Ethical Theory and Moral Practice (2023) 26:127–152
https://doi.org/10.1007/s10677-022-10359-9

# Self-fulfilling Prophecy in Practical and Automated Prediction

Owen C. King[1] · Mayli Mertens[2]

Accepted: 20 December 2022 / Published online: 16 January 2023

## Abstract

A self-fulfilling prophecy is, roughly, a prediction that brings about its own truth. Although true predictions are hard to fault, self-fulfilling prophecies are often regarded with suspicion. In this article, we vindicate this suspicion by explaining what self-fulfilling prophecies are and what is problematic about them, paying special attention to how their problems are exacerbated through automated prediction. Our descriptive account of self-fulfilling prophecies articulates the four elements that define them. Based on this account, we begin our critique by showing that typical self-fulfilling prophecies arise due to mistakes about the relationship between a prediction and its object. Such mistakes—along with other mistakes in predicting or in the larger practical endeavor—are easily overlooked when the predictions turn out true. Thus we note that self-fulfilling prophecies prompt no error signals; truth shrouds their mistakes from humans and machines alike. Consequently, self-fulfilling prophecies create several obstacles to accountability for the outcomes they produce. We conclude our critique by showing how failures of accountability, and the associated failures to make corrections, explain the connection between self-fulfilling prophecies and feedback loops. By analyzing the complex relationships between accuracy and other evaluatively significant features of predictions, this article sheds light both on the special case of self-fulfilling prophecies and on the ethics of prediction more generally.

**Keywords** Self-fulfilling prophecy · Reflexive prediction · Predictive analytics · Accountability · Feedback loop

✉ Owen C. King
owen@owencking.net

Mayli Mertens
mayli@sund.ku.dk

[1] School of Information and Library Science, University of North Carolina, Chapel Hill, NC, USA

[2] Centre for Medical Science and Technology Studies, Department of Public Health, University of Copenhagen, Copenhagen, Denmark

## 1 Self-fulfilling Prophecies in Practical and Automated Prediction

### 1.1 Introduction: Problems with True Predictions

It is difficult to fault a true prediction. Even if the prediction was unscrupulous and turned out true by lucky accident or coincidence, it is hard to get too concerned. Luck runs out eventually, and unscrupulous epistemic behavior tends to catch up with us, one way or another. It is even more difficult to fault a streak of true predictions. A history of accuracy seems to commend a predictor, even when the explanation of success is obscure. Nevertheless, this article addresses a class of true predictions we ought not so readily regard as faultless.

A *self-fulfilling prophecy* (SFP) is not simply a true prediction, but a prediction that somehow brings about its own truth. As such, an SFP is, more generally, a *reflexive prediction*, a prediction that somehow affects the outcome predicted. As we will see, SFPs stand out as significant, indeed problematic, among reflexive predictions, because of their distinctive tendency, not just to affect the object of prediction, but to do so while deflecting scrutiny and defying accountability.

SFPs have been a curiosity since antiquity, and a subject of steady scholarly interest since the middle of the twentieth century. They have attracted renewed and special attention in recent years with the increasing automation of prediction. The automation of cognitive tasks, especially prediction, is a main practical goal in the burgeoning field of artificial intelligence (AI). It makes sense, then, that SFPs would be one focal point within the nebulous collection of moral worries about AI. AI is significant for enabling prediction at unprecedented scale and scope, and SFPs are the predictions with the most direct and pronounced effects on the world.

Scholars of various stripes have observed, bemoaned, and protested SFPs in automated prediction across governmental, educational, and commercial settings. For instance, in a much-discussed law review article about automated credit scoring—automated prediction of an individual's likelihood of defaulting on a loan—Danielle Citron and Frank Pasquale write:

> Scores can become self-fulfilling prophecies, creating the financial distress they claim merely to indicate. The act of designating someone as a likely credit risk (or bad hire, or reckless driver) raises the cost of future financing (or work, or insurance rates), increasing the likelihood of eventual insolvency or un-employability. When scoring systems have the potential to take a life of their own, contributing to or creating the situation they claim merely to predict, it becomes a normative matter, requiring moral justification and rationale.[1]

What this complaint shares with the many other critiques of particular self-fulfilling prophecies is an emphasis on the undesirable outcomes that the predictions brought about—in this case the unfortunate financial and occupational consequences for the people subject to prediction. Most critiques of SFPs are similarly focused on their undesirable, usually unintended, consequences, and are punctuated with calls for greater accountability for those consequences. The predominating concern is about perpetuation of systemic injustice, further disadvantaging already disadvantaged groups. SFPs, especially when

[1] Citron and Pasquale 2014, 18.

automated, threaten to forge a direct link from bias in prediction to the perpetuation of a correspondingly unfair reality.

Hence, critiques invoking SFPs are both timely and pressing. Yet extant critiques do not show what is especially worrisome about SFPs per se. So one may wonder: If the main concern is an SFP's production of harm or injustice, then why so emphatically point out the involvement of an SFP? Beyond the badness (or goodness, as the case may be) of the specific outcomes predicted, what else might be wrong with outcomes realized this way? The goal of this article is to answer that question by providing a general critique of SFPs, showing what makes them distinctively problematic, thus supporting and clarifying the specific critiques in which they figure. Our general critique of SFPs weaves together several strands: We describe mistakes due to insufficient respect for the diverse relations between predictions and the practical endeavors that call for them. We explain how SFPs hide mistakes in the ways predictions are made and used. And we show how these problems impede accountability for SFPs and their outcomes.

The article has three parts. The rest of Part 1 situates our discussion among other work on self-fulfilling and reflexive predictions, and then narrows our scope to predictions in practical endeavors, with special attention to automated prediction. Part 2 provides a descriptive account of SFPs, describing the four essential elements of SFPs and offering a succinct definition. Relying on this descriptive account, Part 3 presents our general critique of SFPs, focusing on the mistakes they involve and the unaccountability they engender, culminating with an explanation of how feedback loops perpetuate SFPs. We close with a few reflections on the complex ethics of practical prediction.

## 1.2 Prior Preoccupations with Self-fulfilling Prophecies

SFPs attracted widespread scholarly interest in the mid-twentieth century,[2] under the influence of sociologist Robert K. Merton. In 1948, Merton published the classic paper coining the term "self-fulfilling prophecy," defining an SFP as "a false definition of the situation evoking a new behavior which makes the originally false conception come true."[3] Merton's aim was to point out that our characterizations of social phenomena shape those very phenomena. Scores of research articles in the social sciences, along with casual commentaries, have invoked Merton's notion of an SFP to characterize and critique various social phenomena. However, social scientists have not aspired to a general critique of SFPs per se, nor to a conceptual analysis of SFPs detailed enough to support such a critique.

That is not to suggest that SFPs have received no careful philosophical attention. Spurred by Merton and contemporaneous work by Karl Popper,[4] philosophers of science have debated whether the potential for reflexivity in scientific prediction could mark the distinction between the social and natural sciences.[5] However, we can largely bypass this

[2] John Venn was perhaps the first influential scholarly commentator on reflexive predictions. In *The Logic of Chance*, in 1866, Venn noted that publishing general observations about human conduct may influence that very conduct (Venn 1866, 345).

[3] Merton 1948, 195.

[4] Popper 1957, 12-16. Popper coins the phrase "the Oedipus Effect" to refer to the effects of reflexive predictions.

[5] A short path through this debate could start with Merton 1948, and Popper 1957. From there, the main arc goes through Grünbaum 1956, Nagel 1961, Buck 1963, and Romanos 1973.

debate, because the predictions of interest here are not, in the first place, scientific predictions, made with the goal of testing a theory in terms of the predictions it yields.

Our interest is in practical predictions. Sorting types of predictions by their motivations, Wesley Salmon observes, “we sometimes find ourselves in situations in which some practical action is required, and the choice of an optimal decision depends upon predicting future occurrences.”[6] We understand practical predictions to be the predictions made with this type of motivation.[7] We take practical predictions to include the everyday predictions we make constantly—out of necessity, optimality, convenience, or habit—as we conduct our affairs and undertakings into the future.[8] Among the SFPs in practical prediction, we count nearly all of those that social scientists have studied in the wake of Merton’s classic article. These include self-fulfilling social judgments mediated by stereotypes of race[9] and gender,[10] self-fulfilling forecasts of technological innovation,[11] self-fulfilling electoral predictions,[12] self-fulfilling medical prognoses,[13] among many others.[14]

Perhaps the most studied and influential example of SFPs has been in education. In 1968, Robert Rosenthal and Lenore Jacobson described the “Pygmalion Effect,” whereby predictions of student success and ensuing teacher expectations affect student performance.[15] For use as a touchstone example, we reproduce a summary of Rosenthal and Jacobson’s famous Oak School experiment:

> All of the children in an elementary school were administered a nonverbal intelligence test disguised as a test that would predict intellectual “blooming” ... [A]pproximately 20 percent of the children were chosen at random to constitute the experimental group. Each teacher was given the names of the children from her class who were in the experimental condition and told that these children had scores on the “test for intellectual blooming” indicating that they would show remarkable gains in intellectual competence during the next eight months of school. The only systematic difference between the experimental group and the control group children, then, was in the mind of the teacher. All the children were retested eight months later with the same IQ test. Considering the school as a whole, those children whom the teachers

[6] Salmon 1981, 116.

