

Self-fulfilling Prophecies

*Michael Biggs**

Chapter 13 of *The Oxford Handbook of Analytical Sociology*

ed. Peter Bearman and Peter Hedström

(Oxford University Press, 2009)

Introduction

The term “self-fulfilling prophecy” (SFP) was coined in 1948 by Robert Merton to describe “a *false* definition of the situation evoking a new behavior which makes the originally false conception come *true*” (Merton 1968: 477). He illustrated the concept with a run on a bank (a fictitious “parable”); his main application was to racial discrimination. The term has since entered social science and even everyday English, a rare feat for a sociological neologism. The concept has been subsequently rediscovered or renamed as the “Oedipus effect” (Popper 1957), “bootstrapped induction” (Barnes 1983), or “Barnesian performativity” (MacKenzie 2006). SFP has been discerned in a congeries of processes (e.g. Henschel 1978): within an individual, as with placebo response; in relations between individuals, such as teacher and student; in relations between collective actors, like states; underlying institutions, such as banks and financial markets; and, most provocatively, between social theory and social reality.

SFP is a particular type of dynamic process. It is not the truism that people’s perceptions depend on their prior beliefs (Rydgren, this volume). Nor is it the truism that beliefs, even false ones, have real consequences. To count as SFP, a belief must have consequences of a peculiar kind: consequences that make reality conform to the initial belief. Moreover, I argue that there is an additional defining criterion. The actors within the process—or at least some of them—fail to

understand how their own belief has helped to construct that reality; because their belief is eventually validated, they assume that it had been true at the outset. This misapprehension is implicit in Merton's account. His examples are social pathologies, but not merely in the sense that they are socially undesirable. They are "pathological" for being predicated on misunderstanding. Depositors fail to realize that their own panicked withdrawals cause the bank to collapse; whites fail to realize that their own racial discrimination makes African Americans seem intellectually inferior.

The mirror image of SFP is the "suicidal prophecy" (Merton 1936), in which the initial belief leads to behavior that makes the belief untrue. This kind of dynamic process has attracted little attention, although it may have considerable importance, as for example in the pursuit of positional goods. It is excluded due to limitations of space.

The chapter opens by proposing an explicit definition of SFP. I argue that the concept is best used to demarcate a narrow range of social processes rather than to encompass all of social life (Barnes 1983). Conceptualization underlines just how difficult it is to empirically demonstrate the existence and significance of SFP. The next two sections reflecting my conviction that analytical sociology must be empirical as well as theoretical. A summary of methods for investigating SFP is followed by a review of systematic evidence for selected phenomena. The final two sections offer explanations for why self-fulfilling prophecies occur: why false or arbitrary beliefs are formed and why they are subsequently fulfilled. My aim is to demonstrate commonalities among substantively diverse phenomena, and to identify the conditions which are most likely to give rise to SFP.

1. Concept

Conceptualizing SFP will clarify its essential elements and differentiate it from other kinds of dynamic processes. There are two criteria. The first is a causal sequence like the following:

- (1) X believes that "Y is p"
- (2) X therefore does b

(3) Because of (2), Y becomes p

To illustrate, X is a teacher, Y is a student, and p stands for someone with great academic ability; b is the behavior—perhaps providing better teaching or expressing greater emotional warmth—that actually causes Y to fulfill the expectations of high achievement. The third step requires it to be true that Y becomes p rather than X incorrectly perceiving that “Y becomes p.” SFP is not merely confirmation bias.

The second criterion is a misapprehension of this causal sequence. X and Y (or at least one of them) wrongly assume that the causal order is as follows:

- (0) Y is p
- (1) Because of (0), X believes that “Y is p”
- (2) X therefore does b
- (3) Because of (0), Y manifests p

To continue the illustration, teacher and student assume that the teacher’s expectations simply reflect the fact that the student has great ability. They assume that this ability—and not the teacher’s behavior—causes the subsequent high achievement. This assumed causal sequence will be called the “inductively-derived prophecy” (IDP); this term is inelegant but will prove useful shorthand.

So SFP comprises a causal sequence whereby an actor’s belief motivates behavior that turns it into reality, while at the same time the actor(s) misapprehend the causal sequence as one whereby belief simply reflects reality. This misapprehension is not unreasonable: after all, IDP is common.

The distinction between the two causal sequences can be sharpened by specifying a counterfactual. SFP requires that alteration of X’s belief would alter the outcome:

- (1) X believes that “Y is q”
- (2) X therefore does c
- (3) Because of (2), Y becomes q

If the teacher were to believe that the student is mediocre, then the teacher would behave very differently, and this would cause the student to perform poorly. By contrast, the counterfactual in IDP requires that a different belief held by X has no effect on Y:

- (0) Y is p
- (1) Despite (0), X believes that “Y is q”
- (2) X therefore does $\underline{c}$
- (3) Because of (0), Y manifests p

In this case the teacher’s belief would have no effect on the student’s performance.

I have defined the causal sequence of SFP without specifying Y’s ‘real’ state at the outset (corresponding to step 0 of IDP). Merton’s stipulation that X’s belief be false raises epistemological difficulties in cases where it is hard to conceive of Y’s state prior to interaction with X (Miller 1961; Krishna 1971). Consider police indiscriminately attacking an unruly crowd, who then react by fighting back (Stott and Reicher 1998). Does it make sense to conceive the crowd as inherently peaceable? This difficulty is circumvented by resort to the counterfactual: if the police had instead shown restraint, then the crowd would have dispersed without violence. Therefore I will refer to X’s beliefs as false or *arbitrary*: arbitrary because a different belief would create a different reality.

Thus far I have considered SFP involving two individual or collective actors (X and Y). This is readily generalized to a causal sequence involving multiple individuals (X_i) and an abstract social institution (Y):

- (1.1) X_1 believes that “Y is p”
- (2.1) X_1 therefore does b
- (1.2) Because of (2.1), X_2 believes that “Y is p”
- (2.2) X_2 therefore does b
- (1.3) Because of (2.1) and (2.2), X_3 believes that “Y is p”
- (2.3) X_3 therefore does b
- ...
- (3) Because of (2.1), (2.2), (2.3)...., Y becomes p

In Merton’s parable of a bank run, X_i are the depositors, Y is the bank, and p stands for insolvency; b is the behavior—withdrawing money—that actually causes the bank to fail. The iteration of the first two steps in the sequence is a process of positive feedback: as more depositors withdraw their money from the bank, the remaining depositors will be more likely to believe that the bank is

insolvent.

