understanding of reflexive measurement, it is helpful to consider the commonly used approach of 'de-biasing' systematic measurement error. The systematic component of measurement error always occurs, with the same value, when the instrument is used in the same way in the same case (see Tal 2019). Thus, for example, we might say the systematic component of measurement error for a poorly worded survey question on political attitudes occurs when the responses are, for example, 'X% more left/right-leaning'.

Example 5 (De-biasing measurement error) *Consider the case of a survey question asking about presidential approval in the US, which was answered by N respondents. The data $X_1, \dots, X_N$ are assumed to come from a Gaussian distribution with unknown mean μ and variance σ. A scientist might then want to learn the value of μ. The statistical error associated with each random variable X_i is decomposed into a random and systematic component:*

$$\epsilon_i = \epsilon_i^{random} + \epsilon^{systematic} = \mu - X_i$$

Given further knowledge of the particulars of this domain of social inquiry, the scientist might impose additional assumptions about, for example, the shape of the distribution of the errors, or their covariance structure. These assumptions capture some of the flaws associated with a particular measurement instrument or measurement process. Ultimately, a scientist could make a post-hoc correction for measurement error by subtracting off (often called

[23] This language is commonly used by researchers in computer science. Indeed, Perdomo et al. 2020 explicitly tie this concept to performativity. Here, instead, I make a similar connection to reflexivity.

'de-biasing') the systematic component of the error:

$$X_i^{new} = X_i + \epsilon^{systematic}$$

Which will provide a more accurate (i.e., less biased) estimate of μ.

This example is paradigmatic of how measurement error is handled in the social sciences. The measurement instrument is biased and interacts with those it's intended to measure in a uniform manner[24]. The post-hoc measurement error correction[25] can be read as, effectively, claiming the underlying data generating process measured by the instrument is actually a Gaussian distribution with mean $\mu - \epsilon^{\text{systematic}}$ and variance σ once the causal effect of the instrument on the data generating process is accounted for.

However, this kind of correction presumes the underlying sample distribution remains unchanged in the face of reflexivity except for a difference in means. In many settings, this may be a reasonable and accurate assumption. However, this framework of de-biasing error cannot account for changes in the higher moments (e.g., variance, skew, etc.) of the underlying distribution. Moreover, the type of statistical distribution itself might change, often considerably, as a result of the measurement. In example 4 above, some Hispanic subgroups who felt threatened by the increased condemnation of undocumented immigrants might be entirely missing from the resulting data. The problem of reflexive measurement is a deeper one than the framework of measurement error allows. This understanding of reflexivity and measurement error also applies more generally to de-biasing corrections in reflexive prediction (Cejka 2022). Sometimes these are appropriate responses to the problem of reflexivity in measurement, however, they are not a substitute for a general understanding of the problem of distribution shift induced by a measurement.

How best then to correct the sample distribution shift that results from a measurement? I think it is instructive to return to the observation by psychologist R. Duncan Luce who, when he advocated the

[24] In the words of political scientist Christopher Achen, "measurement error is primarily a fault of the instruments, not of the respondents" (Achen 1975, p1229).

[25] Since $\epsilon^{\text{systematic}}$ is unknown a correction is only possible with an estimate of this quantity. Ascertaining whether or not this estimate is unbiased with respect to reflexive measurement effects is non-trivial. For a related discussion of this problem in the context of reflexive prediction see (Cejka 2022).

'non-oxymoron criterion' for theory-testing discussed above, noted that studies should be designed such that it "behooves the subjects to reveal their true preferences" (Luce 1995, p9). The account of reflexivity in the previous section focused on the causal pathway of agents' awareness, coupled with their goals, cognitive capabilities, and the actions available to them. Explicit concern with what agents want facilitates the possibility of designing measurements that induce a distribution shift such that accurate data are collected because it is in the study respondent's best interests.

Thus, we can think about whether measurements are, in a loose sense, incentive-compatible[26]. Incentive-compatible measurements induce minimal distribution shift such that the data collected are reliable evidence for the underlying phenomenon. This idea is related to that of *performative optimality* developed in (Perdomo et al. 2020) to capture the distribution shift caused by performative predictions[27]. Two differences being: the account here presupposes neither a model nor some form of model retraining. A single measurement (a survey, a laboratory experiment, etc) should be designed in such a way as to be *reflexively optimal*. It should induce a distribution that is reliable evidence for the phenomenon under investigation (i.e., the sample distribution should be an accurate representation of the population distribution). Thus, it should incentivize truth-telling, discourage withholding relevant information, etc.