[7] Similarly, George Romanos distinguishes between predictions which have a “conspicuously informative character and purpose” (i.e., practical predictions) and predictions which have a “scientific motivation” (1973, 99). Note, however, that the distinction between scientific and practical motivation is not always sharp, especially in medical research; see Mertens 2018, 288-290.

[8] Regarding the inextricable role of practical prediction in human life, see Rescher 1998, 2-4.

[9] See, e.g., Word et al. 1974.

[10] See, e.g., Snyder et al. 1977.

[11] MacKenzie 1996, chs. 3-4; Van Lente 2000, 54-59.

[12] See, e.g., Simon 1954; Marsh 1984.

[13] Christakis 1999, ch. 6; Wilkinson 2009. Some cases of self-fulfilling medical prognosis, when predictions and their consequences affect the patient independently of standard therapies, may count as placebo or nocebo effects (Christakis 1999, 145). When a prognosis is self-fulfilling, it falls within the scope of our analysis. However, our analysis is not intended to cover intentional inducement of placebo effects.

[14] We leave aside notable examples of reflexive prediction in finance and economics, which have attracted sustained attention and become integrated into the relevant disciplines. See, e.g., Diamond and Dybvig 1983, and Lucas 1976. We focus on contexts of prediction which do not yet subsume reflexivity into the basis and activity of prediction.

[15] Rosenthal and Jacobson 1968.

> had been led to expect greater intellectual gains showed significantly greater gains in IQ than did the children of the control group.[16]

Here the predictions of student improvement, as communicated to the teachers, were borne out by tests. The hypothesized explanation was *expectancy effects*: that the teachers' expectations influenced how they interacted with different students, in manner, assignments, feedback, etc., eventually affecting scholastic performance.[17] The predictions that caused these expectations and their effects—although enacted artificially in the Oak School experiment[18]—exemplify the practical predictions of interest here and exhibit how such predictions can be self-fulfilling.

### 1.3 Self-fulfilling Prophecies with Predictive Analytics

As practical endeavors encourage practical prediction, they also encourage development of new techniques for prediction. A popular book on predictive analytics begins with the premise, "A little prediction goes a long way," and explains, "A hazy view of what's to come outperforms complete darkness by a landslide."[19] Predictive analytics answers the demand for more plentiful practical predictions, and is at the center of AI-based data science.[20] In general, *predictive analytics* is the use of statistical and computational techniques to analyze and model data (typically large volumes of data) about observed instances of some phenomena, with the goal of making predictions about unobserved instances.[21] Thus construed, predictive analytics includes both predictions by humans using statistics and computation, and also the similarly derived predictions generated automatically by computers. The new wave of practical predictions facilitated by predictive analytics comes with a new wave of SFPs. We already mentioned SFPs in automated credit scoring.[22] SFPs in automated prediction have been noted also in criminal risk assessment,[23] hiring,[24] student curricular advising,[25] product recommendation,[26] and medical prognostication.[27]

Perhaps the most widely recognized SFPs in automated prediction have been in policing. The example is both urgent and instructive. Like many practical challenges, policing

[16] Rosenthal and Rubin 1978, 378.

[17] See Rosenthal and Rubin 1978, 377. Cf. Jussim and Harber 2005, reviewing research on SFPs involving teacher expectations and cautioning against overstatement of effect size.

[18] This experiment raises several ethical concerns. Fortunately, to avoid negatively impacting any students, the researchers induced only expectations of improved performance, not expectations of reduced performance (Rosenthal and Jacobson 1968, 65). Another concern is that teachers were misled about the experimental condition, and apparently neither the students nor their parents were informed at all. This speaks to a general difficulty with empirically studying reflexive predictions without purposefully misleading the subjects involved.

[19] Siegel 2016, 14.

[20] Dhar 2013, 66.

[21] Cf. Siegel 2016, 15. Although the term 'predictive analytics' comes from the field of business intelligence (Watson and Wixom 2007), it also describes the same techniques in scientific research (Shmueli and Koppius 2011).

[22] Citron and Pasquale 2014, 18. Zarsky 2014, 1405-1408. Pham and Castro 2019, 122.

[23] Harcourt 2007, 29-36.

[24] Kroll et al. 2016, 684.

[25] Carmel and Ben-Shahar 2017, 92.

[26] King 2020, 30-34. Rakova and Chowdhury 2019, 3.

[27] Geocadin et al. 2019, e524.

requires careful allocation of scarce resources—including the officers themselves. Ideally, officers would patrol exactly those places at exactly those moments where the most undesirable criminal activity will (or otherwise would) take place. Efficiently deploying patrols to approximate this ideal depends on accurate, fine-grained predictions. Assuming criminal activity follows patterns, predicting the time and location of crime "hotspots" has appeared to many police departments as an opportunity to benefit from predictive analytics.[28] Hence, since 2011, many U.S. cities have adopted place-based *predictive policing* systems, computer systems that predict locations and times that will have high levels of crime.[29]

One criticism of predictive policing is that it produces SFPs. An attorney for the Electronic Frontier Foundation, explains, "by increasing police focus on certain people and areas, the prediction that someone will commit crime or that some communities will have more crime almost becomes a self-fulfilling prophecy, because when the number of police is increased in a given area, it almost always results in more arrests."[30] To explicate: The analytics system predicts that a particular area will have a high crime rate at a particular time. In response, extra patrols are deployed to that area. With extra patrols, more crime is recorded there than would have been otherwise. Hence, the prediction is realized. In such a case, the prediction does not necessarily affect the background level of criminal activity—the sorts of occurrences which, if observed, might prompt a police report—but does affect the recorded crime rate by influencing what the officers do.[31]

With such self-fulfilling practical predictions in mind, we proceed to our general descriptive account.

## 2 What are Self-fulfilling Prophecies?

### 2.1 First Element: Treating the Prediction as Credible

Intuitively, predictions are statements, judgments, or messages, intended to be informative about matters yet uncertain. In the Oak School case, the communication by the administrators to the teachers that particular students would "bloom" that year was a prediction. With predictive policing, a prediction occurs when the computer system displays its indications of future hotspots. In ordinary speech, the term 'prediction' is ambiguous, sometimes referring to the action or event of predicting, sometimes the proposition predicted, sometimes the informational signal[32] expressing that proposition, sometimes more than one of these. For most purposes, ours included, this ambiguity is benign.

Although an intuitive conception of prediction largely suffices for our purposes, a few particular features of predictions will be crucial. First, we can speak of a prediction as

[28] Regarding the motivations for adopting predictive policing, see Goode 2011, and Berg 2014.

[29] PredPol has been a common predictive policing system in the U.S. since the 2010's. See Brayne 2017, 989-990, regarding use of PredPol. Place-based predictive policing—predicting the time and location of criminal activity—is different both from traditional police profiling and from systems that predict which individuals might commit crimes. See Ferguson 2017, 1123-1144, for a helpful break-down.

[30] Lynch 2016. See also Miller 2014, 124; Ferguson 2017, 1178.

[31] It has long been recognized that SFPs with this same structure occur with police profiling, especially racial profiling. See, e.g., Swett 1969, 93; Johnson 2000, 104-107; Harcourt 2007, 154-156.

[32] Since we will sometimes have in mind automated predictions, it is useful to think in terms of signals instead of utterances and inscriptions, which are more anthropocentric. Regardless, any signal expressing a prediction must have a syntax that allows some entity to interpret it as a prediction. Cf. Romanos 1973, 105.

turning out true or false, according to the truth or falsity of the associated proposition. Second, a prediction is supposed to be informative, not just by having content, but also by being about something yet unobserved, uncertain, or undetermined. Predictions may be about states of affairs that are determined but unverified, as well as about those not (yet) determined.[33] Third, a prediction—and here we mean a practical prediction—is supposed to be suitable to be used or relied upon. Not only does a prediction say that something is (or will be) the case[34]; a prediction, like an assertion, has greater strength than a supposition, hypothesis, or mere guess—purporting at least minimal suitability for practical reliance.

Our account of SFPs begins by considering what happens once a prediction has been made. A prediction may be made with all the force of an assertion, only to meet disinterest or incredulity. Sometimes, though, a prediction finds a receptive audience. A person or artificial system *treats a prediction as credible,* or *grants it credibility*, by treating it as providing information eligible for reliance. When, but only when, a prediction has been granted credibility does it actually become usable as a prediction. Until then, there can be no step from the prediction to its employment, and so no self-fulfillment.

What factors disposed the teachers in the Oak School experiment to treat as credible and rely on the predictions that certain students would "bloom" that year? What explains why a police captain would grant credibility to the predictions from a computer system and deploy officers accordingly? These are empirical questions, of course, but just raising them prompts us to notice many possible answers:

- *Persuasion*: This prediction, or the way it has been presented, is persuasive.
- *Deference*: The situation requires deference to the predictor.
- *Reputation*: The reputation or purported expertise of the predictor is impressive.
- *Wishful thinking*: This prediction is found agreeable.
- *Herd instinct*: Others are relying on this prediction.
- *Availability*: Any prediction would be helpful, and this one is available.
- *Process design*: The process is designed to employ this prediction.

We think of these as possible *credentials* for a prediction. For some employer of a prediction, the prediction's credentials encompass the circumstances or considerations—whether psychological, institutional, or procedural—that explain why that employer was disposed to treat it as credible.[35] The credentials supporting a single prediction may vary across time, circumstances, and different prospective employers of the prediction.