Having offered a precise definition, I will now put it to use. Firstly, let me show how this definition helps to distinguish SFP from superficially similar dynamic processes. Take the first criterion, causal sequence. According to Merton, whites deprived African Americans of education because they believed them to be intellectually inferior; given such inadequate education, blacks became academically inferior. Is the belief (step 1) sufficient to explain the behavior (step 2)? One suspects that this belief was more of a rationalization for hatred and fear of blacks. The question may be sharpened by considering the counterfactual: is altering those beliefs in intellectual inferiority, leaving everything else unchanged, conceivable? If whites came to believe that blacks were equally intelligent, would they have devoted resources to providing equal educational facilities? Only if the answer is affirmative would this fall under my definition of SFP.

Equally important is the second criterion, misapprehension. This distinguishes SFP from self-enforcing conventions and similar phenomena. Take the example of driving on the left in Great Britain. Each driver believes that others will drive on the left and therefore drives on that side of the road. Leaving aside the difficulty of envisaging a plausible counterfactual (how many drivers would have to simultaneously switch their beliefs to reverse the convention?), there is no misapprehension of causation. Drivers are quite aware that they drive on the left only because everyone else in Britain does. After all, they have no difficulty comprehending that the convention reverses as soon as they arrive in France.

My definition thus serves to demarcate a particular type of dynamic process. It nevertheless encompasses a diverse range of phenomena. This chapter selects a few of these for consideration, summarized in Table 1. Interpersonal expectancy and reactive conflict are straightforward.¹ Placebo response is rather different. It is SFP when caused by response expectancy—“anticipations of one’s own automatic responses to various stimuli” (Kirsch 2005: 794)—rather than by unconscious conditioning (Stewart-Williams and Podd 2004). While placebo response occurs “within” the

individual undergoing treatment, it is important to note that the individual's belief derives from situational encounters with someone assumed to possess expertise.² Placebo response deserves attention because it has been investigated so extensively and rigorously.

[INSERT TABLE 1 HERE]

The bank run exemplifies SFP with positive feedback. Positive feedback also occurs in the investment bubble, though the causal sequence is more elaborate. As more people invest in the scheme, their investment increase the scheme's credibility. Moreover, subsequent investment enables the scheme to provide initial investors with the promised returns; these returns further enhance the credibility of the scheme. All this leads to further investment. The process can last for some time, but it eventually proves unsustainable. The ultimate outcome is the same as in the bank run: financial failure. The crucial difference is that in the investment bubble this outcome refutes rather than fulfills the prophecy. What is self-fulfilling—albeit temporarily—is the belief that the investment yields high returns.

The final example to be considered is self-fulfilling theory. What distinguishes social science from natural science is the potential for reality to be altered by theory. A theory of society could, in principle, prove self-fulfilling. Marxism predicts that capitalism is fated to end in revolution; if many people believe in the theory, then they could foment revolution. For self-fulfilling theory, misapprehension is more ambiguous than for other phenomena, for individuals may be partly cognizant of their own role in making reality conform to theory. Nevertheless, the ability of theory to motivate action surely requires a conviction that the theory expresses some underlying truth about Society, however imperfectly instantiated in actually existing social arrangements.

These varied phenomena are alike in that SFP is very hard to distinguish from IDP. Indeed, social life provides plenty of genuine examples of the latter: patients experience pain relief because they have been given morphine; students perform well because they have genuine talent; enemies attack because they are aggressive; banks fail because they are insolvent; investments yield high

returns because the asset is productive; theories are confirmed because they are true. In all these cases, beliefs follow reality. And because such cases exist, actors are naturally prone to misapprehend the causal sequence of SFP: they mistakenly think that their belief follows reality, when in fact their belief creates it.

2. Method

Identifying SFP in reality poses an immense challenge. If the actors involved cannot discriminate between IDP and SFP, how can a sociologist know any better? One method is to manipulate actors' beliefs in order to see whether those beliefs have causal effect. The experimental method (see Bohnet, this volume) is worth discussing for its probative value, and also because it reveals how beliefs are not easily manipulated—a point of some sociological interest.

Placebo research exemplifies the experimental method. Randomly assign patients to two groups. Convince the first that they are getting a painkiller such as morphine, while injecting an inert substance. Leave the second group untreated altogether; this control group is necessary to exclude spontaneous remission and regression to the mean. The difference in pain experienced by the two groups (placebo minus control) is placebo response. There is also an alternative experimental design. Administer the painkiller to the first group, as with any treatment. Covertly administer the painkiller to the second group without them realizing what has happened; in practice, this can be done when patients are constantly hooked to an intravenous drip (Levine et al. 1981). The difference in pain experienced by the two groups (treatment minus covert) is placebo response.³ Each design allows the effects of subjective belief to be disentangled from the inherent physical powers of the treatment. Similar manipulation is used to study interpersonal expectancy. The most informative experiments are those conducted outside the psychology laboratory, in natural social situations. Randomly select a small proportion of students in a class. Convince the teacher that these students have especially high

potential. At the end of the school year, compare this select group with the control group. The difference in performance measures the effect of the teacher's positive expectations (Rosenthal and Jacobson 1968).

The logic of experimental design is straightforward; putting it into practice is surprisingly difficult. The norm of informed consent constrains manipulation in medical research. In the double-blind clinical trial, patients in the so-called placebo group are *not* deceived into thinking that they are getting the treatment; they know that placebo and treatment are equally likely. In fact, blinding is rarely if ever achieved; well over half the patients correctly guess whether they are getting placebo or treatment (Shapiro and Shapiro 1997). Informed consent must attenuate placebo response.

