The Amazing Success of Statistical Prediction Rules

Judgment problems great and small are an essential part of everyday life. What menu items will I most enjoy eating? Is this book worth reading? Is the boss in a good mood? Will the bungee cord snap? These and other common judgment problems share a similar structure: On the basis of certain cues, we make judgments about some target property. I doubt the integrity of the bungee cord (target property) on the basis of the fact that it looks frayed and the assistants look disheveled and hungover (cues). How we make and how we ought to make such evidence-based judgments are interesting issues in their own right. But they are particularly pressing because such predictions often play a central role in decisions and actions. Because I don't trust the cord, I don't bungee jump off the bridge.

Making accurate judgments is important for our health and happiness, but also for the just and effective operation of many of our social institutions. Judgments about whether someone will become violent can determine whether that person loses their freedom by being involuntarily committed to a psychiatric institution. Predictions about whether a prisoner if set free will commit violence and mayhem can determine whether he is or is not paroled. Judgments about a student's academic abilities play a role in determining the quality of medical school or law school she goes to, or even whether she gets to study law or medicine at all. Judgments about a person's future financial situation can determine whether they receive loans to make large purchases; such judgments can also determine whether they receive the most attractive loans available. And most

everyone who has ever held a job has had others pass judgments about their trustworthiness, intelligence, punctuality, and industriousness.

It is hard to overestimate the practical significance of these sorts of social judgments. Using reasoning strategies that lead to unreliable judgments about such matters can have devastating consequences. Unnecessarily unreliable judgments can lead to decisions that waste untold resources, that unjustly deprive innocent people of their freedom, or that lead to preventable increases in rape, assault, and murder. There is a difference between cancer and horseshoes, between prison and a good shave. For many reasoning problems, ''close enough'' isn't good enough. Only the best reasoning strategies available to us will do. Ameliorative Psychology is designed to identify such strategies, and the primary tasks of a useful epistemology are to articulate what makes a reasoning strategy a good one and to carry that message abroad so that improvements can be implemented. This chapter is the prologue to that epistemological message.

Who could possibly deny that those charged with making high-stakes decisions should reason especially carefully about them? Consider, for example, predictions about violent recidivism made by parole boards. Who could deny that members of parole boards should scrupulously gather as much relevant evidence as they can, carefully weigh the different lines of evidence, and on this basis come to a judgment that is best supported by the entirety of the evidence? Actually, we deny this. We contend that it would often be much better if experts, when making high-stakes judgments, ignored most of the evidence, did not try to weigh that evidence, and didn't try to make a judgment based on their long experience. Sometimes, it would be better for the experts to hand their caseload over to a simple formula that a smart 8-year-old could solve and then submit to the child's will. This is what Ameliorative Psychology counsels. (Of course, discovering such a formula takes some expertise.)

For the past half century or so, psychologists and statisticians have shown that people who have great experience and training at making certain sorts of prediction are often less reliable than (often very simple) Statistical Prediction Rules (SPRs). This is very good news, especially for those of us who like to do hard work without having to work hard. Of course, the philosophical literature is full of fantastic examples in which some simple reasoning strategy that no reasonable person would accept turns out to be perfectly reliable (e.g., ''believe all Swami Beauregard's predictions''). But we are not engaged here in Freak Show Philosophy. Many SPRs are robustly successful in a wide range of real-life reasoning problems—including some very high-stakes ones. Further, the success of

some SPRs seems utterly miraculous. (In fact, when we introduced one of the more shocking SPR results described below to a well-known philosopher of psychology who is generally sympathetic to our view, he simply didn't believe it.) But there are general reasons why certain kinds of SPRs are successful. We turn now to describing their success. Later, we'll try to explain it.

1. The success of SPRs

We have coined the expression 'Ameliorative Psychology' to refer to the various empirical work that concerns itself with passing normative judgments on reasoning strategies and prescribing new and better ways to reason. In this chapter, we will introduce what we take to be the two main branches of Ameliorative Psychology. In section 1, we will describe some of the shocking findings of the predictive modeling literature; and in section 2, we will try to explain some of these findings. In section 3, we will briefly explore the other main branch of Ameliorative Psychology—the psychological investigation into how people tend to reason about everyday matters.

1.1. Proper linear models

A particularly successful kind of SPR is the *proper linear model* (Dawes 1982, 391). Proper linear models have the following form:

$$P = w_1c_1 + w_2c_2 + w_3c_3 + w_4c_4$$

where c_n is the value for the n^{th} cue, and w_n is the weight assigned to the n^{th} cue. Our favorite proper linear model predicts the quality of the vintage for a red Bordeaux wine. For example, c_1 reflects the age of the vintage, while c_2, c_3, and c_4 reflect climatic features of the relevant Bordeaux region. Given a reasonably large set of data showing how these cues correlate with the target property (the market price of mature Bordeaux wines), weights are then chosen so as to best fit the data. This is what makes this SPR a *proper* linear model: The weights optimize the relationship between P (the weighted sum of the cues) and the target property as given in the data set. A wine predicting SPR was developed by Ashenfelter, Ashmore, and Lalonde (1995). It has done a better job predicting the price of mature Bordeaux red wines at auction (predicting 83% of the variance)

than expert wine tasters. Reaction in the wine-tasting industry to such SPRs has been "somewhere between violent and hysterical" (Passell 1990).