It is intuitive but naive to suppose that a prediction's credentials must be rooted in the evidence supporting it. Ideally, the credibility granted to a prediction is tied to the reliability of the predictor or the quality of evidence, but often that is not realistic. Practical endeavors commonly call for prediction despite—or even because of—great uncertainty. In the face of uncertainty, the main question about a prediction is less about its truth and more about whether to use it. Typically, for a prediction to be used, it must be at least plausible

[33] It is common to think of predictions as being about the future. See, e.g., Rescher 1998, 38; Searle 1976, 5. However, we allow that predictions can be about unobserved conditions, past, present, or future. This accords with the discourse of predictive modeling: A model based on observed past instances of some phenomenon can generate predictions about unobserved past instances.

[34] Cf. Searle 1976, 10-11.

[35] Cf. Biggs 2009, 306. However, *pace* Biggs, we do not assume self-fulfillment occurs only with predictions that are arbitrary or false.

enough not to arouse immediate objection,[36] but, beyond baseline plausibility, the credentials may have as much to do with the circumstances as with the evidence. The role of the credentials, whatever they happen to be, is to permit and motivate the employment of the prediction.

## 2.2 Second Element: Employing the Prediction

Treating a practical prediction as credible does not require actually using it; granting credibility just makes it eligible for use. To *employ* a prediction is to rely upon it or put it to use. An *employer* of a prediction is a party or entity that employs it. Of course it may be that the employer and the predictor are the same. An *employment* of a prediction consists of actions or operations performed on the basis of the prediction.

An ordinary way to employ a prediction is to perform some overt action in response to it. Teachers may employ predictions of students' performance by taking those predictions into account when crafting assignments or selecting exercises. With predictive policing, a police captain might employ predictions of crime hotspots by directing officers to patrol particular locations, and the officers might employ the same predictions by exercising extra vigilance when in those areas. Employment does not always involve human action; the process may be completely automated. For example, a prediction of a consumer's interest in a particular product may be employed by a computer automatically displaying a relevant advertisement.

A common way of overtly employing a prediction is to disseminate it, for instance, by announcing or publishing it.[37] In practice, there may not be much to distinguish the original making of a prediction from its employment through dissemination, though the choice of how and to whom to disseminate it is a choice about employment. Once a prediction has been disseminated, the parties on the receiving end may (or may not) treat it as credible and employ it further.

In contrast to these overt modes of employment, a prediction can be employed as information for further planning, inferring, or processing. This could be a simple, deductive inference, or, in predictive analytics, a prediction could be input for another prediction task. Even prior to further planning, inferring, or processing, a prediction may be employed simply by holding onto it, as information to be relied upon later. A person may do this by coming to expect the predicted outcome, forming a mental state of belief or acceptance.[38] Indeed, after granting a prediction credibility, expectation formation commonly follows. The computational equivalent is simply storing the predicted value for subsequent reference.

[36] See Alfano 2013, 91.

[37] This mode of employment has sometimes been supposed essential to self-fulfillment. See Buck 1963, 360-362. Similarly, Merton emphasizes "the public definitions of a situation" (1948, 195). See Romanos 1973, 102-104, for a persuasive argument against restricting the notion of reflexive prediction to cases where the effects of the prediction are mediated by its dissemination.

[38] Forming a belief would be most likely when the prediction's credentials were tied to the persuasiveness of its justification. When the credentials are more local to the practical endeavor at hand, the person may not form a full-fledged belief, but perhaps something more like a context-relative acceptance. See Bratman 1992, 9-11. Oddly, Biggs (2009) defines SFPs in terms of beliefs not predictions, thereby obscuring the relationship between the two and ignoring cases where self-fulfillment of a prediction is not mediated by actual belief in the proposition predicted.

Again, in general, to employ a prediction is simply to rely upon or use it somehow, for further action or processing. For a prediction to become self-fulfilling, its employment is a basic and essential requirement. Without employment, a prediction is inert.

### 2.3 Third Element: Employment-Sensitivity

If we think of credentials as opening a door, then employment is a step across the threshold. On the other side lies the subject of prediction—the things the prediction is about, the system in which the prediction might be fulfilled.

Here we speak of systems, as opposed to particular objects, to capture complex interactions between employment and the outcome predicted, along with the broad range of factors that may mediate those interactions. SFPs do not always unfold by the prediction's employment operating directly on the objects or properties the prediction is about. For example, a teacher's employment of a prediction of student improvement may affect the student's performance in a roundabout, indirect way, due to the educational environment as well as the activities, actors, and processes within it.

Self-fulfillment of a prediction requires that the prediction's employment affect a system that is sensitive to such employment. In general, the very essence of predictive reflexivity is that the system where the objects of a prediction reside is sensitive to the employment of the prediction.[39] We can say that an SFP, or any reflexive prediction, depends on an *employment-sensitive* system—a system potentially affected by the way the prediction has been employed.

We distinguish two types of employment-sensitivity. Consider two ways a prediction about a particular student's success might be realized. In one scenario, a teacher employs the prediction by assigning more challenging exercises to the student. The student is sensitive to this employment: rising to the occasion, learning a lot, performing well on the exam, and receiving a high score. The system exhibits *substantive employment-sensitivity* with respect to that object: the object the prediction is about is affected by the prediction's employment.

In an alternative scenario, the teacher employs the prediction, not by adjusting her teaching or assignments, but simply by forming an expectation that this student will succeed. The teacher herself is then sensitive to this employment, to the expectation she formed, when she grades exams. Even though the student behaves exactly as if no prediction had been made or employed, the teacher assesses the student's work differently than she would have, maybe due to a confirmation bias. The student performs adequately, the teacher grades charitably, and the student receives a high score. The prediction is self-fulfilling due to how its employment affects the way success is registered. Thus, the system exhibits *interpretive employment-sensitivity*: the measure of the feature the prediction is about is affected by the prediction's employment.[40]

[39] Hence, a prediction can be shown to be not reflexive if the system is not sensitive to the employment of the prediction. According to this rationale, Richard Henshel (1982) attempts to circumscribe classes of social prediction that cannot be reflexive. Unfortunately, Henshel considers only a narrow range of possible employments, primarily communication, without considering reflexivity due to other kinds of employment and employment-sensitivity.

[40] Cf. Jussim 1989, 470. Jussim draws the same distinction between two ways that predictions might yield student success through teacher expectations. He reserves the term 'self-fulfilling prophecy' for cases involving substantive sensitivity, and describes interpretive sensitivity in terms of perceptual bias.

With substantive sensitivity, the object is affected; with interpretive sensitivity, the interpretation or measure of the object's properties is affected. A prediction of student success can be self-fulfilling either way. Whether a given prediction's self-fulfillment—or reflexivity more generally—can be due to interpretive sensitivity depends on the formulation of the prediction. Formulating a prediction about a student in terms of the student's *success* allows reflexivity through interpretive sensitivity because *success* may be relative to assessment. In grading essays, the teachers are, to a great degree, the arbiters of success and have some flexibility regarding what they deem successful. In contrast, consider a prediction that a student would answer correctly at least forty out of fifty multiple choice questions. Assuming the grading key is not in dispute, there is little room for interpretation, and little possibility of reflexivity through interpretive sensitivity. Nonetheless, such a prediction could still be reflexive through substantive sensitivity.

This distinction between two types of employment-sensitivity illuminates what may be counterintuitive about the predictive policing SFP: that more patrolling in a predicted hotspot may *increase* the crime rate there. Intuitively, more patrolling should decrease crime. However, if the *crime rate* is a function of arrests or police reports, then the system the predictions are about is interpretively sensitive to increased patrolling. Sending more officers to a predicted hotspot may not change the behavior of people in that area, but more of that same behavior may be noticed by police, yielding a higher crime rate.

These simple examples provide just a glimpse of the diverse ways a system may be sensitive to the employment of predictions about it. The more complex the system, the greater the potential for reflexivity.

## 2.4 Fourth Element: Realization of the Predicted Outcome

The three elements already described suffice for a prediction to be reflexive. They do not, however, suffice for a prediction to be self-fulfilling, for two reasons: First, self-fulfillment requires that the prediction actually be fulfilled, and, second, this fulfillment must be due to the prediction itself.

As already noted, not all reflexive predictions are self-fulfilling. Employment of a prediction may influence the outcome predicted without bringing it about. The outcome may even be thwarted, yielding a *self-defeating prophecy*.[41] Consider how this might work with predictive policing. As before, suppose a hotspot prediction is employed by sending additional patrols to that area. Unlike before, suppose the crime rate is somehow measured independently of the patrols, so the system lacks the interpretive employment-sensitivity described before. Rather, suppose there is substantive employment-sensitivity: When an area is patrolled more heavily, members of the policed population go elsewhere or change their behavior. Thus, more patrols would cause a *lower* crime rate in the area, and the prediction of a high crime rate would be a self-defeating prophecy. Thus, clearly there can be reflexivity without self-fulfillment.[42]

[41] Although there is widespread agreement about the adjective "self-fulfilling," several adjectives have been used for predictions that bring about their own falsity: "suicidal" (Merton 1948, 196), "self-defeating" (Merton 1968, 183), "self-frustrating" (Buck 1963, 359), "self-stultifying" (Grünbaum 1963, 370), and "self-refuting" (Alfano 2013, 95).

[42] More puzzling would be reflexivity without either self-fulfillment or self-defeat. In what sense would the prediction (or its employment) affect the outcome predicted? See Kopec (2011) for a probabilistic answer.