Even without this normative constraint, manipulation is difficult to accomplish. Experimental subjects expect to be duped. In a typical experiment looking at the effects of caffeine, one third of the subjects in the control group—given placebo and told exactly that—believed incorrectly that they had received caffeine (Kirsch and Rosadino 1993). Similar problems may attend experiments in natural settings. The initial findings on teachers' expectations were widely reported in the press. In subsequent replications, teachers may well have become suspicious when educational psychologists arrived with a list of students deemed to be exceptionally talented.

The difficulties of manipulating belief do not vitiate experimental findings, of course. Rather, we should consider these findings as establishing a *lower bound* on the causal effect of belief. The point is illustrated by a novel surgical technique for angina pectoris, which initially demonstrated impressive success in reducing pain (Beecher 1961; Benson and McCallie 1979). Subsequent placebo surgery demonstrated that the operation was physically ineffective; it also induced a less potent placebo response. In other words, being treated “for real” by an enthusiastic surgeon amplified placebo response. Consider another example, an experiment showing that listeners are more likely to download a song from a website if they believe it is popular with other users (Salganik and Watts 2007). We may infer that the effect of popularity would be far greater if someone believed that the

song was the number one hit in the country.

The experimental method provides the strongest proof that false or arbitrary beliefs can become self-fulfilling. In most processes of concern to sociologists, however, experiment is not possible; even where it is feasible in some fashion, there may be doubts about its ecological validity. An alternative is to find a 'natural experiment,' where real life has provided a situation where we know—in retrospect—that beliefs were false, analogous to the surgery for angina pectoris. Ponzi schemes offer a neat natural experiment for investment bubbles, inasmuch as the investment is inherently incapable of generating value. The name comes from a scheme created by Charles Ponzi in Boston after the First World War (Zuckoff 2005). He initially promised a 40% return after 90 days; the rate of return was soon raised to 50% in 45 days. This spectacular profit supposedly came from arbitraging the value of international reply coupons in different countries. He would buy coupons in Italy, where postal costs were low, then ship the coupons to the United States and exchange them (somehow) for dollars. This premise was so absurd that Ponzi made no attempt to implement it.

An alternative method is statistical analysis of longitudinal data. For interpersonal expectancy, use teachers' expectations of ability at time t_1 to explain performance at t_2 , controlling for performance at t_0 . This has been done with data from schools in Michigan, which include many potentially confounding variables; the statistical identification of the causal effect of teachers' beliefs is therefore convincing (Jussim, Eccles, and Madon 1996; Madon, Jussim, and Eccles 1997). Statistical analysis is possible even if the data do not include measures of beliefs. Economic historians use data on banks to test whether failure can be predicted from the institution's previous financial characteristics (e.g., the extent to which liquidity is maintained by borrowing) and from exogenous economic shocks. SFP is thus treated as a residual, plausible if failure cannot be predicted by these economic variables (e.g. Calomiris and Mason 1997, 2003).

3. Evidence

Empirical investigation must answer two questions. The first is whether SFP can occur. An affirmative answer requires evidence that a false or arbitrary belief can have a causal impact on behavior, so as to make that behavior conform to the belief. It is one thing to discern a causal effect in the requisite direction, another thing to demonstrate that the effect's magnitude is substantial. If SFP can occur, there follows a less obvious but equally important question: whether it occurs often enough—outside the context of experimental manipulation—to be socially significant. An affirmative answer requires evidence that false or arbitrary beliefs are prevalent. Such an answer cannot be provided by experiments, and so in this respect statistical evidence is more valuable. I will review three bodies of systematic research—on placebo response, interpersonal expectancy in education, and bank runs—and then note evidence on Ponzi schemes and music popularity.

There is a huge body of research on placebo. Evidence is compelling for pain, depression, and anxiety. The outcome is subjective experience as reported by the patient (e.g. pain on a scale from 1 to 10); there are sometimes associated physical measurements, as with facial inflammation after dental surgery (Hashish et al. 1988). For pain, a meta-analysis of conventional clinical trials, comparing the placebo group with the control group, finds a standardized mean difference of .27 (Hróbjartsson and Gøtzsche 2001). Thus placebo reduced pain by about a quarter of a standard deviation, which is modest. But these trials undermine placebo response by informed consent, as mentioned above. A meta-analysis of experimental studies, where subjects in the placebo group had been told that they were getting an effective analgesic, finds a standardized mean difference of .85 (Vase, Riley, and Price 2002). This is substantively large. For depression, a meta-analysis of clinical trials of anti-depressants finds that the placebo group experienced about 80% of the benefits of the treatment group (Kirsch et al. 2002).⁴

Placebo response is ubiquitous in medical treatment. What is interesting for the sociologist is that “treatments” *without* inherent physical powers can survive thanks to placebo response, along

with spontaneous remission and other factors that cannot be discerned without experimental control. Until the nineteenth century at the earliest, the vast majority of medical treatments had no inherent physical powers—or at least none that were beneficial (Shapiro and Shapiro 1997). Treatments that flourished were surely those that induced a powerful placebo response; for instance cures composed of rare or repugnant ingredients, like viper’s flesh. Today, the massive market for “alternative” therapies can be attributed in many cases to placebo response. Less obviously, perhaps, technologically advanced equipment may also induce a powerful placebo response (Kaptchuk et al. 2000). Ultrasound apparatus, for example, relieves muscular pain equally well when it is turned off (Hashish et al. 1988).

There is an extensive literature on interpersonal expectancy in education (Jussim and Harber 2005). The outcome is typically the result of a standardized written test. It is sometimes an IQ test, but this need not imply that expectations alter “intelligence”; test-taking performance is sufficiently important, given its impact on students’ life chances. A meta-analysis of experiments using an IQ test finds that under the most favorable conditions (when expectations are manipulated at the beginning of the school year) the standardized mean difference between the “high expectations” group and the control group is .32 (Raudenbush 1984). In other words, 63 percent of the former group performed above the latter group’s median. Similarly modest effects were found in statistical analysis of students learning mathematics in Michigan. These data also show, however, a much greater effect of expectations on lower-class and especially African-American students (Jussim, Eccles, and Madon 1996). For the latter, the standardized mean difference was 1.35. To put this more concretely, moving from the lowest to the highest expectation is predicted to raise the student’s grade from C to B+. For trainees in the Israeli army, a meta-analysis of several experiments finds a standardized mean difference of .86 (Eden 1990; Keirein and Gold 2000). All these figures are averages across all levels of ability. The less able are more sensitive to teachers’ expectations than the more able (Keirein and Gold 2000; Madon, Jussim, and Eccles 1997). In addition, expectations that are overestimated appear

more powerful than expectations that are underestimated. Happily, then, SFP is most likely to *raise* the performance of those who are *below* average.