Whining wine tasters might derive a small bit of comfort from the fact that they are not the only experts trounced by a mechanical formula. We have already introduced *The Golden Rule of Predictive Modeling*: When based on the same evidence, the predictions of SPRs are at least as reliable as, and are typically more reliable than, the predictions of human experts for problems of social prediction. The most definitive case for the Golden Rule has been made by Grove and Meehl (1996). They report on an exhaustive search for studies comparing human predictions to those of SPRs in which (a) the humans and SPRs made predictions about the same individual cases and (b) the SPRs never had more information than the humans (although the humans often had more information than the SPRs). They

> found 136 studies which yielded 617 distinct comparisons between the two methods of prediction. These studies concerned a wide range of predictive criteria, including medical and mental heath diagnosis, prognosis, treatment recommendations and treatment outcomes; personality description; success in training or employment; adjustment to institutional life (e.g., military, prison); socially relevant behaviors such as parole violation and violence; socially relevant behaviors in the aggregate, such as bankruptcy of firms; and many other predictive criteria. (1996, 297)

Of the 136 studies, 64 clearly favored the SPR, 64 showed approximately equivalent accuracy, and 8 clearly favored the clinician. The 8 studies that favored the clinician appeared to have no common characteristics; they "do not form a pocket of predictive excellence in which clinicians could profitably specialize" (299). What's more, Grove and Meehl argue plausibly that these 8 outliers are likely the result of random sampling errors (i.e., given 136 chances, the better reasoning strategy is bound to lose *sometimes*) "and the clinicians' informational advantage in being provided with more data than the actuarial formula" (298).

There is an intuitively plausible explanation for the success of proper linear models. Proper linear models are constructed so as to best fit a large set of (presumably accurate) data. But the typical human predictor does not have all the correlational data easily available; and even if he did, he couldn't perfectly calculate the complex correlations between the cues and the target property. As a result, we should not find it surprising that proper linear models are more accurate than (even expert) humans. While

this explanation is intuitively satisfying, it is mistaken. To see why, let's look at the surprising but robust success of some *improper* linear models.

1.2. Bootstrapping models: Experts vs. virtual experts

A *proper* linear model assigns weights to cues so as to optimize the relationship between those cues and the target property in a data set. Improper linear models do not best fit the available data. Bootstrapping models are perhaps the most fascinating kind of improper linear models. These are proper linear models of a person's judgments. Goldberg (1970) constructed the classic example of a bootstrapping model. Many clinical psychologists have years of training and experience in predicting whether a psychiatric patient is neurotic or psychotic on the basis of a Minnesota Multiphasic Personality Inventory (MMPI) profile. The MMPI profile consists of 10 clinical (personality) scales and a number of validity scales. Goldberg asked 29 clinical psychologists to judge, only on the basis of an MMPI profile, whether a patient would be diagnosed as neurotic or psychotic. Goldberg then constructed 29 proper linear models that would mimic each psychologist's judgments. The predictor cues consisted of the MMPI profile; the target property was the psychologist's predictions. Weights were assigned to the cues so as to best fit *the psychologist's judgments* about whether the patient was neurotic or psychotic. So while a bootstrapping model is a proper linear model of a human's judgments, it is an improper linear model of the target property—in this case, the patient's condition.

One might expect that the bootstrapping model would predict reasonably well. It is built to mimic a fairly reliable expert, so we might expect it to do nearly as well as the expert. In fact, *the mimic is more reliable than the expert*. Goldberg found that in 26 of the 29 cases, the bootstrapping model was more reliable in its diagnoses than the psychologist on which it was based! (For other studies with similar results, see Wiggins and Kohen 1971, Dawes 1971.) This is surprising. The bootstrapping model is built to ape an expert's predictions. And it will occasionally be wrong about the expert. But when it is wrong about the expert, it's more likely to be right about the target property!

At this point, it is natural to wonder why the bootstrapping model is more accurate than the person on which it is based. In fact, it seems paradoxical that this could be true: If the bootstrapping model "learns" to predict from an expert, how can the model "know" more than the expert?

This way of putting the finding makes it appear that the model is adding some kind of knowledge to what it learns from the expert. But how on earth can that be? The early hypothesis for the success of the bootstrapping model was not that the model was adding something to the expert's knowledge (or reasoning competence), but that the model was adding something to the expert's reasoning performance. In particular, the hypothesis was that the model did not fall victim to performance errors (errors that were the result of lack of concentration or a failure to properly execute some underlying predictive algorithm). The idea was that bootstrapping models somehow capture the underlying reliable prediction strategy humans use; but since the models are not subject to extraneous variables that degrade human performance, the models are more accurate (Bowman 1963, Goldberg 1970, Dawes 1971). This is a relatively flattering hypothesis, in that it grants us an underlying competence in making social judgments. Unfortunately, this flattering hypothesis soon came crashing down.

1.3. Random linear models

Dawes and Corrigan (1974) took five bootstrapping experiments and for each one constructed a *random* linear model. Random linear models do not pretend to assign optimum weights to variables. Instead, random weights are assigned—with one important caveat: All the cues are defined so they are positively correlated with the target property. They found that the random linear models were as reliable as the proper models and more reliable than human experts. Recall we said that there was an SPR finding that was denied by a well-known philosopher of psychology. This is it. This philosopher is not alone. Dawes has described one dominant reaction to the success of random linear models: "[M]any people didn't believe them—until they tested out random . . . models on their own data sets" (Dawes 1988, 209, n. 17).

The resistance to this finding is understandable (though, as we shall later argue, misguided). It is very natural to suppose that people who make predictions are in some sense "calculating" a suboptimal formula. (Of course, the idea isn't that the person explicitly calculates a complex formula in order to make a prediction; rather, the idea is that there will be an improper formula that simulates the person's weighing of the various lines of evidence in making some prediction.) Since we can't calculate in our heads the optimum weights to attach to the relevant cues, it's understandable that *proper* models outperform humans. This picture of humans "calculating" suboptimal formulas, of implicitly using improper models,

also fits with the optimistic explanation of the bootstrapping effect. A bootstrapping model approximates the suboptimal formula a person uses—but the bootstrapping model doesn't fall victim to performance errors to which humans are prone. So far, so good. But how are we to understand random linear models outperforming expert humans? After all, if experts are calculating some sort of suboptimal formula, how could they be defeated by a formula that uses weights that are both suboptimal and random? Surely we must do better than linear models that assign just any old weights at all. But alas, we do not. Without a plausible explanation for this apparent anomaly, our first reaction (and perhaps even our well-considered reaction) may be to refuse to believe this could be true.