Among reflexive predictions that are actually fulfilled, we can distinguish two kinds of genuine SFPs, and then further distinguish these genuine SFPs from similar predictions which were fulfilled but not *self*-fulfilled.

In the most striking cases of SFPs, the predicted outcome would not have been realized if the prediction had not been made and employed. For example, suppose that a particular student would not have succeeded without a prediction of success and employment of that prediction, but, due to the prediction and its employment, the student *does* succeed. The SFP flips the outcome. Not only did the prediction bring about its own realization; its realization *depended* upon the prediction and its employment. Call an SFP with this characteristic, a *transformative self-fulfilling prophecy*.[43]

In contrast, with some SFPs, the realization of the outcome predicted is *due to* the prediction and its employment but does not *depend on* it.[44] Consider a prediction that a student will succeed and a teacher who employs that prediction by assigning challenging exercises. The student is sensitive to this treatment, rises to the occasion, and succeeds for that reason. However, suppose that if, on the contrary, the teacher had *not* employed the prediction, then the student would have developed other sources of motivation, and would have succeeded for that reason. As the scenario actually played out, the teacher's employment of the prediction was *operative* in producing the outcome predicted, but that outcome would have been realized even without the prediction and its employment. Call this an *operative self-fulfilling prophecy*.

Transformative and operative SFPs are both genuine SFPs because, not only is the predicted outcome realized, the explanation of its realization—the mechanism or process that brings it about—lies in the interaction between the prediction's employment and an employment-sensitive system. An operative SFP is a genuine SFP, in the same way that a careless bonfire is the cause of a forest fire, even if lightning soon would have ignited a similar blaze. Counting operative SFPs as genuine SFPs is justified by a focus on the prediction's role in its own fulfillment. Furthermore, in practice, it is often easier to identify self-fulfillment than to say whether it was transformative or merely operative.[45] That is because the difference between transformative and operative depends on what would have happened had the prediction not been employed, and evaluating such counterfactuals can be challenging.

In contrast to the two types of genuine SFPs, we can distinguish predictions that merely ensure the outcome predicted without actually bringing it about. For example, imagine a student predicted to succeed, but who is also robustly talented and will perform very well, regardless of how she is treated. Now, recall the version of the student success SFP involving interpretive sensitivity: The teacher employs a prediction of success by forming an expectation that the student will succeed, which affects her disposition to grade charitably. But suppose that this student's work is excellent by any standard, and the teacher accordingly awards a high score, without her grading bias ever being triggered. In this situation, the predicted outcome is ensured or secured by, yet produced independently of, the prediction and its employment. We call such a prediction a *self-securing prophecy*.

Self-securing prophecies are similar to operative SFPs in that the prediction does not change whether the predicted outcome is realized: If the predictions were not employed at all, the predicted outcomes still would be realized. But the key difference remains: With

[43] Kopec (2011, 1252) labels reflexive predictions that are transformative *strongly reflexive predictions*.

[44] Here we hope to draw an intuitive distinction, while sidestepping complexities in the relations among causation, explanation, and counterfactual dependence.

[45] In other work, we highlight the epistemic significance of operative SFPs and the danger of nonchalance about them. See Mertens et al. 2022.

an operative SFP, as with a transformative SFP, but not with a self-securing prophecy, the prediction and its employment actually bring about the predicted outcome.

### 2.5 Definition of Self-Fulfilling Prophecy

We have described four elements of predictions that are self-fulfilling: credentials, employment, employment-sensitivity, and realization. From these elements we can synthesize a definition:

> A *self-fulfilling prophecy* is a prediction, treated as credible enough to be employed, with its outcome realized due to how the employment of the prediction affected a system sensitive to such employment.

When the four elements of the account are considered as criteria or conditions of self-fulfillment, we find some redundancy. After all, a prediction cannot be genuinely employed as a prediction unless it has been granted credibility. Furthermore, a prediction cannot be realized due to its employment unless the relevant system is employment-sensitive. Eliminating the redundancy, we are left with a less informative but extensionally equivalent definition, which we might consider a minimal definition: A self-fulfilling prophecy is a prediction with its outcome realized due to how it was employed. Technically correct though the minimal version may be,[46] the informativeness of the full definition makes it preferable for most purposes.

Our definition—in both the full and minimal versions—departs from Merton's classic definition in a few respects. First, regarding the basic ontology, our definition regards a self-fulfilling prophecy as a *prediction of an outcome*, rather than a *definition of a situation*.[47] Second, since it recognizes operative SFPs, our definition does not require that a false claim become true, or even that the outcome would not have been realized without the SFP.[48] Finally, our definition does not presuppose that the employment of the prediction must be its publication or communicative dissemination.

With this descriptive account of SFPs in place,[49] we are now in a position to examine the significance of SFPs.

## 3 What is Wrong with Self-fulfilling Prophecies?

### 3.1 Unintended Outcomes due to Process Mistakes

Our first step toward a critique of SFPs is an explanation of how and why typical SFPs produce unintended consequences. When predictions of student improvement are

[46] This minimal definition is roughly Romanos' definition of reflexive prediction but restricted to self-fulfilling predictions (1973, 106).

[47] Merton's peculiar description of an SFP as a "definition of the situation" instead of a "prediction" manifests the influence of sociologists William I. Thomas and Dorothy Swain Thomas and the so-called Thomas Theorem, which says, "if men define situations as real, they are real in their consequences" (Thomas and Thomas 1928, 572).

[48] Merton's definition has been criticized on this point. See Krishna 1971, 1104-1105.

[49] A fuller account would cover *collective self-fulfilling prophecies*, which rely on multiple instances of prediction and employment. Bank runs are the classic case of collective SFPs (Merton 1948, 194-196; Diamond and Dybvig 1983). Collective SFPs are also found in socio-technical systems, arguably exemplified by Moore's Law (Van Lente 2000, 59).

self-fulfilling, it is not the intention of teachers who employed those predictions that a few students flourish while the rest comparatively languish, but those are the outcomes. Moreover, in any ordinary (i.e., not experimental) context where educators or administrators make self-fulfilling predictions of student success, they do not *intend* the outcomes they predict. Similarly, the intention of developers or operators of predictive policing software is not to skew crime rates. Nor is that outcome intended by officers who employ the predictions. But, when the predictions are self-fulfilling, that is the result.

The outcomes of these SFPs are unintentional in that none of the constituent actions—in either prediction or employment—is performed with the intention to bring about the predicted outcome as such. The goal of prediction is to identify likely outcomes despite unknowns, not to steer interventions to bring about those outcomes. The goal of employment is to proceed with a practical endeavor by relying on predictions, enabling the employer to treat otherwise uncertain outcomes as givens, not to actually accomplish those outcomes. Thus, the general aims of prediction and employment are at odds with intentions to bring about the outcomes predicted.

Although all actions have unintended consequences, SFPs are unusual in that their unintended consequences are the very outcomes predicted. This can happen when predictors and employers fail to recognize potential reflexivity, conceiving of their practical endeavors without regard to the ways that employing predictions may affect employment-sensitive systems. We will say that, for a predictor, employer, or other party involved in a practical endeavor, to make a *process mistake* is to treat as uninfluenced by the endeavor, features of a system that are actually subject to its influence. More specifically, and now focusing on the role of prediction, to make a *prospective process mistake* is to treat as largely imperturbable and subject to prediction, features that are sensitive to ways the prediction might be employed. Although retrospective process mistakes will become relevant later, we now focus exclusively on the prospective variety because they explain the unintended outcomes of SFPs.

The key point is that once a process mistake has been made—and the tasks of prediction and employment have been delineated according to it—mechanisms of reflexivity may operate without the intention, or even the awareness, of the parties involved. Even in the case of an operative SFP, where the predicted outcome would have been realized anyway, the predictors and employers in the grip of a process mistake suppose this realization will be independent, not brought about through their own actions. Thus, SFPs arising in the wake of process mistakes produce outcomes that, although expected, were never intended.[50]

In the SFP of student success, the teachers fell into a process mistake by considering a student’s path of improvement as predictable, intrinsic to the student and indicated by a test, rather than something heavily influenced by the teachers’ actions. In employing this prediction, the teachers selectively swayed student success without intention or awareness. Of course, in the actual Oak School experiment, the teachers’ process mistake was induced by researchers who portrayed the students as more predictable and less sensitive to influence than the researchers actually expected them to be. Thus,

[50] Biggs (2009, 311) claims that the occurrence of process mistakes (both prospective and retrospective) is an essential criterion of SFPs, thus building the problems of SFPs into the very definition of SFP. In contrast, we offered a purely descriptive definition of SFP, in terms of which we aim to explain the distinctive role of SFPs within a nexus of mistakes and other problems.

the researchers did not make a process mistake, but rather aimed to experiment with the effects of one.

In the predictive policing SFP, the parties involved treat *crime rate* at a particular time and place as subject to prediction, independent of the activity of police officers at that time and place. Considering the crime rate this way is a process mistake, because the crime rates registered are directly affected by police activity. While making this mistake, the police can skew the crime rate without intention or awareness.

For one more example, return to the SFP in credit scoring, noted in our introduction. Credit scoring agencies, as predictors, and the lending institutions, as employers of the predictions, make a process mistake by treating a person's status as a "credit risk" as predictable, rather than as affected by how the person is classified. In the grip of this process mistake, lenders and credit scorers may unintentionally and unknowingly make a person more of a credit risk.