Further departing from (Perdomo et al. 2020), it is helpful to consider reflexive optimality an equilibrium notion[28] in the game-theoretic sense (e.g. Nash 1950). This is helpful for two reasons. First, it allows scientists to give an explicit model of agents' motivations and reasoning and how they interact

[26] This has a specific, technical meaning in the context of mechanism design. Here, I use it in the informal sense.

[27] Their proposed definition is one of iterative convergence from model retraining (Perdomo et al. 2020, Definition 2.3). The use of the word 'prediction' should not confuse philosophers: the problem they consider is simultaneously a problem of measurement. Data are collected for retraining after each iteration of the model is deployed.

[28] The equilibrium notion of (Perdomo et al. 2020, p1) "coincide[s] with the stable points of [model] retraining" and does not reflect an understanding of why agents act they way they do. In contrast, recent work (Oesterheld et al. 2023) conceives of a truth-telling equilibrium in a performative prediction game as one induced by the self-interest of the participants. In the author's view, this latter contribution is the more promising approach to tackling the problem of reflexive measurement.

with a measurement. As argued above, this is a key facet of understanding how reflexive science causally affects its target of study. Secondly, it facilitates the application of the techniques of mechanism design[29] to the problem of designing reflexively optimal measurement instruments. This turns out to be closely related to an active area of research in theoretical computer science called incentive-compatible learning. Here, the choice of statistical estimator or algorithm itself can induce people to adapt their behavior. Thus, it is possible to re-frame the choice of estimator or algorithm as one that induces truth-telling on behalf of those data are collected from. To better understand this approach, consider the following example from (Caragiannis, Procaccia, and Shah 2016):

Example 6 (Incentive-Compatible Mean Estimation) *A statistician is trying to estimate the mean preferred temperature of occupants of a building. A sample of occupants are randomly selected and asked their preferred temperature. Consider the following scenarios.*

In one case, each person sampled is told that the estimator the statistician will use for their estimate of the population mean is the sample mean. Notice that if you have a preference for, say, warmer temperatures, you are best off lying about your preferred temperature to raise the sample average. This is because more extreme values will raise the sample average.

Suppose the statistician instead uses the sample median as his estimate of the population mean and this is communicated to each person in the sample. Even if you have a preference for much warmer temperatures, you no longer gain by lying since the median is robust to large outlier values (see Caragiannis, Procaccia, and Shah 2016 for extended discussion of this result).

In example 6 above, there is an explicit model of the relationship between the statistical estimator and the data collected in terms of benefits to people (i.e., their utility). The choice of statistical estimator (i.e., measurement instrument) is recast as a game theoretic problem whereby the statistician and the people in the sample play a game. The statistician wants to estimate the population mean. People in the sample will report their preferred temperature truthfully if they stand to benefit from it or can't benefit from lying. Thus, the statistician can use the *sample median* to estimate the *population mean* to achieve reflexive optimality. Note that despite the game-theoretic formulation of the problem the goal of statistical inference remains the same.

[29] See (Börgers 2015) for an introduction to the topic of mechanism design.

In this example, the sample distribution changes as a function of the estimator yet the underlying phenomenon of interest (people's preferred temperature) remains unchanged throughout. The measurement instrument (i.e., estimator) is chosen so as to induce a distribution shift which is more reliable evidence for people's preferred temperature. This framework of incentive-compatible estimators and algorithms has been extended to explicitly causal settings (Toulis et al. 2015), forecasting problems (Roughgarden and Schrijvers 2017), and even bandit-type exploration algorithms (Mansour, Slivkins, and Syrgkanis 2019). It makes the strong assumption that study participants know the functional form of the estimator and possess an ability to reason about how the actions they can take ultimately affect their welfare. However, it explicitly models the incentive structures faced by agents whose behavior is measured. This captures the key idea of the proposal that opened the section: scientists need to understand how their measurements affect the incentives of agents they collect data about.