The prevalence of inaccurate expectations is another matter. Systematic evidence suggests that teachers' expectations are usually accurate, in the sense that they conform to objective indicators of prior performance such as test scores (Jussim and Harber 2005). One might be tempted to argue that prior performance is simply the result of prior expectations, and so ultimately students' trajectories depend on the expectations of their initial teacher. To the contrary, the effects of expectations seem to diminish rather than accumulate over subsequent years (Smith, Jussim, and Eccles 1999). What is most important is whether inaccurate expectations are correlated with social categories such as race or class, which would vindicate Merton's original application of SFP. The Michigan data show that teachers were biased neither against African Americans nor lower-class students: differences in expectations for these groups matched differences in prior performance (Jussim, Eccles, and Madon 1996).⁵ We cannot, of course, conclude that biased expectations have never existed. Systematic studies on teachers have been conducted in America in recent decades, in a time when racism was becoming unacceptable in the teaching profession. One might doubt whether the same results would have been obtained in Michigan classrooms in the 1950s, for example.

There are a number of statistical studies of bank runs in the United States before the adoption of Federal deposit insurance. Bank failure was far more common than in other countries, due to the prohibition of large-scale banks. A crucial distinction is between suspension and failure: some banks that suspended withdrawals eventually reopened, while others failed. There is clear evidence that a run could lead to suspension. Event-history analysis of banks during the panic of 1893 cannot distinguish between those that were suspended but reopened and those that remained unscathed (Carlson 2005). This suggests that runs beset banks that were fundamentally sound. An indirect approach is to examine the effect of state deposit insurance (introduced by some states and not others), which should reduce or eliminate runs. In the 1920s, it is estimated that insurance halved the

number of suspensions due to runs, as diagnosed by regulators at the time (Chung and Richardson 2006). These suspensions, however, represented only a small proportion of all suspensions. When it comes to failure, moreover, there is scant evidence that runs often caused solvent banks to fail. Event-history analysis of banks in the 1930s—including the Chicago panic of 1932, the setting for Merton’s parable—reveals fundamental pre-existing differences between those that survived and those that failed (Calomiris and Mason 1997, 2003). Solvent banks could survive (even after suspension) because they maintained the confidence of large depositors, who had better information about their financial position.

The literature on financial bubbles is huge and contentious. It is not a simple matter to establish that an investment boom is really a bubble. To take a salutary example, “tulipmania” has passed into social scientific folklore as the exemplary bubble, and yet the original evidence can be interpreted as showing rational investment in an intrinsically valuable commodity (Garber 1989). Therefore I will focus on the original Ponzi scheme as a natural experiment proving the possibility of a pure bubble, inflated by SFP. As mentioned above, the underlying investment scheme—arbitrage on postal vouchers—was not feasible let alone productive. Yet the scheme succeeded in attracting a huge volume of investment. Ponzi appointed an Italian grocer as agent (on 10% commission). By January 1920, the agent had persuaded seventeen individuals, his customers and friends, to invest an average of a hundred dollars each. Crucially, the trickle of investors continued over the following month, so that Ponzi could pay the returns owed to the original investors. Once investors gained such extraordinary high yields, they told their friends and family; they also often reinvested what they had gained. Figure 1 traces the total invested each month. (Unfortunately reinvestment is not distinguished.) Investment grew exponentially, represented by a straight line on the graph’s logarithmic scale; it quadrupled every month. In July, about twenty thousand people invested in the scheme. By the end of that month, people waited in line for hours outside Ponzi’s office, eager for their chance to get rich quick. Under investigation by the financial authorities, he voluntarily halted

deposits. Even then, however, there was no panic; most investors did not withdraw their money. The scheme collapsed some weeks later when a newspaper discovered his earlier arrest for fraud. Without these two interventions, it seems that the scheme would have continued for longer.

Ponzi schemes should be most viable where people have little or no experience in investment, like the shopkeepers and workers in 1920. Therefore these schemes flourished in Russia and Eastern Europe in the 1990s, during the transition to a market economy; Albanian investors lost the equivalent of half the country's annual GDP (Bezemer 2001). More surprising is the recent discovery that thousands of experienced investors in the United States and Western Europe had been similarly fooled. As this chapter went to press, Bernard Madoff admitted running a Ponzi scheme from the early 1990s to 2008, which consumed many billions of dollars (statement to the U.S. District Court in New York, 12 March 2009). His fraud departed from the classic Ponzi model, however, in that the returns were not extraordinarily high and there was no wholesale recruitment of investors. Aside from pure Ponzi schemes, conventional investment booms—like internet stocks in the 1990s or house prices in the 2000s—owe something to self-fulfilling expectations (Shiller 2000).

Finally, a recent study is noteworthy for exploiting the internet to conduct an experiment in a natural setting (Salganik and Watts 2007). Songs from real music groups were made available for listening and downloading. The popularity of each song, indicated by the number of previous downloads, was visible to users. In one condition, popularity was inverted part way through the experiment: users were deceived into thinking that the most popular song was the least popular, and so forth. This dramatically affected subsequent usage, making the “worst” (indicated by previous downloads) songs more popular and the “best” songs less popular. Over time, however, the very best songs gradually rose in rank, showing that appreciation depends on intrinsic qualities as well as perceived popularity

4. Explanations: belief

I now turn to the overarching theoretical problem, explaining *why* SFP occurs. The problem can be broken down into two questions. First, why does X form false or arbitrary beliefs (step 1 of the causal sequence) and behave accordingly? Second, why does X's behavior subsequently fulfill those beliefs (step 3)? Answering these questions involves, at the same time, describing the conditions in which SFP is likely to arise. There is no single answer to each question, of course, because SFP encompasses such diverse processes. However, we can identify explanations of sufficient generality to apply to more than one substantive phenomenon.