### 3.2 Process Mistakes, Complex Endeavors, and Unaccountability

The more complex the practical endeavor—especially if there are many diverse and not fully anticipated ways predictions may be employed—the easier it is to make a process mistake. Nevertheless, we may hope that conscientious predictors and employers will catch process mistakes as their endeavors proceed. After all, a familiar fact about practical activities is that we seldom understand exactly how they will unfold until they are actually underway. Once an endeavor is underway, unforeseen connections between predictions' employments and the predicted outcomes may become more apparent. A reasonable prescription, then, for catching process mistakes is to increase coordination between predictors and employers, with some responsible parties keeping the whole system in view.

Unfortunately, the trend appears to be in the opposite direction. Indeed, predictive analytics is promoted as a means to cope with complex endeavors by modularizing prediction tasks, increasing the separation between prediction and employment. A *Harvard Business Review* piece called "Making Advanced Analytics Work for You" blithely explains, "The key is to separate the statistics experts and software developers from the managers who use the data-driven insights."[51] The authors offer a telling illustration of this strategy:

> One large industrial company, for instance, sought to better forecast workforce needs to reflect local market variations. Historically, as the company had tried to keep labor costs low, it had often found itself short-staffed in some markets, leading to significant overtime costs and service snafus. To remedy the problem, the company convened a small working group of analysts and IT programmers who developed a series of predictive models that forecast workforce availability on the basis of factors such as vacation time, absenteeism, and work rules in labor contracts. The models incorporated millions of new data points on thousands of employees across dozens of locations. But rather than providing managers with reams of data and complex

[51] Barton and Court 2012, 82.

> models, they created a simple visual interface that highlighted projected workforce needs and necessary actions.[52]

We see here a division of cognitive labor, with the data scientists tossing their predictions over the wall to the managers who will use them, diminishing rather than improving coordination between predictors and employers. Hence, any mistakes and reflexive predictions can easily go unnoticed.

In complex endeavors, especially when prediction and employment are separate, process mistakes make it challenging to hold anyone accountable for the consequences of reflexivity. Consider some instance of reflexive prediction, and imagine asking the predictor to justify bringing about the outcome predicted.[53] The predictor might reply that she was just engaged in prediction, and it is no part of that activity to influence the things predicted. The employer would be equally befuddled, because her choices were based on an expectation of the outcome, not anticipation of influencing it.[54] Each responds from the perspective of a process mistake. This makes it awkward, and perhaps not entirely reasonable, to ask either one to justify the outcome. The difficulty is not merely a problem of many hands,[55] though it may be that too. The underlying difficulty is that, in the grip of process mistakes, the parties' conceptions of prediction and employment orient each party to be not straightforwardly answerable for bringing about the outcome.[56]

Recognition of this impediment to accountability renders sensible the feeling of accomplishment sometimes evinced by commentators when pointing out SFPs. Flagging what might otherwise be a subtle evasion of accountability is a benefit of social criticism. However, just flagging deficient accountability is not the same as rectifying it, and there is no quick fix. When we contemplate how to actually improve accountability for reflexive predictions, a comparison to the problem of implicit bias is instructive. Like reflexive predictions, implicit biases often produce undesirable consequences, including perpetuation of patterns of inequality. Furthermore, as with reflexive predictions, the consequences of implicit bias are often "unintentional, unendorsed, and perpetrated without awareness."[57] However, even if people are not fully accountable for their biases, they can still strive to become increasingly aware of them and how they function.[58] We recommend a similar forward-looking perspective—supported by critique and increased awareness—for improving accountability for SFPs and other reflexive predictions.

[52] Barton and Court 2012, 82-83.

[53] Here we are thinking of accountability in terms of Angela Smith's notion of answerability. An agent is *answerable* for bringing about some outcome if it is intelligible to ask her for reasons that might justify having brought it about (Smith 2015, 103).

[54] Cf. Rubel et al. 2019, 1021-1022, similarly invoking answerability to analyze failures of accountability in algorithmic decision-making.

[55] See Nissenbaum 1996, 28-32.

[56] With both transformative and operative SFPs, the parties involved are the unwitting causal agents of the outcome. Arguably, moral responsibility for the outcomes, if any there be, is less with operative SFPs, since the outcomes would have been realized anyway. We note, however, that the parties' unwittingness of their causal responsibility, and hence the impediment to answerability, is equal for operative and transformative SFPs.

[57] Holroyd et al. 2017, 2.

[58] Saul 2013, 55.

### 3.3 Unmistaken and Intentional Self-fulfilling Prophecies

We have just highlighted two prominent features of our standard cases of SFPs: They occur due to process mistakes, and they produce unintended outcomes. Although most SFPs share these two features, we also must recognize those special cases of SFPs that lack one or both of these features.

First, consider SFPs that lack both features: Some party involved makes *no* process mistakes and intentionally helps bring about a predicted outcome by way of an SFP. This definitely can happen. When a participant in a practical endeavor recognizes the potential for reflexivity, and especially if she recognizes the potential for others to fall into a process mistake, this opens the door to an intentional SFP. For instance, a predictor may disingenuously offer a prediction, anticipating that it will be employed as an ordinary non-reflexive prediction. In other words, a predictor may lure others into a process mistake. Indeed, this is exactly what was done, for experimental purposes, in the Oak School study. Employers too may be intentional perpetrators of SFPs: An employer may intentionally employ a prediction to bring about the outcome predicted. We can imagine a devious police captain or officer, perhaps wanting certain hotspot predictions to be borne out, demanding extra-vigilant patrolling, while everyone else remains in the grip of process mistakes.[59]

Such cases of intentional SFPs, as effected by the predictor, employer, or whichever party, involve making use of the potential for reflexive prediction, while depending on others to make process mistakes. However, this can go only so far. Enlarging the group privy to the reflexivity—those enabling or allowing it, not being fooled by the would-be process mistake—diminishes the very coherence of intentional reflexivity. When everyone involved recognizes that a prediction has been made based on the assumption that it will be self-fulfilling, or treats the prediction as credible only on the condition that its employment will produce the predicted outcome, then the “prediction” does not function as a genuine prediction. When the “prediction” is only a means to realize the outcome predicted, *prediction* is no longer the right concept.[60] At that point, the “prediction” is not an attempt to reflect an independent (possibly future) reality, but rather a step toward shaping that reality. Hence, genuine self-fulfilling predictions cannot be deliberately enacted by a *fully* aware and informed group. When an SFP is enacted intentionally by some party, this depends on others remaining confused or oblivious about it.[61]

Second, and in contrast to cases in which an SFP is intended by someone *not* making a process mistake, try to imagine an SFP due to a process mistake but in which the outcome *was* nevertheless intended. The person would, somehow, have to be mistaken in thinking the outcome was beyond the influence of her endeavor, and yet nevertheless intend the endeavor to bring about that very outcome. In essence, she would need to intend for something to happen in a way that she assumed it could not. This would be odd, if not downright incoherent, and it is difficult to imagine an example. Hence, we set aside this class of cases.

[59] If, while intentionally inducing reflexivity, a predictor or employer denies her agency in realizing the relevant outcome, she can be accused of agency laundering (Rubel et al. 2019, 1021-1024).

[60] We can appeal here to the thought that a prediction’s direction of fit is *word fits world*, not *world fits word*. Cf. Searle 1976, 3-4.

[61] So, when a single person both makes and employs the prediction, it is difficult to have an intentional SFP. Absent some self-deception, intentional self-fulfillment would require the person not to treat the “prediction” as a genuine prediction. Hence, we disagree with some popular self-help literature that describes optimistic “fake it ’til you make it” strategies as SFPs. We are not criticizing such strategies, just noting that it is misleading to call them SFPs.

Third and finally, consider cases of SFPs which are neither intentional nor due to a process mistake. An example from medical practice is illustrative.[62] Consider doctors treating a patient who is in a coma. On the basis of neurological tests, the patient is predicted to have a "poor outcome," which could be death or a severe disorder of consciousness. In light of this prognosis and the patient's interests, the medical team decides to withdraw life support, allowing the patient to die. Death is the realization of a poor outcome for the patient, and so the prognosis is self-fulfilling.

This example need not involve any process mistake: All parties recognize both that the prognosis was intended to reflect the patient's medical condition in relation to available treatments, and that the employment of a poor prognosis will be the decision to withdraw life support. Nor was the outcome intentional: No one intended a poor outcome per se; the intention, rather, was to avoid the *most* undesirable outcomes. Similar cases are possible when a credible prediction of a particular outcome calls for planning for that outcome in ways that everyone involved recognizes may determine the outcome. The prediction may then, as in this medical example, be employed to expedite the outcome, choose the form it takes, or control (or avoid) aspects of the process that will yield it. This need not involve any process mistake or intention for that outcome per se.

In contrast to our paradigm cases, the sorts of SFPs described throughout this section do not arise from process mistakes. Rather, at least some party involved exhibits some *process awareness*, as we might put it, which affords some measure of accountability for the outcomes produced. Nevertheless, these SFPs are still subject to the further problems we will now describe.

### 3.4 Errors Without Error Signals

We have already seen that, when an SFP is due to a process mistake, it may be that no one is well-positioned to answer for the outcome produced. However, regardless of whether an SFP came from a process mistake, accountability for its outcome is elusive for a further reason: There may be no call for anyone to answer for it at all. Since, with an SFP, the outcome eventually observed is the outcome originally predicted, anyone who grants credibility to the prediction in the first place will not be surprised by the outcome. Thus, when a prediction is self-fulfilling, any mistakes involved will be *errors without error signals*.[63] Whatever mistakes may have taken place, an SFP is nevertheless a true prediction, and its truth deters scrutiny.