7 Conclusion

I have argued for a novel conception of reflexivity that puts the sociological practice of science at the center of our understanding of reflexivity. This move facilitates the consideration of the multitude of ways science and scientists causally affect their target of study. Reflexivity concerns a kind of causal effect that science has on its target of study where agents that comprise the phenomenon of interest are aware of the scrutiny they are subjected to. In the case of measurement, we can distinguish this measurement-as-intervention view by virtue of whether the measurement causally affects the underlying phenomenon or the data collected about it. In the latter case, we might be able to mitigate the effects of misreporting, lying, and withheld data by designing the measurement so that truth-telling benefits those whose data are being collected. If this goal is achieved, the measurement is reflexively optimal, and the data collected are reliable evidence for the phenomenon in question.

This work makes a number of philosophical and scientific contributions to the study of reflexivity. Existing accounts of reflexivity have neglected the causal role of measurement as a scientific practice that affects what it investigates. The proposed account of reflexive measurement provides a nuanced picture of the ways this causal effect occurs in practice: sometimes affecting the underlying phenomenon and sometimes only affecting the data collected. The view of science that animates this account extends beyond economics and sociology to include experimental psychology and contemporary research in computer science. The result is a clearer understanding of the challenges that face the social sciences and where these challenges exist in natural scientific domains like ecology. Finally, the philosophical project of conceptualizing reflexive measurement gives rise to concrete recommendations on how to design scientific measurement instruments.

A few points are worth emphasizing. Firstly, scientists who warn of the consequences of observer effects or self-fulfilling science often do not narrowly focus on either measurement or prediction but instead explore and investigate cases where they co-occur. In psychology, researchers (Gergen 1973; Luce 1995) consider how revealing a theoretical finding during a laboratory experiment can result in the study participants falsifying (or confirming) the experiment at will. In computer science, researchers have considered the effects of predictions on subsequent data collection (Perdomo et al. 2020). The account presented here offers a more general characterization of reflexivity which is more faithful to its varied manifestations. In my view, large swathes of science are entirely reflexive and yet, perhaps surprisingly, reflexive effects are often quite minor. Thankfully, scientists are increasingly aware of this phenomenon and have begun investigating specific occurrences of reflexivity in a far more thorough capacity than philosophers (e.g., Rothschild and Malhotra 2014; Westwood, Messing, and Lelkes 2020).

Additionally, an upshot of the characterization given in section 3 is that the terrain of the discussion concerning the presence of reflexivity, reactivity, and performativity in a scientific domain should shift from an antiquated "social" versus "natural" science framing to one instead marked by a deeper appreciation for the nature of agency. The question 'is a science reflexive?' is transformed into 'what is the nature of agency?' in virtue of the concern with awareness as the causal pathway along which reflexive effects materialize. Where opinions differ on the nature of agency, so too will they differ on the designations of reflexivity. Examples like missiles (Grünbaum 1956) and thermostats (Beinhocker 2013) make this point abundantly clear. In my view, this is a welcome change. It moves from treating an existing academic division of labor as a primordial categorization of scientific practice to one instead informed by careful study of differing targets of inquiry.

Two extensions to the line of research initiated here are clear. Firstly, it is worth noting a significant omission from the present account is that of qualitative social science. Qualitative research techniques across the social sciences have become increasingly sophisticated (see, for example, King, Keohane, and Verba 2021). Further research outlining how the account presented here interacts with scientific practices like structured interviews and ethnographic research would be welcome. Secondly, the concept of reflexive optimality is developed exclusively in the context of measurement. It is an interesting and challenging proposition to consider what reflexive optimality might look like for prediction and theory development. As introduced here, the concept is bound up with data collection, and extending the account here for other scientific practices may yield insights that aid scientists in overcoming the reflexive effects of science.

Almost half a century ago, political scientist Christopher Achen lamented the lack of understanding social scientists possess concerning how their measurement instruments investigate the world. He wrote:

> "[m]ajor improvements in our understanding of political thinking may therefore come to depend upon a considerably more advanced theoretical knowledge of our measuring instruments than we have yet mustered." (Achen 1975, p1231)

I have argued that part of the toolkit of modern science fails in the presence of a particular, pervasive type of measurement concern. It is my hope that this essay constitutes an accurate philosophical diagnosis of the problem of reflexivity coupled with a concrete proposal for addressing it in the context of measurement.

8 Acknowledgements

In no particular order: Dejan Makovec, Marina Dubova, Brendan Fleig-Goldstein, Kevin Zollman, Tim Cejka, Adrian K. Lee, Bele Wollesen, Agne Sabaliauskaite, Robin Mejia.