The first question—how to explain false or arbitrary beliefs—is thrown into relief by contrasting SFP with IDP. In IDP, X's belief about Y is derived from evidence about Y. In SFP, by contrast, X formulates a false or arbitrary belief without waiting for evidence. The former scenario may seem naive (Rydgren, this volume), but it is what the actor imagines is happening. The teacher, for example, considers her belief about the student's ability to be based on observation. I will identify three explanations for behavior based on false or arbitrary beliefs.

One explanation is that X has power over Y or that Y accepts X's expertise. X then has considerable latitude in forming beliefs about Y, whether or not they are justified by evidence. Y cannot challenge false or arbitrary beliefs due to this power imbalance, or—more insidiously—Y accepts those beliefs as true. At one extreme, consider the example of a prisoner who is tortured as long as he refuses to admit being a terrorist; at the other extreme, consider a student who accepts the teacher's judgment that he is poor at mathematics. In both cases, X's belief may be completely false and yet will remain uncontested; X has therefore no reason to revise or even question it. The belief can survive long enough to become self-fulfilling.

A second explanation is that the cost of making a mistake is asymmetric and the cost of waiting is high. This situation is exemplified by the bank run. If a depositor is uncertain whether to believe a rumor that the bank is about to collapse, there is a vast difference in the cost of making a mistake. On one hand, withdrawing savings is a relatively low-cost action, even if the bank turns out to be sound.

On the other, inaction will mean losing everything if the rumor is true. Waiting for more evidence of the bank's financial condition simply increases the likelihood, if it really is unsound, of losing everything. Therefore the rational action is to withdraw savings immediately.⁶ The same logic applies to reactive conflict, insofar as the cost of mistaken aggression is outweighed by the cost of mistaken passivity. When police are deciding whether to attack an unruly crowd, the cost (to them) of attacking people who would have dispersed peacefully may be far lower than the cost of allowing a violent mob to get out of control.

A third explanation is that someone stands to gain by inculcating false or arbitrary beliefs. Purveyors of medical remedies or Ponzi schemes are obvious examples. Politicians too may gain from provoking international conflict. Where the instigator intentionally decides to mislead others, this still counts as SFP insofar as it requires others to misapprehend the causal sequence. The politician may cynically provoke another state, but gains only if the populace believe that their country is the victim of unprovoked aggression. While this explanation depends on intentional deception, it is also possible to suggest a more elaborate version where "deception" is unintended. After all, the most convincing advocates are those who genuinely believe. It is crucial, however, for such an explanation to specify how rewards shape action in the absence of intentionality; in short, to avoid the problem of functionalism. A selectionist explanation may be plausible. Consider several homeopaths dispensing (ineffective) remedies to patients. Those who are most enthusiastic about the potency of a remedy will—by inculcating a false belief in the patient and hence inducing placebo response—enjoy the best results. Over time, those homeopaths will thrive while their more cautious colleagues fail. In this scenario, the fact that homeopaths are rewarded for inculcating false beliefs helps to explain why this particular SFP occurs, but there is no conscious deception by practitioners.

These three explanations are sufficiently general to cover a range of SFP. They do not, however, exhaust the possible explanations for false or arbitrary beliefs. Two processes— investment bubbles and self-fulfilling theories—demand other explanations. While the logic of asymmetric costs helps to

explain a bank run, it makes an investment bubble more puzzling. If the investment turns out to be bogus, then the investor will lose everything. How then could such an investment be explained? It is possible to conceive of an inordinately sophisticated investor who knows that the investment is a bubble and who is able to calculate when it will burst; such an investor can exit with a handsome profit. A bubble created by such sophisticates alone would not, of course, fall within my definition of SFP. There is ample evidence that bubbles attract naive investors in considerable numbers. This can be demonstrated for Ponzi's scheme, because after Ponzi closed it to new depositors he continued to redeem existing certificates. Anyone knowing that it was a bubble would have immediately withdrawn their money. Because the vast majority of investors did not, they must have believed that the scheme was intrinsically capable of generating extraordinary returns. To explain this belief, it is necessary to refer to learning. They had seen other investors receiving—in cash—the promised rate of return; some had experienced this for themselves on previous iterations. "Each satisfied customer became a self-appointed salesman," recalled Ponzi (Zuckoff 2005: 116). "I admit that I started a small snowball downhill. But it developed into an avalanche by itself."

Self-fulfilling theory demands a particular explanation. It is easy to see why social scientists adhere to theories that are false or arbitrary; the interlocking system of propositions is almost impervious to countervailing evidence, and scientists are far more committed to their theory than are (say) depositors to their belief about the bank. What is harder to discern is why some theories motivate their adherents to act in the world outside academia.⁷ Powerful motivation is created by the conflation of descriptive and normative claims. Neoclassical economics and Marxism share this character. On one hand, they are ostensibly hard-headed scientific theories of reality, possessing unique rigor. On the other hand, they specify—and also justify—a normative ideal, immanent in capitalism or imminent in history. The combination seems logically contradictory, but it is nonetheless highly productive. Markets are natural, and so we should create them! The revolution is inevitable, and so we should foment it! Motivation can also spring from a more venal source.

Neoclassical economics promises tangible gain to those who deploy it in the real world. Anyone who believes its predictions about prices should use those predictions in their own market trading.

5. Explanations: fulfillment

After explaining why X behaves on the basis of false or arbitrary beliefs, there follows a second question: why does that behavior cause Y to fulfill those beliefs? After all, action on the basis of incorrect beliefs is widespread—but it is rare for the consequences of action to bring reality into alignment with the initial beliefs, as with SFP. The explanation may be straightforward where Y is an institution. A bank by definition does not have sufficient liquidity to repay all its depositors simultaneously; a Ponzi scheme can obviously thrive if the volume of new investment is sufficient to repay earlier investors. Explanation poses a much greater challenge where Y is an actor. Potential explanations can be classified according to the degree to which Y accepts X's belief about Y. I will focus on three types of explanation.