Thinking about a scenario will be helpful. Imagine two police captains, each in charge of different precincts in a large municipality that has just hired some data scientists to implement predictive policing. Each captain will receive daily predictions from the data scientists about which locations will have high crime rates. Suppose that both captains will treat these predictions as credible. However, the two captains have different plans for employing the predictions: One captain—call her *Captain Deployment*—aims to use the predictions in the usual way, deploying more officers to patrol areas predicted to have higher crime rates. The other captain—call her *Captain Rotation*—opts for an alternative. Instead of directing greater attention to hotspots, her goal is to distribute the challenging

[62] The following sort of case has been discussed in biomedical ethics (Wilkinson 2009) and in debates about neuroprognostic research (Geocadin et al. 2012, 979-980). See also Mertens et al. 2022.

[63] Regarding a similar problem of errors without error signals, see King 2020, 29-30.

work of patrolling hotspots more evenly among the officers. So, she rotates different officers through the projected hotspots on different days, without actually adjusting how heavily any area is patrolled. Finally, suppose that, unbeknownst to the data scientists and the police, a software update introduces a bug, making the hotspot predictions essentially random. The captains then use the system for a month.

For Captain Deployment, the month unfolds as our earlier sketch of predictive policing would lead us to expect: Many of the hotspot predictions are transformative SFPs because having more officers in an area causes extra arrests and police reports in that area. So, at least in aggregate, the predictions are borne out, with more crime registered in the predicted hotspots than in otherwise comparable areas. In contrast, for Captain Rotation, the predictions are not self-fulfilling. Overall, the crime rate in the predicted hotspots is no different than it would have been, and not significantly different from the crime rate in other areas.

In this scenario, the two captains receive equally faulty predictions. However, only Captain Deployment's employment introduces significant reflexivity and skews crime rates. This is due to a process mistake, treating the *crime rate* as something holding independently and subject to prediction, even though it is susceptible to influence by the very mode of employment she intends. In contrast, Captain Rotation does not introduce reflexivity or skew crime rates. Perhaps she remains prone to some kind of process mistake, but her actual employment of the predictions does not result in their fulfillment.

Now, to see the special problem with SFPs, consider the captains' perspectives at the end of the month. We might regard Captain Rotation as epistemically fortunate. At the end of the month, she is surprised that the crime rates do not match the predictions. For her, the mismatch of predictions and outcomes is an *error signal*—an indication that something has gone awry—which prompts her to look for mistakes and make corrections.[64] She might reconsider how she grants credibility to the data scientists' predictions. Perhaps she stops relying on those predictions altogether, or simply tells the data scientists that something is wrong. In general, an error signal can alert the participants in a practical endeavor that the endeavor is not well-served by the predictions employed. Thus, error signals create opportunities for adjustment.

Captain Deployment, by comparison, is not so fortunate. Many of the hotspot predictions she employed were self-fulfilling. Of course, not all the predictions came true, since the presence of officers does not guarantee that crime will be recorded. Still, she observes a general alignment between hotspot predictions and areas with high crime rates. So, Captain Deployment receives no error signal, or, at most, a much weaker error signal than Captain Rotation receives. Indeed we can imagine her pleased to see the predictions largely borne out, never led to suspect either the randomness of the predictions or her process mistake. Furthermore, though her actions actually elevated the crime rate registered in the predicted hotspots, neither she nor anyone else will likely be called to answer for that. In general, a consequence of an SFP, beyond just the production of the predicted outcome, is the concealment of process mistakes and other faults that may have affected the predictions.[65] As

[64] Thus, we conceive of error signals in practical endeavors as roughly analogous to prediction error signals in predictive coding models of neural processes.

[65] Along these lines, Robin Hogarth mentions SFPs as a potential cause of a *wicked learning structure*, a situation in which, due to missing or misleading feedback, it is difficult for a person to learn from experience (2001, 84-90).

a result, no one will be held accountable for the outcome or even prompted to notice the problems.

It is crucial to recognize the contrast here between SFPs and other reflexive predictions. SFPs are the exceptional case, posing a special problem because they cover up errors. Ordinarily, reflexivity will be an interfering or perturbing factor, *reducing* the likelihood that a predicted outcome will be realized. Hence, process mistakes and the ensuing reflexivity typically *do* produce error signals. This is especially striking in cases of self-defeating prophecies, where the reflexivity thwarts the outcome predicted.

For illustration, suppose there is a third captain—call her *Captain Avoidance*—who, in the interest of reducing officers' job-related stress, employs the hotspot predictions by sending officers *away* from predicted hotspots. The predictions Captain Avoidance employs are self-defeating, because they leave few officers around to register crime in the areas where it is predicted. After a month of employing hotspot predictions, Captain Avoidance, like Captain Rotation, notes that the predictions were not accurate. Thus, the recognition of false predictions is an error signal for Captain Avoidance, alerting her that something has gone wrong, perhaps putting her in a position to notice her process mistake. Whereas Captain Deployment is not called to notice or answer for how her actions skewed crime rates, Captain Avoidance is prompted to take responsibility and adjust.

What distinguishes SFPs from other reflexive predictions is that, by bringing outcomes into alignment with predictions, they avoid error signals. Process mistakes responsible for the reflexivity, as well as any other mistakes in prediction, remain hidden. Thus, SFPs keep us from catching—and learning from—our mistakes.

## 3.5 Specious Quality Assurance

Regarding our example of the police captains, one might reasonably argue that Captain Deployment and Captain Avoidance should have anticipated how their uses of the predictions would generate reflexivity. Even without error signals, they could have, perhaps should have, caught their process mistakes. After all, the mechanism of reflexivity is fairly straightforward. All of that may be correct.

However, other cases of reflexivity are more subtle, and may offer little hope of catching errors, except by way of error signals yielded by evidently false predictions. A sufficiently complex system or endeavor may be inscrutable, and careful attention to a track record of predictive accuracy may be the best quality assurance available. But, as we have seen, SFPs defy this mode of quality assurance. A history of operative and transformative SFPs might be reckoned a uniform success, though it should have been assessed as spotty at best. Hence, we have a problem: With SFPs in a complex endeavor, the parties involved may *appear* to exercise responsible, diligent quality assurance, maybe even the best quality assurance available—by fastidiously checking predictions against actual outcomes—and yet still fail to catch and address mistakes.

To see why this problem is pressing, keep in mind that practical prediction is often deemed desirable in complex situations that are not well-understood, situations where reflexivity would not be obvious. Practical endeavors may favor reliance on predictions despite absence of any clear idea about what mechanisms would explain the outcomes predicted. This is unsettling for those of us who find comfort in knowledge and certainty, but it comes with the territory: We rely on predictions where certainty is unavailable.

Once again, trends in predictive analytics exacerbate the problem, this time by increasing the volume of potentially useful predictions unaccompanied by explanations of predictive success. Predictive analytics based on machine learning prioritizes and excels at discovering patterns in data,[66] leaving the identification of causal or explanatory mechanisms secondary.[67] Hence, significant progress in predictive accuracy may be achieved without parallel progress in understanding the objects of prediction.[68] This often leaves statistical measures of accuracy—based on checking the truth of predictions once outcomes are apparent—as the only viable means of quality assurance. But this is precisely the check that SFPs vitiate.

Of course, sometimes it may be possible to run controlled studies, specifically studies with control groups in which the relevant predictions are not employed. If so, this extra step could reveal SFPs and other reflexive predictions. Indeed, it is hardly an exaggeration to say that the entire point of running blinded randomized controlled trials is to screen off the effects of reflexivity. However, practical endeavors are often not scientific endeavors. Practical endeavors often call for predictions—and indeed call for those predictions to be employed—before more rigorous experimentation is feasible. When practical urgency will not permit scientific rigor, retrospective quality assurance will remain the best (really, the only) option, and then SFPs will get a pass.

The result is a further impediment to accountability. Earlier we saw how process mistakes complicate accountability by misaligning intentions and outcomes. Then we observed how SFPs, because they produce no error signals, prompt no calls for accountability. The additional problem we have just seen goes a step further. It is not just that SFPs do not reveal flaws. It is that they can yield a specious appearance of diligent quality assurance, discouraging reexamination and encouraging perpetuation.[69]

## 3.6 Retrospective Process Mistakes and Predictive Feedback Loops

We have just seen that unaccountable processes may emerge when apparently diligent regimens of retrospective quality assurance fail to flag flaws. Further mistakes ensue if assessments of predictive success encourage spurious commendations of flawed processes. These further mistakes drive predictive feedback loops. Through predictive feedback loops, mistakes and biases in sources of predictions may be repeatedly projected, often with increasing intensity, out into the world.

Return once more to our example of the police captains. Consider the perspective of Captain Deployment after a month of reliance on the hotspot predictions. Having checked and confirmed that crime rates were high where predicted, and not recognizing that the predictions' accuracy was due to her own actions, it is natural for her confidence in the predictions to swell. In giving *greater* credibility to a source of predictions, even though she influenced the accuracy of past predictions from that source, she extends her original prospective process mistake.

[66] Dhar 2013, 66.