1.4. Unit weight models

Among the successful improper linear models, there is one that tends to be a bit more reliable and easier to use than the others. Unit weight models assign equal weights to (standardized) predictor cues, so that each cue has an equal "say" in the final prediction. Our favorite example of a unit weight model is what we might call the "F minus F Rule." Howard and Dawes (1976) found a very reliable, low-cost reasoning strategy for predicting marital happiness. Take the couple's rate of lovemaking and subtract from it their rate of fighting. If the couple makes love more often than they fight, then they'll probably report being happy; if they fight more often than they make love, then they'll probably report being unhappy. Howard and Dawes tested their hypothesis on data compiled by Alexander (1971) in which 42 couples "monitored when they made love, when they had fights, when they had social engagements (e.g., with in-laws), and so on. These subjects also made subjective ratings about how happy they were in their marital or coupled situation" (Dawes 1982, 393). The results were interesting: "In the thirty happily married couples (as reported by the monitoring partner) only two argued more often than they had intercourse. All twelve of the unhappily married couples argued more often" (478). The reliability of the F minus F Rule was confirmed independently by Edwards and Edwards (1977) and Thornton (1977).

The F minus F Rule exhibits three advantages of unit weight SPRs. First, it requires attention to only a slim portion of the available evidence. We can ignore the endless variety of psychological and behavioral quirks and incompatibilities that married people can exhibit and instead focus on two relatively simple, straightforward (though personal) cues. Second, the

F minus F Rule is very simple to use. There is no need to try to weigh different complex cues against each other. For example, there is no need to guess whether the (presumably) negative sign that the partners have different approaches to finances is outweighed by the (presumably) positive sign that both had happily married parents. Third, the F minus F Rule is known to be quite reliable.

Given the success of unit weight models, Paul Meehl has said, ''In most practical situations an unweighted sum of a small number of 'big' variables will, on the average, be preferable to regression equations'' (quoted in Dawes and Corrigan 1974, 105). Dawes and Corrigan succinctly state the cash value of these results: To be more reliable than expert humans in the social arena, ''the whole trick is to know what variables to look at and then know how to add'' (1974, 105).

1.5. SPRs vs. Humans: An unfair test?

Before we turn to an explanation for the success of SPRs, we should consider a common objection against the SPR findings described above. The objection proceeds as follows: ''The real reason human experts do worse than SPRs is that they are restricted to the sort of objective information that can be plugged into a formula. So of course this tilts the playing field in favor of the formula. People can base their predictions on evidence that can't be quantified and put in a formula. By denying experts this kind of evidence, the above tests aren't fair. Indeed, we can be confident that human experts will defeat SPRs when they can use a wider range of real world, qualitative evidence.''

There are three points to make against this argument. First, this argument offers no actual *evidence* that might justify the belief that human experts are handicapped by being unable to use qualitative evidence in the above examples. The argument offers only a speculation. Second, it is possible to quantitatively code virtually any kind of evidence. For example, consider an SPR that predicts the length of hospitalization for schizophrenic and manic-depressive patients (Dunham and Meltzer 1946). This SPR employs a rating of the patients' insight into their condition. Prima facie, this is a subjective, nonquantitative variable because it relies on a clinician's diagnosis of a patient's mental state. Yet clinicians are able to quantitatively code their diagnoses of the patient's insight into his or her condition. The clinician's quantitatively coded diagnosis is then used by the SPR to make more accurate predictions than the clinician. Third, the

speculation that humans armed with "extra" qualitative evidence can outperform SPRs has been tested and has failed repeatedly. One example of this failure is known as the *interview effect*: Unstructured interviews degrade human reliability (Bloom and Brundage 1947, DeVaul et al. 1957, Oskamp 1965, Milstein et al. 1981). When gatekeepers (e.g., hiring and admissions officers, parole boards, etc.) make judgments about candidates on the basis of a dossier and an unstructured interview, their judgments come out worse than judgments based simply on the dossier (without the unstructured interview). So when human experts and SPRs are given the same evidence, and then humans get more information in the form of unstructured interviews, clinical prediction is *still* less reliable than SPRs. In fact, as would be expected given the interview effect, giving humans the "extra" qualitative evidence actually makes it easier for SPRs to defeat the predictions of expert humans. To be fair, however, there are cases in which experts can defeat SPRs. We will discuss these exceptions below.

2. Why do SPRs work?

There is an aura of the miraculous surrounding the success of SPRs. But even if there is no good explanation for their relative success, we ought to favor them over human judgment on the basis of performance alone. After all, the psychological processes we use to make complex social judgments are just as mysterious as SPRs, if not more so. Further, there is no generally agreed upon explanation for why our higher-level cognitive processes have the success that they do. (Indeed, there is even disagreement about just how successful they are; see, for example, Cohen 1981 and Piatelli-Palmarini 1994.) It might be that given our current understanding, replacing human judgment with an SPR may inevitably involve replacing one mystery for another—but the SPR is a mystery with a better track record.

2.1. The flat maximum principle

Let's suppose we have an explanation for the success of *proper* linear models. It would be natural to suppose we still had a lot of work to do coming up with an explanation for the success of *improper* linear models. But that's not true. Interestingly enough, it turns out that anyone who explains the success of *proper* linear models for problems of human and social prediction gets for free the explanation of the success of *improper* linear models. That's because for certain kinds of problem, the success of

improper models rides piggy-back on the success of proper models. Recall the passage quoted above in which Dawes reports that many people didn't believe his results concerning the success of improper linear models. Here it is in its entirety:

> The results when published engendered two responses. First, many people didn't believe them—until they tested out random and unit models on their own data sets. Then, other people showed that the results were trivial, because random and unit linear models will yield predictions highly correlated with those of linear models with optimal weights, and it had already been shown that optimal linear models outperform global judgments. I concur with those proclaiming the results trivial, but not realizing their triviality at the time, I luckily produced a "citation classic"—and without being illustrated with real data sets, the trivial result might never have been so widely known. (1988, 209, n. 17)

The reason some people argued that Dawes's results were trivial was because of a fascinating finding in statistics called *the flat maximum principle* (for a good nontechnical explanation, see Lovie and Lovie 1986; for a more technical introduction, see Einhorn and Hogarth 1975). (Einhorn and Hogarth in fact show there are not uncommon situations in which the improper unit weight models will be *more* reliable than the proper models. This is in part the result of the overfitting problem; i.e., the proper model "fits" some of the random, unrepresentative peculiarities of the data set on which it is constructed and is therefore less accurate on future data points than an improper model.)