One type of explanation depends on Y believing in the validity of X's expectations. In some circumstances, this invidious belief could simply have its effect by means of altering Y's perceived payoffs. A student accepting the teacher's low opinion of his ability would rationally choose to reduce the time spent on study or practice. A more intriguing explanation is "response expectancy," originally proposed to explain placebo response. Here Y's belief has its effect beneath the actor's consciousness. Believing that you have been given morphine leads to pain relief, but you are not aware of the causal connection. Researchers on pain have actually discovered a chemical substance that enhances placebo response, by interacting with the brain's endogenous opioid system (Colloca and Benedetti 2005). This substance has no inherent physical powers to relieve pain—demonstrated by the fact it has no effect when administered covertly. It simply amplifies the effect of belief. While the precise mechanisms can be left to biologists, response expectancy may have unappreciated

significance for sociology. Anxiety and depression are subject to placebo response, and these have socially important effects. A student who accepts the teacher's low expectations may suffer from anxiety when taking a test, and anxiety will necessarily degrade performance.⁸

At the opposite extreme is a type of explanation that does not require Y to accept (or even know about) X's belief. Rather, X's behavior alters the payoff structure for Y. Torture will almost inevitably elicit a confession even when the victim knows that he is innocent. Reactive conflict has the same logic: once the police attack the crowd, they will fight back even though they do not accept the police's characterization of them as a violent mob. In some circumstances, Y's behavior need not involve a decision. A student who does not accept the teacher's low opinion is nonetheless hindered by being assigned less challenging material. For a less hypothetical example, consider the observation that a gang member's status predicted his bowling performance when the gang bowled together, even though some of the low-ranking individuals showed considerable skill when bowling alone (Whyte 1943). If someone with low rank began performing well, other gang members (especially those at the bottom of the hierarchy) would deliberately throw him off his game. It is no wonder, then, that the gang's beliefs proved self-fulfilling. Response expectancy may also have contributed, if the individual himself expected poor performance and this interfered with his concentration.

A third type of explanation falls in between the other two. There is a gap between Y's self-image and (Y's perception of) X's belief about Y; this gap leads Y to behave in ways that conform to X's belief. In short, Y intentionally lives up—or down—to X's expectations. When X has higher expectations, it is easy to envisage how Y would be motivated not to disappoint X. Experiments in the Israeli Army show that when the instructor (falsely) believed that trainees were promising, he behaved in ways that made the trainees feel that they were being treated better than they deserved; this motivated them to increase their own effort in return, which in part explains their superior performance (Eden 1990). When X has lower expectations, however, it is not altogether clear (at least to me) why Y would fulfill those expectations rather than striving to reverse them. There is

experimental evidence, however, that trust lives down as well as up to expectations; those who believe that they are mistrusted are less likely to behave in a trustworthy manner (Bacharach, Guerra, and Zizzo 2007).

Self-fulfilling theory again demands separate consideration. Theory can be made real by shaping institutions (outside academia) which in turn direct the action of numerous individuals. As an example, neoclassical economics postulates that firms maximize profits, or equivalently shareholder value (Blinder 2000). Half a century ago this proposition was contradicted by empirical evidence. Some economists modified the theory to make it more realistic (e.g. maximizing revenue or satisficing), but strict neoclassical economists argued that it was reality that needed revising: managers should be given appropriate incentives to maximize profits, namely stock options. Thanks in part to their argument, companies now adopt this form of compensation. Moreover, it is now taken for granted that the sacred duty of managers is to maximize shareholder value. As a result, short-term profit maximization—as neoclassical theory predicts—is surely more often true now than it was half a century ago. Even when it is not true, the fact that managers invariably claim to be maximizing shareholder value gives the theory superficial plausibility.

Theory can also be made real when its predictions are used to guide self-interested action. Neoclassical economists insist that prices in capital markets incorporate all relevant information; this is the “efficient market hypothesis.” Yet initial econometric testing revealed significant anomalies, like the finding that small firms provided higher returns than large firms. Anyone who believed this finding could make excess profits by investing in smaller firms, and that is exactly what some market participants—including the author of the hypothesis—did. As a result, the anomaly disappeared in subsequent empirical investigation (MacKenzie 2006; Schwert 2003). This example is structurally similar to the bank run: the belief provides incentives for action that makes it become true. It is less demanding than the bank run, however, because the effect can be produced by a tiny minority of market participants. Another example is analogous to the investment bubble, in that it enjoyed only

temporary success. Economists formulated a model of stock-option prices before a formal market was institutionalized, and indeed the model helped to legitimize the creation of such a market. Traders in this new market then used the model to calculate expected returns, assisted by datasheets sold by one of the model's authors. Option prices, when investigated econometrically, were eventually found to closely match the model's predictions. All this changed with the 1987 crash, when traders discovered to their cost that the model underestimated the volatility of stock prices (MacKenzie 2006). On this account, the model was self-fulfilling for a time, in part because it enjoyed such widespread use as a predictive tool.

Conclusion

This chapter has proposed a conceptualization that explicates the essential characteristics of SFP; it has reviewed systematic empirical evidence on the possibility and prevalence of SFP; and it has sketched a variety of explanations for SFP, which at the same time suggest conditions under which this process is likely to occur. All this could be condensed into checklist for analyzing a dynamic process as SFP. First, establish the causal sequence. This can be checked by considering the counterfactual: is it plausible that a change in the initial belief would alter the outcome? Second, establish misapprehension by at least some of the participants. Sketch their assumed IDP. Enlightenment can be employed as a diagnostic: is it plausible that making participants understand the actual causal sequence would alter the outcome? Third, gather evidence on the magnitude and prevalence of SFP. Even if systematic evidence is not available, it is certainly helpful to broaden the focus to include IDP as well SFP in this particular context. To clarify this with an example, consider not only the bank run causing a solvent institution to fail but also the bank run that responds to the institution's real insolvency. Fourth, explain why the false or arbitrary belief was formed. Why did the individual(s) act on the basis of this belief rather than waiting for evidence (as with IDP)? Fifth, explain why this belief became self-fulfilling. Remember that most incorrect beliefs do not have such

an effect.