[67] Lipton 2018, 37. For a comparison of the deliverances of explanatory statistical models and predictive analytics, see Shmueli and Koppius 2011, 554-557.

[68] Research on *explainable AI* responds to this deficit between high predictive accuracy and the low intelligibility of the underlying predictive models. See Adadi and Barrada 2018, for an introduction and overview.

[69] As Tal Zarsky explains regarding financial credit scoring: "These dynamics will not be self-corrected, as they are misunderstood by the analysts studying the feedback of the scoring practices as mere reassurance of the scoring system's precision." See Zarsky 2014, 1405.

Earlier we said that, in making a process mistake, someone regards as independent features of a situation that are actually under the influence of the practical endeavor at hand. And we examined how prospective process mistakes give rise to reflexive predictions. Now we can now see what a *retrospective process mistake* would be: to treat as providing independent credentials or grounds for further predictions, features of a system that may have been affected by ways previous predictions were employed.

Captain Deployment makes a retrospective process mistake in how she thinks about the credentials of the predictions at her disposal. She grants increasingly more credibility to the predictions by the data scientists, even though the explanation of predictive success is her employment of the predictions, not their quality. Similarly, if the data scientists feed the skewed crime rates from Captain Deployment's precincts back into their predictive model, that also would be a retrospective process mistake. In general, a retrospective process mistake occurs whenever the basis or credentials for new predictions depend on unacknowledged reflexivity in prior predictions.

Even though retrospective process mistakes are most common in the aftermath of prospective process mistakes, they can occur even without them. Even for those atypical SFPs not caused by prospective process mistakes, ongoing unaccountability—as the practical endeavor coasts along smoothly without any error signals—may coax employers of predictions into retrospective process mistakes. To extend our medical example described above, suppose that a particular prognostic test has been the basis of many predictions of poor patient outcome, and, in each case, life support has been withdrawn. Then, in each case, the death of the patient eliminates the possibility of an error signal for an erroneous prognosis. The resulting lack of negative feedback *should* be understood as exactly that: a lack of feedback. As such, it does not give any indication of the quality of the test.[70] If, however, that lack of negative feedback encourages the physicians to become more comfortable or confident with the relevant test, then they have fallen into a retrospective process mistake.

An especially troubling characteristic of retrospective process mistakes is that they can serve to reinforce even the most arbitrary credentials of predictions. Note that, with SFPs, the explanation of predictive accuracy may be entirely divorced from the credentials that motivated the employer to use the prediction. Nevertheless, recognition of that predictive accuracy may seem to commend those very credentials. Thus, through an SFP, *any* credentials can be seemingly validated and thereby reinforced. This includes highly objectionable credentials, including those based on discriminatory biases or prejudices.

When SFPs reinforce the credentials of a class of predictions, motivating continued or intensified reliance on those predictions, yielding more SFPs, this amounts to a *credibility feedback loop*. We can imagine a credibility feedback loop occurring with any of the examples of SFPs we have considered. It would simply involve the employer repeatedly interpreting the outcomes of prior SFPs as indicators of predictive success, and consequently becoming increasingly disposed to grant credibility to and rely on similar predictions. With each iteration, reliance on such predictions may become more expansive.[71]

A credibility feedback loop is a relatively tight loop, with just the credentials and employment of predictions iteratively reinforced. But SFPs may also produce wider feedback loops that encompass the basis and content of future predictions. This occurs when

[70] For elaboration, see Mertens et al. 2022.

[71] Regarding feedback loops like these in predictive policing, see Brayne 2017, 998. See also Richardson et al. 2019, 43-45, for a discussion of similar "confirmatory feedback loops" which reinforce the stereotypes and assumptions used to justify disproportionate policing of marginalized communities.

the outcomes of prior SFPs are taken as data to inform new predictions. Along these lines, in an influential paper on machine learning and predictive policing, Danielle Ensign and colleagues present the following critique:

> Once police are deployed based on these predictions, data from observations in the neighborhood are then used to further update the model. . . Since such [observations] only occur in neighborhoods that police have been sent to by the predictive policing algorithm itself, there is the potential for this sampling bias to be compounded, causing a runaway feedback loop.[72]

We call this wider kind of loop, which integrates observed outcomes of prior SFPs within a predictor's basis for future predictions, hence affecting the content of those predictions, an *outcome feedback loop*.[73] An outcome feedback loop involves the retrospective process mistake of treating data tainted by prior reflexivity as though it were suitable evidence for further predictions that are not conditioned on such reflexivity. As with the process mistakes associated with credibility feedback loops, these retrospective process mistakes may occur even if the original SFP was not due to a prospective process mistake.

The primary difference between credibility feedback loops and outcome feedback loops is in whether the apparent evidential basis for new predictions is affected. A credibility loop operates primarily on the employer, strengthening the employer's disposition to treat certain predictions as credible, but the predictions themselves may continue just as before. Such a loop may indeed bring about the predicted outcomes repeatedly and increasingly, but these outcomes are not necessarily fed back into the basis for prediction. If, however, data about these outcomes *are* fed back into the basis for prediction, thus affecting the content of future predictions—e.g., the magnitude of some phenomena or the classification of certain items—then we have an outcome feedback loop. Thus, the distinguishing feature of an outcome feedback loop is that it is mediated by the actual effects of SFPs on an employment-sensitive system. As predictors iteratively pick up and then project the patterns shaped by prior SFPs, the employment-sensitive system is iteratively transformed.

Both types of predictive feedback loops perpetuate and amplify SFPs, systematically projecting errors and biases into the world. Furthermore, through feedback loops, especially outcome feedback loops, the consequences of SFPs may ramify. For example, progressively increasing the policing of particular communities affects those communities in myriad consequential ways. Similarly, sorting students into inappropriate categories affects their lives more broadly. Thus, feedback loops, especially outcome feedback loops, emerge as drivers of social transformation—drivers which become more potent as prediction is automated.

## 3.7 Conclusion: Problems with True Predictions

Our critique has woven together several strands. Roughly, and oversimplifying a bit, we can say: Failures to recognize how predictions might be reflexive, along with the way SFPs cloak their defects, yield unaccountability and then feedback loops.

[72] Ensign et al. 2018, 2. For similar descriptions, see Lum and Isaac 2016, 16; Ferguson 2017, 1178.

[73] Outcome feedback loops are a recurring target of criticism in Cathy O'Neil's popular book documenting the harms of unscrupulous predictive modeling. See O'Neil 2016, 6-7, for general comment, and 2016, 87, on predictive policing. What Zarsky, following Citron and Pasquale, calls a "negative spiral" is a kind of outcome feedback loop. See Zarsky 2014, 1405-1408, for an illuminating discussion, touching on both credibility and outcome feedback loops. Similarly, what Bernard Harcourt calls a "ratchet effect" is an outcome feedback loop (2007, 147-160).

To summarize with a bit more nuance: We have examined how SFPs, and reflexive predictions more generally, typically arise due to conceiving of predictions without sufficient regard to the relations among prediction, employment, and employment-sensitive systems. Prospective process mistakes were central in that discussion. We also have examined the distinctiveness of SFPs among reflexive predictions, in the way that they elude scrutiny, hiding mistakes by taking cover in their truth. Central to that discussion was the observation that SFPs generate no error signals. Each of these discussions supported and illuminated the suspicion that SFPs undermine accountability for the outcomes they produce. Next, we saw how SFPs circumvent ordinary modes of quality assurance for predictions, thus permitting faulty processes to proceed despite their faults. As a faulty process proceeds, producing a spurious history of predictive success, SFPs may give rise to new mistakes: retrospective process mistakes. Finally, we observed how these new mistakes encourage repetition of SFPs, yielding feedback loops, through which the consequences of SFPs are multiplied.

These problems with SFPs are due to failure on the part of those involved in a practical endeavor to respect the complex relations among predictions, employment of those predictions, and employment-sensitive systems. We should expect, then, that SFPs are most treacherous in practical endeavors structured to modularize the role of prediction, treating prediction merely as a means of coping with uncertainty, dismissing its dependencies and impacts on other parts of the practical endeavor and the larger system in which the endeavor is situated.

In laying out the "problems," "mistakes," and "failures" involved in SFPs, we have not limited ourselves to just a moral critique or just an epistemological one. Rather, we have drawn out and developed a collection of related lines of substantial criticism. Some of these have a more moral complexion, such as the points about unaccountability for practical consequences. Others, such as the failure to notice and learn from mistakes, are more epistemological. However, rigidly sorting problems this way may ultimately hinder inquiry into the ethics of prediction. Prediction serving a practical endeavor is, after all, an attempt to grasp the world in order to influence it. So, epistemological and moral issues are always intertwined.

There is a great deal more to say about the complex ethics of prediction. In closing, we simply wish to stress that adequate assessments of predictions cannot be concerned merely with their truth, but must encompass the structure, operation, and consequences of the practical endeavors that depend on them. An exaggerated estimation of the value of predictive accuracy blinds us to self-fulfilling prophecies, the mistakes that accompany them, and their often undesirable effects on the world.

**Acknowledgements** For helpful input and feedback, we wish to thank Berend Alberts-de Gier, Sander Beckers, Maren Behrensen, Marianne Boenink, Justin D'Arms, Dai Heide, Olya Kudina, Jonne Maas, Alan Rubel, Mark Ryan, David Skillicorn, Brandt van der Gaast, and Nils Wagner. We are grateful for discussions with audiences at the University of Twente, Northeastern University, and Queen's University, where earlier versions of this article were presented. We also benefited from discussions of parts of this project at several conferences, including the Aachen Emerging Technology & Evolving Responsibility Workshop, Computer Ethics—Philosophical Enquiry, and the OZSW Annual Conference. Early work on this article was done with support from the Netherlands Organisation for Scientific Research (NWO), under project numbers 652.001.003 (King) and 313-99-309 (Mertens).