The flat maximum principle says that for a certain class of prediction problems, as long as the signs of the coefficients are right, any linear model will predict about as well as any other. It is important to recognize that the flat maximum principle is restricted to certain kinds of problems. In particular, it applies only to problems in which the following conditions obtain:

1. The judgment problem has to be difficult. The problem must be such that no proper model will be especially reliable because the world is messy. Perhaps the best way to understand this is to visualize it. A linear model tries to draw a line through a bunch of data points. Suppose the points are quite spread out so that no single line can get close to all of them. Two things are intuitively obvious: (a) The best line through those points won't be *that* much better than lots of lines close to it. (b) The best line through those points might not be the best line through the next set of spread-out data points that comes down the pike. For example, consider the attempt to predict what an applicant's academic

performance in college might be. Even the best models are not exceptionally reliable. *No one* can predict with great accuracy who is and who is not going to be academically successful in college. A big part of the reason is colloquially expressed: Stuff happens. Two candidates who are identical on paper might have quite different academic careers for a multitude of unpredictable reasons.

2. The evidential cues must be reasonably predictive. The best cues for predicting academic performance (GPA, test scores) are reasonably predictive. Certainly, you'll do better than chance by relying on these cues.
3. The evidential cues must be somewhat redundant. For example, people with higher GPAs tend to have higher test scores.

Problems of social judgment—who is going to succeed in a job, who is going to commit another violent act, what football teams are going to win next weekend—tend to share these features. As a result, for problems of social judgment, improper models will be about as reliable as proper models.

Okay, so the success of improper linear models rides piggy-back on the success of proper linear models for problems of social prediction. So then what explains the success of proper linear models?

2.2. Condorcet to the rescue?

Condorcet's jury theorem, in its simplest form, says that if a jury is facing a binary choice and each jury member makes her decision independently and has a better-than-even chance of making the right decision, a simple majority of the jurors is likely to make the right decision, and this will tend toward certainty as the number of jurors tends toward infinity. We can think of the successful linear models we have introduced as a jury: The jury must make a binary decision about a target, and each jury member makes her decision on the basis of a single piece of evidence. Each piece of evidence correlates positively with the target; so each juror's decision is going to be right more often than not. And the linear model adds together each juror's judgment to come to a final decision about the target. The only difference between the different types of models is that some weigh certain lines of evidence more than others. Putting this in terms of our jury analogy, some models have more jurors focusing on certain lines of evidence than others. So given Condorcet's jury theorem, we should expect linear models to predict reasonably well. (Thanks to Michael Strevens and Mark Wunderlich for suggesting this explanation.)

The Condorcet explanation leaves open at least two questions. First, many successful linear models consist of a small number of cues (sometimes as few as two). But Condorcet's jury theorem suggests that high reliability usually requires many jurors. So the success of linear models still seems a bit mysterious. Second, why are linear models, particularly those with a very small number of cues, more reliable than human experts? After all, if human experts are able to use a larger number of reliable cues than simple linear models, why doesn't the Condorcet explanation imply that they will typically be more reliable than the models? We will address these questions in section 3. But for now, let's turn to a different explanation for the success of linear models.

2.3. An alternative hypothesis: The world we care about consists of mostly monotone interactions

Reid Hastie and Robyn Dawes have offered a different account of the success of linear models (2001, 58–62; see also Dawes 1988, 212–15). Their explanation comes in three parts. Since we embrace and elaborate on the third part of their explanation in section 3, we will focus only on the first two parts of their explanation here. The first part of their explanation for the success of SPRs is a principle about the relationship between proper linear models and the world: *Proper linear models can accurately represent monotone (or "ordinal") interactions.* We have already introduced linear models—they are models in which the judgment made is a function of the sum of a certain number of weighted variables. The best way to understand what monotone interactions are is to consider a simple example. Suppose a doctor has told you to reduce your body fat, and she recommends a special diet D and an exercise regime E. Now, let's suppose that D alone, without the exercise regime, is effective at reducing body fat. This would be the diet's *main effect.* Suppose also that the exercise regime alone, without the diet, is also effective at reducing body fat. Again, this would be the *main effect* of exercise. Now let's suppose Sam goes on the diet D and the exercise regime E. If Sam gets the benefits of both—the main effect of D and the main effect of E—then the interaction of D and E is monotone. If, however, Sam gets the main effects of both plus an extra benefit, then the interaction is not monotone. The extra benefit is often called an *interaction effect.*

If we continue this absurdly simplistic example, it will be easy to see why proper linear models can accurately represent monotone interactions.

Suppose that for a certain population of people, D will bring a loss of $\frac{1}{2}$ pound per week while E will bring a loss of $\frac{3}{4}$ pound per week. The following linear model will predict how much weight loss one can expect:

$$W = \frac{1}{2}d + \frac{3}{4}e$$

where W is the number of pounds lost, d is the number of weeks on the diet, and e is the number of weeks on the exercise regimen. It should be clear that a proper linear model will do a reasonably good job of predicting interactions that are not monotone, but for which the interaction effects are not strong.

The second part of the Hastie-Dawes explanation is a speculation about the world: *In practical social settings (where linear models have proven most successful), interactions are, near enough and in the main, monotone.* Those who study complex systems, nonlinear dynamics, and catastrophe theory will note that not all of the world we're interested in consists of monotone interactions. The idea is that as long as we are not looking for SPRs to predict the performance of nonlinear systems, linear models may perform well—better than human experts. By restricting the explanation of the success of linear models to practical, social settings, Hastie and Dawes can take advantage of the flat maximum principle. From the reliability of *proper* linear models, they can employ the flat maximum principle to infer the reliability of *improper* linear models as well.