9 References

Achen, Christopher H. (1975). "Mass Political Attitudes and the Survey Response". In: *American Political Science Review* 69.4, pp. 1218–1231. DOI: 10.2307/1955282.

Basu, Tanya (Oct. 4, 2015). "New Google Parent Company Drops 'Don't Be Evil' Motto". In: *Time*. URL: https://time.com/4060575/alphabet-google-dont-be-evil/ (visited on 06/10/2023).

Beinhocker, Eric D. (2013). "Reflexivity, complexity, and the nature of social science". In: *Journal of Economic Methodology* 20.4, pp. 330–342. DOI: 10.1080/1350178X.2013.859403. URL: https://doi.org/10.1080/1350178X.2013.859403.

Blake, Aaron (Jan. 18, 2022). "The audacious timeline of Trump's failed plot on the census and citizenship". In: *The Washington Post*. URL: https://www.washingtonpost.com/politics/2022/01/18/audacious-timeline-trumps-failed-plot-census-citizenship (visited on 06/05/2023).

Bondt, Werner F. M. De and Richard Thaler (1985). "Does the Stock Market Overreact?" In: *The Journal of Finance* 40.3, pp. 793–805. URL: http://www.jstor.org/stable/2327804 (visited on 06/12/2023).

Börgers, Tilman (2015). *An Introduction to the Theory of Mechanism Design*. Oxford University Press.

Buck, Roger C. (1963). "Reflexive Predictions". In: *Philosophy of Science* 30.4, pp. 359–369. DOI: 10.1086/287955.

Cai, Yang, Constantinos Daskalakis, and Christos Papadimitriou (June 2015). "Optimum Statistical Estimation with Strategic Data Sources". en. In: *Conference on Learning Theory*. ISSN: 1938-7228. PMLR, pp. 280–296. URL: http://proceedings.mlr.press/v40/Cai15.html (visited on 07/29/2021).

Caragiannis, Ioannis, Ariel Procaccia, and Nisarg Shah (2016). "Truthful Univariate Estimators". In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, pp. 127–135.

Cejka, Timotej (May 2022). "Reflexivity of Predictions as a Statistical Bias". In: URL: http://philsci-archive.pitt.edu/21326/.

Colman, Andrew M. (2008). *A Dictionary of Psychology*. Ed. by Andrew M. Colman. 3rd. Oxford University Press. ISBN: 9780199534067.

Crofoot, Margaret C. et al. (2010). "Does watching a monkey change its behaviour? Quantifying observer effects in habituated wild primates using automated radiotelemetry". In: *Animal Behaviour* 80.3, pp. 475–480. DOI: https://doi.org/10.1016/j.anbehav.2010.06.006.

Friedman, M. (1953). *Essays in Positive Economics*. A Phoenix book. Business economics. University of Chicago Press. ISBN: 9780226264035.

Gergen, Kenneth J. (1973). "Social Psychology as History". In: *Journal of Personality and Social Psychology* 26.2, pp. 309–320.

Golembiewski, Robert T., Keith Billingsley, and Samuel Yeager (1976). "Measuring Change and Persistence in Human Affairs: Types of Change Generated by OD Designs". In: *The Journal of Applied Behavioral Science* 12.2, pp. 133–157. DOI: 10.1177/002188637601200201.

Goodhart, C. A. E. (1984). "Problems of Monetary Management: The UK Experience". In: *Monetary Theory and Practice: The UK Experience*. London: Macmillan Education UK, pp. 91–121. ISBN: 978-1-349-17295-5. DOI: 10.1007/978-1-349-17295-5_4. URL: https://doi.org/10.1007/978-1-349-17295-5_4.

Goodwin, C.J. (2009). *Research In Psychology: Methods and Design*. John Wiley & Sons. ISBN: 9780470522783.

Groves, R.M. et al. (2011). *Survey Methodology*. Wiley Series in Survey Methodology. Wiley. ISBN: 9781118211342.

Grünbaum, Adolf (1956). "Historical Determinism, Social Activism, and Predictions in the Social Sciences". In: *The British Journal for the Philosophy of Science* 7.27, pp. 236–240. URL: http://www.jstor.org/stable/685878.

— (1963). "Comments on Professor Roger Buck's Paper "Reflexive Predictions"". In: *Philosophy of Science* 30.4, pp. 370–372. DOI: 10.1086/287956.