## Declarations

**Conflict of Interest** The authors declare that they have no conflicts of interest.

## References

Adadi A, Berrada M (2018) Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access 6:52138–52160

Alfano M (2013) Character as moral fiction. Cambridge University Press

Barton D, Court D (2012) Making advanced analytics work for you. Harv Bus Rev 90(10):78–83

Berg N (2014) Predicting crime, LAPD-style. The Guardian. https://www.theguardian.com/cities/2014/jun/25/predicting-crime-lapd-los-angeles-police-data-analysis-algorithm-minority-report. Accessed 15 Nov 2021

Biggs M (2009) Self-fulfilling prophecies. In: Bearman P, Hedström P (eds) The Oxford handbook of analytical sociology. Oxford University Press, Oxford, pp 294–314

Bratman ME (1992) Practical reasoning and acceptance in a context. Mind 101(401):1–15

Brayne S (2017) Big data surveillance: The case of policing. Am Sociol Rev 82(5):977–1008

Buck RC (1963) Reflexive predictions. Philosophy of Science 30(4):359–369

Carmel YH, Ben-Shahar TH (2017) Reshaping ability grouping through big data. Vanderbilt J Entertain Technol Law 20:87

Christakis NA (1999) Death Foretold: Prophecy and Prognosis in Medical Care. University of Chicago Press, Chicago

Citron DK, Pasquale F (2014) The scored society: due process for automated predictions. Wash Law Rev 89:1–33

Dhar V (2013) Data science and prediction. Commun ACM 56(12):64–73

Diamond DW, Dybvig PH (1983) Bank runs, deposit insurance, and liquidity. J Polit Econ 91(3):401–419

Ensign D, Friedler SA, Neville S, Scheidegger C, Venkatasubramanian S (2018) Runaway feedback loops in predictive policing. Proc Mach Learn Res 81:1–12

Ferguson AG (2017) Policing predictive policing. Washington University Law Review 94:1109–1189

Geocadin RG, Peberdy MA, Lazar RM (2012) Poor survival after cardiac arrest resuscitation: a self-fulfilling prophecy or biologic destiny? Crit Care Med 40(3):979–980

Geocadin RG, Callaway CW, Fink EL, Golan E, Greer DM, …, Ko NU (2019) Standards for studies of neurological prognostication in comatose survivors of cardiac arrest: A scientific statement from the American heart association. Circulation

Goode E (2011) Sending the police before there's a crime. New York Times. https://www.nytimes.com/2011/08/16/us/16police.html. Accessed 15 Nov 2021

Grünbaum A (1956) Historical determinism, social activism, and predictions in the social sciences. Br J Philos Sci 7(27):236–240

Grünbaum A (1963) Comments on Professor Roger Buck's Paper "Reflexive Predictions". Philos Sci 30(4):370–372

Harcourt BE (2007) Against prediction: Profiling, policing, and punishing in an actuarial age. University of Chicago Press

Henshel RL (1982) The boundary of the self-fulfilling prophecy and the dilemma of social prediction. Br J Sociol 33:511–528

Hogarth RM (2001) Educating intuition. University of Chicago Press

Holroyd J, Scaife R, Stafford T (2017) Responsibility for implicit bias. Philos Compass 12(3):e12410

Johnson SL (2000) The self-fulfilling prophecy of police profiles. In: Markowitz MW, Jones-Brown DD (eds) The system in black and white: exploring the connections between race, crime, and justice. Praeger, pp 93–108

Jussim L (1989) Teacher expectations: Self-fulfilling prophecies, perceptual biases, and accuracy. J Pers Soc Psychol 57(3):469–480

Jussim L, Harber KD (2005) Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. Pers Soc Psychol Rev 9(2):131–155

King OC (2020) Presumptuous aim attribution, conformity, and the ethics of artificial social cognition. Ethics Inf Technol 22:25–37

Kopec M (2011) A more fulfilling (and frustrating) take on reflexive predictions. Philos Sci 78(5):1249–1259

Krishna D (1971) "The self-fulfilling prophecy" and the Nature of Society. Am Sociol Rev 36(6):1104–1107

Kroll JA, Barocas S, Felten EW, Reidenberg JR, Robinson DG, Yu H (2016) Accountable algorithms. U Pa l Rev 165:633–705

Lipton ZC (2018) The mythos of model interpretability. Commun ACM 61(10):36–43

Lucas RE (1976) Econometric policy evaluation: A critique. In: Carnegie-Rochester conference series on public policy, 1(1), 19–46.

Lum K, Isaac W (2016) To predict and serve? Significance 13(5):14–19

Lynch J (2016) Is predictive policing the law-enforcement tactic of the future. Wall Street J. https://www.wsj.com/articles/is-predictive-policing-the-law-enforcement-tactic-of-the-future-1461550190. Accessed 15 Nov 2021

MacKenzie DA (1996) Knowing machines: Essays on technical change. MIT Press

Marsh C (1984) Back on the bandwagon: The effect of opinion polls on public opinion. Br J Polit Sci 15(1):51–74

Mertens M (2018) Liminal innovation practices: questioning three common assumptions in responsible innovation. J Respons Innov 5(3):280–298

Mertens M, King OC, Van Putten MJ, Boenink M (2022) Can we learn from hidden mistakes? Self-fulfilling prophecy and responsible neuroprognostic innovation. J Med Ethics 48(11):922–928

Merton RK (1948) The self-fulfilling prophecy. Antioch Rev 8(2):193–210

Merton RK (1968) Social theory and social structure, 3rd edn. Simon and Schuster

Miller K (2014) Total surveillance, big data, and predictive crime technology: Privacy's perfect storm. J Technol Law Policy 19(1):105–146

Nagel E (1961) The structure of science: problems in the logic of scientific explanation. Harcourt, Brace & World, New York

Nissenbaum H (1996) Accountability in a computerized society. Sci Eng Ethics 2(1):25–42

O'Neil C (2016) Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books

Pham A, Castro C (2019) The moral limits of the market: the case of consumer scoring data. Ethics Inf Technol 21(2):117–126

Popper K (1957) The Poverty of Historicism. Routledge, London/New York

Rakova B, Chowdhury R (2019) Human self-determination within algorithmic sociotechnical systems. AAAI FSS-19: Human-Centered AI: Trustworthiness of AI Models and Data, Arlington, Virginia. arXiv preprint arXiv:1909.06713

Rescher N (1998) Predicting the future: An introduction to the theory of forecasting. SUNY Press

Richardson R, Schultz JM, Crawford K (2019) Dirty data, bad predictions: how civil rights violations impact police data, predictive policing systems, and justice. New York Univ Law Rev 94:15–55

Romanos GD (1973) Reflexive predictions. Philos Sci 40(1):97–109

Rosenthal R, Jacobson L (1968) Pygmalion in the classroom: Teacher expectations and student intellectual development. Holt, New York

Rosenthal R, Rubin DB (1978) Interpersonal expectancy effects: The first 345 studies. Behav Brain Sci 1(3):377–386

Rubel A, Castro C, Pham A (2019) Agency laundering and information technologies. Ethical Theory Moral Pract 22(4):1017–1041

Salmon WC (1981) Rational prediction. Br J Philos Sci 32(2):115–125

Saul J (2013) Implicit bias, stereotype threat, and women in philosophy. In: Hutchison K, Jenkins F (eds) Women in philosophy: What needs to change. Oxford University Press, Oxford, pp 39–60

Searle JR (1976) A classification of illocutionary acts. Language in society 5:1–23

Shmueli G, Koppius OR (2011) Predictive analytics in information systems research. MIS Q 35:553–572

Siegel E (2016) Predictive analytics: the power to predict who will click, buy, lie, or die, 2nd edn. John Wiley & Sons

Simon HA (1954) Bandwagon and underdog effects and the possibility of election predictions. Public Opin Q 18(3):245–253

Smith AM (2015) Responsibility as answerability. Inquiry 58(2):99–126

Snyder M, Tanke ED, Berscheid E (1977) Social perception and interpersonal behavior: on the self-fulfilling nature of social stereotypes. J Pers Soc Psychol 35(9):656–666

Swett DH (1969) Cultural bias in the American legal system. Law Soc Rev 4:79–110

Thomas WI, Thomas DS (1928) The child in America: Behavior problems and programs. Knopf

Van Lente H (2000) Forceful futures: from promise to requirement. In: Brown R, Rappert B, Webster A (eds) Contested futures. A sociology of prospective techno-science. Routledge, pp 43–63

Venn J (1866) The logic of chance. Macmillan, London

Watson HJ, Wixom BH (2007) The current state of business intelligence. Computer 40(9):96–99

Wilkinson D (2009) The self-fulfilling prophecy in intensive care. Theor Med Bioeth 30(6):401–410

Word CO, Zanna MP, Cooper J (1974) The nonverbal mediation of self-fulfilling prophecies in interracial interaction. J Exp Soc Psychol 10(2):109–120

Zarsky TZ (2014) Understanding discrimination in the scored society. Wash Law Rev 89:1375–1412