We have doubts about the Hastie-Dawes explanation for the success of SPRs. Consider the linear model that represents the monotone weight loss interaction. The reason this linear model is reliable is that it accurately portrays the main causal agents and the relative influence of those agents in subjects' weight loss. But the robust reliability of SPRs can't depend on their reasonably accurate portrayal of causal reality. The reason is quite simply that many SPRs are not even close to accurate portrayals of reality. Consider a linear model that predicts academic performance on the basis of grade point average and test scores. The student's college GPA is not a primary cause of her graduate school performance; same with her test score. Rather, it is much more plausible to suppose that whatever complex of factors goes into a student's GPA and test scores is also heavily implicated in a student's success in graduate school. (Recall that the flat maximum principle is operative when the cues employed by a linear model are redundant.) So it seems unlikely that the success of SPRs depends on their mirroring or reflecting monotone interactions. (Thanks to Michael Strevens for this point.)

We need to be a bit careful here. We're not suggesting that we oppose or doubt the possibility of successful SPRs that identify causes. (Just the opposite.) Nor are we suggesting that successful SPRs do not depend for their success on causal regularities. (Again, just the opposite.) Our point is that even when we can't "read off" anything like the causal structure of the world from an SPR, it can still be highly reliable and worthy of being used. If that's so, then the success of SPRs can't depend on their representing (even approximately) the interactions that produce the item of interest.

3. The foibles of human prediction

In our philosophical circles, we're considered good athletes—well, okay, we used to be considered good athletes. Compared to our nonacademic friends, however, we have always aspired to athletic mediocrity. It may be that the success of SPRs is like our athletic success—apparent only when measured against earnest but rather undistinguished competition. (We could put the point more bluntly, but we're talking about our friends here.) The right question to ask might not be "Why are SPRs so good at prediction?" but rather "Why are we so bad at prediction?" There is a large and fascinating literature on this topic (Nisbett and Ross 1980; Gilovich 1991; Hastie and Dawes 2001). We can hit some of the high points of this literature by noting that in order to develop reliable reasoning strategies for problems of social judgment, it is typically necessary (a) to be able to determine which cues are most predictive, which requires detecting correlations between potential cues and the target property; (b) to be able to attend to and remember all those cues; (c) to be able to combine them appropriately; and (d) to get accurate feedback on one's judgments. As we shall see, we have considerable difficulty with each of these stages.

3.1. Covariation illusions

In order to reason well about social matters, we need to be able to reliably detect correlations. But in a classic series of studies, Chapman and Chapman (1967, 1969) found that we can be quite bad at this on tasks that represent the ordinary challenges facing us. We often don't recognize covariations that exist, particularly when they do not conform to our background beliefs; and we often report covariations where there are none, particularly when we expect there to be covariation. In the past, many psychologists used Draw-a-Person (or DAP) tests to make initial diagnoses.

It was thought that patients' disorders could be diagnosed from their drawings of people. For example, it was thought that paranoid patients would draw large eyes; the drawings of impotent patients would emphasize male genitalia or would be particularly macho. By the mid-1960s, it was well known that DAP tests were bunk. There are no such correlations. And yet clinicians continued to use them. Chapman and Chapman (1967) asked clinicians who used the DAP test to describe the features of patients' drawings they thought were associated with six diagnoses. Once they had these reports, Chapman and Chapman obtained 45 DAP drawings made by patients in a state hospital and randomly paired those drawings with the six diagnoses. Each drawing-diagnosis pair was then presented to introductory psychology students for 30 seconds, and then the students were asked to report which features of the drawings were most frequently associated with each diagnosis. Even though there were no systematic relationships in the data, subjects claimed to detect covariations. Further, they were virtually the same covariations the clinicians claimed to find in real data! It is plausible to suppose in this case that widely shared background assumptions (or perhaps just thoughtless stereotypes) led both expert clinicians and naïve subjects to "see" covariations in data that simply weren't there. Interestingly, when Chapman and Chapman built in massive negative covariations between the features of the drawings and the diagnoses subjects were likely to make, naïve subjects still reported positive covariations—though somewhat reduced in magnitude.

In another fascinating study, Chapman and Chapman focused on the famous Rorschach test. While most of the associations clinicians have believed they detected in Rorschach tests are actually not present, it turns out that two responses to the Rorschach test are correlated with male homosexuality. However, these responses are not particularly "face valid" (i.e., they do not strike most people as particularly intuitive). For example, male homosexuals are not more likely to identify in the Rorschach blots feminine clothing, anuses or genitalia, or humans with confused or uncertain sexes. In fact, homosexual men more frequently report seeing monsters on Card IV and a part-human-part-animal on Card V. (Again, Chapman and Chapman found that clinicians of the day believed there was a significant correlation between the "face valid" signs and homosexuality. Only 2 of the 32 clinicians they polled even listed one of the valid signs.) Naïve subjects (1969) were given 30 cards with traits (homosexual or nonhomosexual) on one side and Rorschach responses on the other (a valid sign, an invalid but "face valid" sign, or a filler sign) and were given 60 seconds to review each card. Even though the cards contained no correlations

between the traits and the Rorschach responses, subjects reported frequent correlations between the "face valid" signs and homosexuality. This finding essentially replicates the DAP test result.

Next, Chapman and Chapman changed the cards so that the valid signs were associated more often with homosexuality than were the other signs. Even when the valid signs were associated with homosexuality 100% of the time, naïve observers failed to detect the covariation. So it's not just that subjects see correlations when there are none. In fact, we often don't see correlations that are actually there, and sometimes we see positive correlations when in fact the correlations are negative.

It should be noted that Chapman and Chapman did not draw particularly pessimistic conclusions from their experiments. Nor do we. In fact, when Chapman and Chapman took out the misleading invalid signs, subjects were capable of detecting the actual covariations in the data. Nisbett and Ross (1980) draw the following conclusion from these experiments:

> [R]eported covariation was shown to reflect true covariation far less than it reflected theories or preconceptions of the nature of the associations that "ought" to exist. Unexpected, true covariations can sometimes be detected but they will be underestimated and are likely to be noticed only when the covariation is very strong, and the relevant data set excludes "decoy features" that bring into play popular but incorrect theories. (97)

When it comes to social judgment, the evidential situation is likely to be quite complex—with many signs that are valid but counterintuitive and other signs that are "face valid" but not predictive. In such an environment, we are not likely to do a particularly good job of detecting covariations. And so, unless the theories, background assumptions, and stereotypes we bring to a particular prediction are accurate, we are not likely to be very good at identifying what cues are most likely to covary with and so predict our target property.