Grunberg, Emile (1986). "Predictability and Reflexivity". In: *The American Journal of Economics and Sociology* 45.4, pp. 475–488.

Hacking, Ian (1995). "The Looping Effects of Human Kinds". In: *Causal cognition: a multi-disciplinary debate*. Ed. by David Premack Dan Sperber and Ann James Premack. Chap. 12, pp. 351–383.

Hardt, Moritz et al. (2016). "Strategic Classification". In: *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*. ITCS '16. Cambridge, Massachusetts, USA: Association for Computing Machinery, pp. 111–122. ISBN: 9781450340571. DOI: 10.1145/2840728.2840730.

King, G., R.O. Keohane, and S. Verba (2021). *Designing Social Inquiry: Scientific Inference in Qualitative Research, New Edition*. Princeton University Press. ISBN: 9780691224640.

Kopec, Matthew (2011). "A More Fulfilling (and Frustrating) Take on Reflexive Predictions". In: *Philosophy of Science* 78.5, pp. 1249–1259.

Krishnaswamy, Anilesh K. et al. (May 2021). “Classification with Strategically Withheld Data”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.6, pp. 5514–5522. DOI: 10.1609/aaai.v35i6.16694. URL: https://ojs.aaai.org/index.php/AAAI/article/view/16694.

Krumpal, Ivar (2013). “Determinants of social desirability bias in sensitive surveys: a literature review”. In: *Quality Quantity* 47, pp. 2025–2047. DOI: https://doi.org/10.1007/s11135-011-9640-9.

Landsberger, H. A. (1958). “Hawthorne Revisited. Management and the Worker, its Critics and Developments in Human Relations in Industry.” In: *Cornell Studies in Industrial and Labor Relations* IX.

Lowe, Charles (2018). “The Significance of Self-Fulfilling Science”. In: *Philosophy of the Social Sciences* 48.4, pp. 343–363. DOI: 10.1177/0048393118767087.

Luce, R. Duncan (1995). “Four Tensions Concerning Mathematical Modeling in Psychology”. In: *Annual Review of Psychology* 46.1, pp. 1–27. DOI: 10.1146/annurev.ps.46.020195.000245. URL: https://doi.org/10.1146/annurev.ps.46.020195.000245.

Mackinnon, Lauchlan (2006). “Appendix B: Reflexive Prediction”. Unpublished PhD thesis. URL: https://www.researchgate.net/publication/37618514_Reflexive_Prediction_A_Literature_Review. PhD thesis. University of Queensland.

Mansour, Yishay, Aleksandrs Slivkins, and Vasilis Syrgkanis (2019). *Bayesian Incentive-Compatible Bandit Exploration*. arXiv: 1502.04147 [cs.GT].

Merton, Robert K. (1948). “The Self-Fulfilling Prophecy”. In: *Antioch Review* 8, pp. 193–210.

Metcalfe, Chloë Alexia, Alfredo Yhuaraqui Yaicurima, and Sarah Papworth (2022). “Observer effects in a remote population of large-headed capuchins, *Sapajus macrocephalus*.” In: *International Journal of Primatology* 43, pp. 216–234. DOI: 10.1007/s10764-021-00264-w.

Miller, John, Smitha Milli, and Moritz Hardt (13–18 Jul 2020). “Strategic Classification is Causal Modeling in Disguise”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 6917–6926. URL: https://proceedings.mlr.press/v119/miller20b.html.

Morgan, Mary S. (Dec. 2001). “Making Measuring Instruments”. In: *History of Political Economy* 33.1, pp. 235–251. DOI: 10.1215/00182702-33-Suppl_1-235.

Nagel, E. (1961). *The Structure of Science: Problems in the Logic of Scientific Explanation*. Donald F. Koch American Philosophy Collection. Harcourt, Brace & World. ISBN: 9780710018823.

Nash, John F. (1950). "Equilibrium points in <i>n</i>-person games". In: *Proceedings of the National Academy of Sciences* 36.1, pp. 48–49. DOI: 10.1073/pnas.36.1.48. URL: https://www.pnas.org/doi/abs/10.1073/pnas.36.1.48.

Oesterheld, Caspar et al. (31 Jul–04 Aug 2023). "Incentivizing honest performative predictions with proper scoring rules". In: *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*. Ed. by Robin J. Evans and Ilya Shpitser. Vol. 216. Proceedings of Machine Learning Research. PMLR, pp. 1564–1574. URL: https://proceedings.mlr.press/v216/oesterheld23a.html.