I have argued for a restricted definition of SFP, which makes misapprehension of the causal process an essential criterion. Clearly this definition excludes many dynamic processes in which false beliefs have important consequences. However, it is more useful to restrict SFP to denote a very particular type of process than to extend it to wide swathes of social life. Indeed, it is arguable that SFP can exist because they are unusual, being “parasitic” on the frequency of IDP. These two processes would stand in the same relation as forged to authentic banknotes; the former are more likely to go undetected when rare. I have also argued for the importance of empirical research. Psychologists and economists (and medical researchers) have done much more than sociologists to systematically demonstrate the occurrence of SFP. To be sure, empirical investigation poses a huge challenge. The most compelling evidence pertains to processes (like placebo response) that are of less interest to sociologists, whereas the grandest claims for SFP (like self-fulfilling theory) are extremely difficult to investigate. This challenge will test the ingenuity of analytical sociologists.

One final lesson can be drawn from SFP. My emphasis on misapprehension underlines the importance of “folk sociology,” how social actors themselves understand the causal processes which lead to their action and which flow from it. This understanding is usually tacit and often confused, but no less significant for all that. The abiding fascination of SFP is the notion that social actors are caught in a web of their own making; they reify social reality, failing to realize that they are responsible for creating it. Analytical sociologists typically wield Occam’s razor to whittle the individual actor down to the simplest decision function. Sometimes explanation requires a more complex—and admittedly less analytically tractable—notion of the individual actor, as someone whose own (mis)understanding of social process has to be taken into account.

NOTES

* Department of Sociology, University of Oxford; michael.biggs@sociology.ox.ac.uk. The editors, Diego Gambetta, John Goldthorpe, Rob Mare, Federico Varese, and Duncan Watts all helped to sharpen the argument.

¹ Worth noting is a variant of reactive conflict where X's behavior creates or at least expands Y as a collective entity. Naming an enemy such as "the mafia" or "Al Qaeda" and attributing to it extraordinary powers is liable to encourage many other groups to jump on the "brandwagon" and identify themselves accordingly (Gambetta 2004).

² It is telling that ultrasound equipment produces a placebo response only when it is applied by someone else; self-massage with the same apparatus has no effect (Hashish et al. 1988).

³ This assumes that the treatment's inherent powers and the placebo response have an additive effect in the treatment group.

⁴ This fraction includes spontaneous remission and regression to the mean as well as placebo response; on the other hand, placebo response is attenuated by informed consent.

⁵ How can that be reconciled with the finding that teachers' expectations had a greater effect on students from these groups? When a teacher underestimated the ability of an African American student, that had a significantly detrimental effect on the student's performance (and a more detrimental effect than on a white student); yet teachers on average were no more likely to underestimate the ability of African Americans than that of white students.

⁶ This situation (as in clinical trials with informed consent) requires a more elaborate specification of belief. Rather than X believing in a simple proposition, X has a probabilistic belief: "there is probability π that the bank is unsound" and hence "there is probability $1-\pi$ that the bank is sound." The decision to withdraw deposits may be rational even if the probability π is small.

⁷ Anyone who adheres to a theory is motivated to seek supporting evidence, of course, but mere confirmation bias is not SFP. To count as such, the theory must alter the social reality that it describes.

⁸ This is not trivial: merely labeling a test an "intelligence test" significantly degrades the performance of African American students relative to white students (Steele and Aronson 1995).

REFERENCES

- Bacharach, M., G. Guerra, and D. J. Zizzo. 2007. The self-fulfilling property of trust: An experimental study. *Theory and Decision*, forthcoming.
- Barnes, B. 1983. Social life as bootstrapped induction. *Sociology*, 17: 524-45.
- Beecher, H. K. 1961. Surgery as placebo: a quantitative study of bias. *Journal of the American Medical Association*, 176: 1102-7.
- Benson, H., and D. P. McCallie, Jr. 1979. Angina pectoris and the placebo effect. *New England Journal of Medicine*, 300: 1424-9.
- Bezemer, Dirk J. 2001. Post-Socialist Financial Fragility: The Case of Albania. *Cambridge Journal of Economics*, 25: 1-23.
- Blinder, A. S. 2000. How the economy came to resemble the model. *Business Economics*, 16: 25.
- Calomiris, C. W., and J. R. Mason. 1997. Contagion and bank failures during the great depression: the June 1932 Chicago banking panic. *American Economic Review*, 87: 863-83.
- — —. 2003. Fundamentals, panics, and bank distress during the depression. *American Economic Review*, 93: 1615-47.
- Carlson, M. 2005. Causes of bank suspensions in the panic of 1893. *Explorations in Economic History*, 42: 56-80.
- Chung, C.-Y., and G. Richardson. 2006. Deposit insurance and the composition of bank suspensions in developing economies: lessons from the state deposit insurance experiments of the 1920s. NBER Working Paper, 12594.
- Colloca, L. and F. Benedetti. 2005. Placebos and painkillers: is mind as real as matter? *Nature Reviews: Neuroscience*, 6: 545-52.
- Eden, D. 1990. *Pygmalion in Management: Productivity as a Self-Fulfilling Prophecy*. Lexington, MA: Lexington Books
- Gambetta, D. 2004. Reason and terror: has 9/11 made it hard to think straight? *Boston Review*, April/May.
- Garber, P. M. 1989. Tulipmania. *Journal of Political Economy*, 97: 535-60.
- Gracely, R. H. et al. 1985. Clinicians' expectations influence placebo analgesia. *Lancet*, 325: 43.
- Hashish, I. et al. 1988. Reduction of postoperative pain and swelling by ultrasound treatment: a placebo effect. *Pain*, 33: 303-11.
- Henshel, R. L. 1978. Self-altering predictions. Pp. 99-123 in *Handbook of Futures Research*, edited by J. Fowles. Westport, CT and London: Greenwood Press.
- Hróbjartsson, A., and P. C. Gøtzsche. 2001. Is the placebo powerless? an analysis of clinical trials comparing placebo with no treatment. *New England Journal of Medicine*, 344: 1594-1602.