3.2. Limits on memory, attention, and computation

In reasoning about social matters, we often attend to a number of different evidential cues. But we have certain cognitive limits, including limits on memory, attention, and computation, that could well be implicated in the relative unreliability of our social judgments. For example, we aren't very good at keeping even medium-sized amounts of information available in attention or memory when solving a problem (Bettman et al. 1990). And

this prevents us from making accurate predictions on the fly. On the received view, we attempt to arrive at a solution to a problem by searching the problem space. For many problems, the size of this space is cognitively unmanageable; the problem space contains more information than the electric flesh between our ears can handle at one time. Take the example of chess. If the goal is to checkmate your opponent, in the early stages of the game the solution search space is enormous. How do people make the problem tractable? They adopt a strategy that navigates a limited path through the search space, a heuristic that identifies a small number of plausible (rather than all possible) strategies to secure a solution (Newell and Simon 1972).

Daily life confirms that our memory is limited. (We seem to get more confirmation as we grow older!) It also confirms that our attention is limited. The so-called ''central limited capacity of attention'' principle has been a basic premise of the last 40 years of research on attention. In the classic divided-attention experiments, observed decrements in performance are explained in terms of limitations on internal processing (van der Heijden, 1998). If limitations on attention and memory produce regrettable performance in simple tasks, why should we suppose that we can, without fear of embarrassment, use the same feeble tools to accurately evaluate complex issues of social judgment?

Even if we knew what cues to look for and we could remember them and we could attend to them, we often find it very difficult to combine those cues effectively. Paul Meehl makes this point starkly by focusing on a familiar example:

> Surely we all know that the human brain is poor at weighting and computing. When you check out at a supermarket, you don't eyeball the heap of purchases and say to the clerk, ''Well it looks to me as if it's about $17.00 worth; what do you think?'' The clerk adds it up. There are no strong arguments from the armchair or from empirical studies . . . for believing that human beings can assign optimal weights in equations subjectively or that they apply their own weights consistently. (Meehl 1986, 372)

Notice that in Meehl's grocery example, we know that a simple addition is the right calculation to apply, and the variable values (i.e., the prices) are usually stamped right on the products. But suppose that the computation required was much more complex. This of course would make matters even worse:

> Suppose instead that the supermarket pricing rule were, ''Whenever both beef and fresh vegetables are involved, multiply the logarithm of 0.78 of the meat price by the square root of twice the vegetable price''; would the clerk

> and customer eyeball any better? Worse, almost certainly. When human judges perform poorly at estimating and applying the parameters of a simple or component mathematical function, they should not be expected to do better when required to weigh a complex composite of those variables. (Dawes, Faust, and Meehl 1989, 1672)

So when it comes to problems of social judgment, we have trouble discovering the right correlations, remembering their values, attending to more than just a few of them, and combining the values appropriately to render a judgment. If this is right, if the basis of our social judgments are riddled with error and limitations, then why do most people seem to have so much success in the social world? The sobering answer is probably that most of us have less success in the social world than we think.

3.3. Lack of reliable feedback

Even if we don't start off making complex social judgments in a reliable fashion, we can at least hope to improve our judgments by receiving and acting on accurate feedback. If we can determine that a depressingly large number of our past judgments were mistaken (or even that a well-defined class of past predictions was mistaken), perhaps we can modify our reasoning strategies and so judge more accurately. (The fact that a person might have made such modifications might lead him to discount the pessimism we seem to be insisting upon here.) Unfortunately, there are a number of quite natural phenomena that keep us from getting accurate feedback on our past judgments and behaviors.

For many irrevocable decisions we make, the feedback we receive on our judgments is almost inevitably incomplete. Consider the grizzled philosopher who has played a major role in hiring a number of junior colleagues and who takes the interviews very seriously. Given the nature of the job market in philosophy, it's quite likely that his junior colleagues are, by and large, a pretty impressive lot. Given this feedback, he is likely to think quite highly of his ability to identify in interviews good young philosophers. The problem here is that the grizzled philosopher doesn't know whether his predictions would have turned out better or worse without the interviews. (And even if he did, it's unlikely he would have a large enough sample size to draw a reasonable conclusion.) Simply put, most gatekeepers don't have control groups to test the effectiveness of their reasoning strategies. After all, the set of junior colleagues who would have been hired without interviews (the control group) might have been even more terrific than his actual set of junior colleagues. The problem is

not just that most gatekeepers don't have control groups—that is often a practical inevitability. The problem is that they don't recognize that this is a serious problem. Most gatekeepers should probably have much more diffidence concerning their powers of prediction—especially in a job market in which most job seekers are something more than competent.

Another problem is that the feedback we get, especially when it comes to social matters, is likely to be highly unrepresentative. Consider the finding that 94% of university professors believe they're better-than-average at their jobs (Gilovich 1991, 77). One reason for this may be that we typically get personal feedback from students who think we were terrific teachers (or at least who say we were terrific teachers). Seldom will students go out of their way to make contact with professors they thought were really mediocre (if for no other reason than, where would they begin?).

The problem of unrepresentative feedback can be made vivid with an example that is likely familiar to everybody. Think about someone who employs mildly (or outright) annoying interpersonal strategies, for example, dominating conversations or name dropping. How likely are you to tell this person that these behaviors are annoying? Some blunt folk might always do so. But most of us, probably as a result of some combination of politeness, pusillanimity, and prudence, let it slide. Of course, we recognize that this behavior is annoying (or worse), and we might judge the person to be annoying (or worse). But given the feedback he has received, he might well go forth into the world confident that he has once again been socially deft, charming, and deeply impressive. (We are inclined to suggest you perform a public service. Supply accurate feedback. Call a bore a bore, a jerk a jerk, a blowhard a blowhard. Just don't do it to us.)