Orne, M. T. (1962). "On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications". In: *American Psychologist* 17.11, pp. 776–783. DOI: 10.1037/h0043424.

Patrick Cantwell Peter Davis, James Mulligan (2012). *DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-G-03*. US Census Bureau. URL: https://www2.census.gov/programs-surveys/decennial/2010/technical-documentation/methodology/g-series/g03.pdf.

Perdomo, Juan et al. (13–18 Jul 2020). "Performative Prediction". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 7599–7609. URL: https://proceedings.mlr.press/v119/perdomo20a.html.

Popper, Karl (1953). *The Poverty of Historicism*. Harper Torchbooks.

Romanos, George D. (1973). "Reflexive Predictions". In: *Philosophy of Science* 40.1, pp. 97–109. DOI: 10.1086/288499.

Rosenthal, Robert (1966). *Experimenter effects in behavioral research*. Appleton-Century-Crofs.

Rosenthal, Robert and Lenore Jacobson (1968). "Pygmalion in the classroom". In: *The Urban Review* 3, pp. 16–20. DOI: https://doi.org/10.1007/BF02322211.

Rothschild, David and Neil Malhotra (2014). "Are public opinion polls self-fulfilling prophecies?" In: *Research & Politics* 1.2, p. 2053168014547667. DOI: 10.1177/2053168014547667.

Roughgarden, Tim and Okke Schrijvers (2017). "Online Prediction with Selfish Experts". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc.

Runhardt, Rosa W. (2023). "Legitimate Reactivity in Measuring Social Phenomena: Race and the Census". In: *Philosophy of the Social Sciences* 53.2, pp. 122–141. DOI: 10.1177/00483931221150487.

Soros, George (2013). "Fallibility, reflexivity, and the human uncertainty principle". In: *Journal of Economic Methodology* 20.4, pp. 309–329. DOI: 10.1080/1350178X.2013.859415.

Stanford Prison Experiment (2023). URL: https://www.prisonexp.org (visited on 07/11/2023).

Stantcheva, Stefanie (Oct. 2022). *How to Run Surveys: A guide to creating your own identifying variation and revealing the invisible*. URL: https://scholar.harvard.edu/files/stantcheva/files/How_to_run_surveys_Stantcheva.pdf (visited on 07/11/2023).

Tal, Eran (2019). "Individuating quantities". In: *Philosophical Studies* 176.4, pp. 853–878. DOI: 10.1007/s11098-018-1216-2.

— (2020). "Measurement in Science". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2020. Metaphysics Research Lab, Stanford University.

Texier, Thibault Le (2019). "Debunking the Stanford Prison Experiment". In: *American Psychologist* 74.7, pp. 823–839. DOI: https://doi.org/10.1037/amp0000401.

Toulis, Panos et al. (June 2015). "Incentive-Compatible Experimental Design". In: *Proceedings of the Sixteenth ACM Conference on Economics and Computation*. DOI: 10.1145/2764468.2764525.

Westwood, Sean Jeremy, Solomon Messing, and Yphtach Lelkes (2020). "Projecting Confidence: How the Probabilistic Horse Race Confuses and Demobilizes the Public". In: *The Journal of Politics* 82.4, pp. 1530–1544. DOI: 10.1086/708682.

Woodward, Jim (1989). "Data and Phenomena". In: *Synthese* 79.3, pp. 393–472.

Zheng, Stephan et al. (2022). "The AI Economist: Taxation policy design via two-level deep multiagent reinforcement learning". In: *Science Advances* 8.18, eabk2607. DOI: 10.1126/sciadv.abk2607.

Zimbardo, P.G. (2008). *The Lucifer Effect: How Good People Turn Evil*. Rider. ISBN: 9781846041037.

10 Appendix 1: Census Non-Response Calculations

The 2020 census post-enumeration survey data can be found in the US Census data tables[30], where the 'Net Coverage Error for the Household Population in the United States by Race and Hispanic Origin' is given by the variable **C_RACEHISUS** and the net coverage error is estimated at -4.99%. The data for the 2010 US Census are not available on the census data tables, however, the official estimated net undercount of Hispanics was -1.54% (Patrick Cantwell 2012, p1).

[30] `https://data.census.gov/table`