- Jussim, L., J. Eccles, and S. Madon. 1996. Social perception, social stereotypes, and teacher expectations: accuracy and the quest for the powerful self-fulfilling prophecy. *Advances in Experimental Social Psychology*, 28: 281-388.
- Jussim, L., and K. D. Harber. 2005. Teacher expectations and self-fulfilling prophecies: knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review*, 9: 131-55.
- Kaptchuk, T. J. et al. 2000. Do medical devices have enhanced placebo effects? *Journal of Clinical Epidemiology*, 53: 786-92.
- Kierein, N. M. and M. A. Gold. 2000. Pygmalion in work organizations: a meta-analysis. *Journal of Organizational Behavior*, 21: 913-28.
- Kirsch, I. 2005. Placebo psychotherapy: synonym or oxymoron. *Journal of Clinical Psychology*, 61: 791-803.
- Kirsch, I. et al. 2002. The emperor's new drugs: an analysis of antidepressant medication data submitted to the U.S. Food and Drug Administration. *Prevention and Treatment*, 5.
- Kirsch, I., and M. J. Rosadino. 1993. Do double-blind studies with informed consent yield externally valid results? An empirical test. *Psychopharmacology*, 110: 437-42.
- Krishna, D. 1971. "The self-fulfilling prophecy" and the nature of society. *American Sociological Review*, 36: 1104-7.
- Levine, J.D. et al. 1981. Analgesic responses to morphine and placebo in individuals with postoperative pain. *Pain*, 10: 379-89
- MacKenzie, D. A., 2006, *An Engine, Not a Camera: How Financial Models Shape Markets*, Cambridge, MA and London: MIT Press
- Madon, S., L. Jussim, and J. Eccles. 1997. In search of the powerful self-fulfilling prophecy. *Journal of Personality and Social Psychology*, 72: 791-809.
- Merton, R. K. 1936. The unanticipated consequences of purposive social action. *American Sociological Review*, 1: 894-904.
- — —. 1968 [1948]. The self-fulfilling prophecy. Pp. 475-90 of *Social Theory and Social Structure*, 2nd ed. New York: Free Press.
- Miller, C., 1961. The self-fulfilling prophecy: a reappraisal. *Ethics*, 72: 46-51.
- Popper, K. 1957, *The Poverty of Historicism*. London and New York: Routledge.
- Raudenbush, S. W. 1984. Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: a synthesis of findings from 18 experiments. *Journal of Educational Psychology*, 76: 85-97.
- Rosenthal, R. and L. Jacobson. 1968. *Pygmalion in the Classroom: Teacher Expectation and Pupils' Intellectual Development*. New York: Holt, Rinehart and Winston.
- Salganik, M. J. and D. J. Watts. 2007. An experimental approach to self-fulfilling prophecies in cultural markets. Unpublished.

- Schwert, G. W. 2003. Anomalies and market efficiency. Pp. 937-72 in *Handbook of the Economics of Finance*, vol. 1A, edited by G. M. Constantinides, M. Harris, and R. Stulz. Amsterdam and London: Elsevier / North-Holland.
- Shapiro, A. K. and E. Shapiro. 1997. *The Powerful Placebo: From Ancient Priest to Modern Physician*. Baltimore: Johns Hopkins University Press.
- Shiller, R. J. 2000. *Irrational Exuberance*. Princeton, N.J.: Princeton University Press.
- Smith, A. E., L. Jussim, and J. Eccles. 1999. Do self-fulfilling prophecies accumulate, dissipate, or remain stable over time? *Journal of Personality and Social Psychology*, 77: 548-65.
- Spitz, H. H. 1999. Beleaguered *Pygmalion*: a history of the controversy over claims that teacher expectancy raises intelligence. *Intelligence*, 27: 199-234.
- Steele, C. M. and J. Aronson. 1995. Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, 69: 797– 811.
- Stott, C. and S. Reicher. 1998. How conflict escalates: the inter-group dynamics of collective football “violence.” *Sociology*, 32: 353-77.
- Steward-Williams, S. and J. Podd. 2004. The placebo effect: dissolving the expectancy versus conditioning debate. *Psychological Bulletin*, 130: 324-40.
- Vase, L., J. L. Riley III, and D. D. Price 2002. A comparison of placebo effects in clinical analgesic trials versus studies of placebo analgesia. *Pain*, 99: 443-52.
- Whyte, W. F. 1943. *Street Corner Society: The Social Structure of an Italian Slum*, 4th ed. Chicago: University of Chicago Press, 1993.
- Zuckoff, M. 2005. *Ponzi's Scheme: The True Story of a Financial Legend*. New York: Random House.

Table 1: Examples of SFP

Placebo response

- (1) Y believes that “I have received a treatment that relieves pain”
- (2) Because of (1), Y experiences pain relief

Interpersonal expectancy

- (1) X believes that “Y has great ability”
- (2) X therefore gives Y challenging material and communicates high expectations
- (3) Because of (2), Y performs well on test

Reactive conflict

- (1) X believes that “Y is aggressive”
- (2) X therefore attacks Y
- (3) Because of (2), Y attacks X

Bank run

- (1.1) X₁ believes that “Y is insolvent”
- (2.1) X₁ therefore withdraws deposits
 - (1.2) Because of (2.1), X₂ believes that “Y is insolvent”
 - (2.2) X₂ therefore withdraws deposits
 - (1.3) Because of (2.1) and (2.2), X₃ believes that “Y is insolvent”
 - (2.3) X₃ therefore withdraws deposits
- (3) Because of (2.1), (2.2), (2.3) ..., Y fails

Investment bubble

- (1.1) X₁ believes that “Y generates high returns”
- (2.1) X₁ therefore invests in Y
 - (1.2) Because of (2.1), X₂ believes that “Y generates high returns”
 - (2.2) X₂ therefore invests in Y
- (3.1) Because of (2.2), Y pays high returns to X₁
 - (1.3) Because of (2.1), (2.2), and especially (3.1), X₃ believes that “Y generates high returns”
 - (2.3) X₃ therefore invests in Y
- (3.2) Because of (2.3), Y pays high returns to X₂
- ...

Social theory

- (1) X₁ ... X_n believe that “Y is a true model of society”
- (2) X₁ ... X_n therefore act accordingly
- (3) Because of (2), society conforms to Y

Figure 1: Sums invested in Ponzi's scheme, 1920
(source: Zuckoff 2005)