Even when the feedback we get is representative and shows that our predictions are mistaken, we will often interpret such feedback in a way that supports our preconceptions. For example, Gilovich (1983) asked people who gambled on football games to tape-record their thoughts about the outcomes of their bets. One might expect the gamblers to remember their wins and repress their losses. In fact, just the opposite occurred:

> [T]hey spent more time discussing their losses than their wins. Furthermore, the kind of comments made about wins and losses were quite different. The bettors tended to make "undoing" comments about their losses—comments to the effect that the outcome would have been different if not for some anomalous or "fluke" element. . . . In contrast, they tended to make "bolstering" comments about their wins—comments indicating that the outcome either should have been as it was, or should have been even more extreme in the same direction. . . . By carefully scrutinizing and explaining

> away their losses, while accepting their successes at face value, gamblers do indeed rewrite their personal histories of success and failure. Losses are often counted, not as losses, but as "near wins." (Gilovich 1991, 55)

One interesting feature of this common interpretative strategy is that the subject cannot be accused of ignoring negative evidence. In fact, the subject is attending more to the negative evidence than to the positive evidence. It's just that he interprets the positive evidence as positive, and the negative evidence as bad luck.

3.4. The basis of epistemic exceptionalism: The overconfidence feedback loop

Let's recap briefly. We aren't especially good at detecting the properties that covary with the target property we want to predict—especially when we have strong background opinions and when the informational situation is complex. We aren't especially good at recalling or attending to lots of different avenues of information. And often, the feedback we get about the quality of our judgments or behavior is unrepresentative (and we don't know it) or incomplete (and we don't see that this is a serious problem). As a result, it is not surprising that we aren't especially reliable in our judgments about complex social phenomena.

Against this background, the sluggish reception SPRs have received in the disciplines whose business it is to predict and diagnose is particularly puzzling. (Resistance to the use of SPRs is particularly strong when it comes to making social predictions. SPRs have found easier acceptance in non-psychiatric medical diagnosis.) In the face of a half century of studies showing the superiority of SPRs, many experts still base judgments on subjective impressions and unmonitored evaluation of the evidence. Resistance to the SPR findings runs deep and typically comes as a kind of *epistemic exceptionalism.* Those who resist the SPR findings take their reasoning powers to be exceptional, and so they defect from the judgments of SPRs when they find what they take to be exceptions to it. They are typically quite willing to admit that *in the long run*, SPRs will be right more often than human experts. But their (over)confidence in their subjective powers of reflection leads them to deny that we should believe the SPR's prediction *in this particular case.*

We suspect that epistemic exceptionalism, which we suggest has led to the sluggish reception of SPRs, is the result of two facts about people. When it comes to prediction, we find the success of SPRs hard to believe,

and we find our lack of success hard to believe. The reason we find our own lack of success hard to believe is that most of the failures of our predictive capacities are hidden from us. We don't see what's gone wrong. We don't detect the right covariations, but we think we do. We can't attend to the relevant complexities, but we think we have. We aren't getting representative feedback on our predictions, but we think we are. As a result, we tend to be overconfident about the power of our subjective reasoning faculties and about the reliability of our predictions (Trout 2002). Our faith in the reliability of our subjective powers of reasoning bolsters our (over)confidence in our judgments; and our (over)confident judgments bolster our belief in the reliability of our subjective faculties (Arkes 1991; Sieck and Arkes [unpublished manuscript]). Let's focus on each side of this overconfidence feedback loop.

The first side of the overconfidence feedback loop consists in overconfidence in our judgments. This overconfidence leads too often to defection from a successful SPR. That we fall victim to an overconfidence bias is one of the most robust findings in contemporary psychology:

> [A] large majority of the general public thinks that they are more intelligent, more fair-minded, less prejudiced, and more skilled behind the wheel of an automobile than the average person. . . . A survey of one million high school seniors found that 70% thought they were above average in leadership ability, and only 2% thought they were below average. In terms of ability to get along with others, *all* students thought they were above average, 60% thought they were in the top 10%, and 25% thought they were in the top 1%! Lest one think that such inflated self-assessments occur only in the minds of callow high-school students, it should be pointed out that a survey of university professors found that 94% thought they were better at their jobs than their average colleague. (Gilovich 1991, 77)

The overconfidence bias goes far beyond our inflated self-assessments. For example, Fischhoff, Slovic, and Lichtenstein (1977) asked subjects to indicate the most frequent cause of death in the U.S. and to estimate their confidence that their choice was correct (in terms of "odds"). When subjects set the odds of their answer's correctness at 100:1, they were correct only 73% of the time. Remarkably, even when they were so certain as to set the odds between 10,000:1 and 1,000,000:1, they were correct only between 85% and 90% of the time. It is important to note that the overconfidence effect is systematic (it is highly replicable and survives changes in task and setting) and directional (the effect is in the direction of over rather than underconfidence). But overconfidence is eliminated or

reversed when the questions are very easy. This phenomenon is known as the difficulty (or hard-easy) effect (Lichtenstein and Fischhoff 1977).

The second side of the overconfidence feedback loop consists of our overconfidence in the reliability of our subjective reasoning faculties. We are naturally disposed to exaggerate the powers of our subjective faculties. A very prominent example that we have already discussed is the interview effect. When gatekeepers avail themselves of unstructured interviews, they actually degrade the reliability of their predictions. Although the interview effect is one of the most robust findings in psychology, highly educated people ignore its obvious practical implication. We suspect that this occurs because of our confidence in our subjective ability to ''read'' people. We suppose that our insight into human nature is so powerful that we can plumb the depths of a human being in a 45-minute interview—unlike the lesser lights who were hoodwinked in the SPR studies. As we have said, a major reason our (over)confidence survives is because we typically don't get systematic feedback about the quality of our judgments (e.g., we can't compare the long-term outcomes of our actual decisions against the decisions we would have made if we hadn't interviewed the candidates). To put this in practical terms, the process by which most working philosophers were hired was seriously and, at the time, demonstrably flawed. This will be of no comfort to our colleagues, employed or unemployed. We expect, however, that the unemployed will find it considerably less surprising.

4. The tempting pleasures of broken legs

It doesn't matter how reliable a reasoning rule might be if a reasoner applies it poorly. There are two things the reasoner must do right. She must execute the strategy correctly (e.g., plug in the right values, perform the calculations properly), and she must apply the strategy to the right sorts of problems. It is not always easy to know whether it is appropriate to use a particular reasoning strategy in a particular case. This has come to be known as the broken leg problem, and here is a classical statement of it:

> Clinicians might be able to gain an advantage by recognizing rare events that are not included in the actuarial formula (due to their infrequency) and that countervail the actuarial conclusion. This possibility represents a variation of the clinical-actuarial approach, in which one considers the outcome of both methods and decides when to supercede the actuarial conclusion. In psychology this circumstance has come to be known as the 'broken leg'

> problem, on the basis of an illustration in which an actuarial formula is highly successful in predicting an individual's weekly attendance at a movie but should be discarded upon discovering that the subject is in a cast with a fractured femur (footnotes deleted). The clinician may beat the actuarial method if able to detect the rare fact and decide accordingly. In theory, actuarial methods can accommodate rare occurrences, but the practical obstacles are daunting. For example, the possible range of intervening events is infinite. (Dawes, Faust, and Meehl 1989, 1670)

The broken leg problem arises because a person who applies a reasoning strategy must judge whether it is appropriate to apply the strategy to this particular case. But there are bound to be difficult cases. The broken leg problem occurs when the person comes to believe she has strong evidence for defecting from the strategy.

4.1. Diagnosing the broken leg problem

The broken leg problem arises when a reasoning strategy that has been proven reliable on a particular class of problems is applied to a problem that is thought (rightly or wrongly) to be outside the range of problems for which the strategy is known to be reliable. For example, the VRAG (Violence Risk Appraisal Guide) test for violent recidivism was developed primarily as the result of research done on a population of violent Canadian psychiatric patients at the Oak Ridge Division of the Penetanguishene Mental Health Care Center (Quinsey et al. 1998, xi). When using the VRAG, one might reasonably wonder whether it is reliable on different subpopulations, such as non-psychiatric patients or criminals in the U.S. (In both cases, it is.) One way to pose the broken leg problem is to ask: Under what conditions is it reasonable to defect from a reasoning strategy that has been shown to be reliable for a particular class of problems?

The broken leg problem is a serious and pressing issue for any theory that embraces the findings of Ameliorative Psychology. On the one hand, it is absurd to suppose that one should never defect from a successful SPR. On the other hand, people have a hard time avoiding the temptations of defection. And excessive defection undermines reliability. After all, whenever an SPR is more reliable than human judgment and the expert and the SPR disagree, the SPR is more likely to be correct. In the long run, reliability is reduced if one insists upon consistently replacing more reliable reasoning strategies with less reliable reasoning strategies.

This intuitively powerful argument has been confirmed a number of times in the laboratory. There are a number of studies in which subjects

are given SPRs and then are permitted to selectively defect from them (i.e., override them), sometimes after having been told that the SPR by itself has been shown to be more reliable than experts. Typically, subjects find more broken leg examples than there really are. As a result, the experts predict less reliably than they would have if they'd just used the SPR (Goldberg 1968, Sawyer 1966, Leli and Filskov 1984). (Interestingly, it doesn't usually seem to matter whether the subjects are experts or not.) Selective defection strategies generally have a poor track record (except when the defectors have expertise in a theory with significant predictive success).

The broken leg problem and the failure of selective defection strategies suggest that any epistemic theory that hopes to take full advantage of the prescriptive power of Ameliorative Psychology must do more than put forward and recommend reliable SPRs. It must include a psychological theory of human judgment that can anticipate the difficulties we will have implementing the best available reasoning strategies. It is an unfortunate fact about humans that we are too often tempted to defect from successful SPRs. A normative theory with prescriptive force needs to predict the ways in which we are likely to deviate from excellent reasoning and perhaps provide methods of preventing such unfortunate deviations. Of course, we don't pretend to have such a theory; accordingly, our discussion of this matter will be tentative and programmatic. But we take this to be a prime example of how a reason-guiding epistemology will essentially depend on, and be informed by, a mature empirical psychology.

4.2. Grounded and ungrounded SPRs

Let's make a rough distinction between two classes of SPRs. Grounded SPRs are SPRs for which we have a theoretical explanation for their success. Ungrounded SPRs are SPRs for which we do not have a theoretical explanation for their success. Basically, we understand why grounded SPRs work, but we don't understand why ungrounded SPRs work. There are two points to note about this distinction. First, it is not hard-and-fast, since we can have better and worse understanding of why an SPR works. Second, for any ungrounded SPR, there may well be a neat causal explanation for its success that we don't yet know. So the distinction is not meant to be a metaphysical one, but an epistemological one. It is a distinction based on the quality of our understanding of SPRs and the subject matters on which they are based.

Consider an ungrounded SPR—the F minus F Rule for predicting marital happiness (discussed in section 1). Why is this rule reliable?

A reasonable assumption is that the correlation between the combined set of predictor cues and the target property is sustained by an underlying, stable network of causes. This is not to say that there is a science that would treat such ensembles of cues as a natural kind; it *is* to say, however, two things. First, their arrangement has a natural explanation. The explanation may not be unified—indeed, it may be so tortured that it is little more than a description of causal inventory—but it is an explanation in terms of causes nonetheless. Second, these arrangements, in general, do not spontaneously vanish.

Whatever specific facts explain the success of SPRs, they are not metaphysically exotic. As predictive instruments, SPRs are not like the occasional "technical" stock market indicators offered by gurus who combine a motley of moon phases, glottal stops, and transfer credits to predict stock movements. The VRAG test for predicting violent recidivism is an ungrounded SPR. In its present form, it consists of twelve predictor variables, and each is scored on a weighting system of (+) or (−). The weights vary from a −5 to a +12. The VRAG requires such information as the person's: Revised Psychopathy Checklist Score, Elementary School Maladjustment Score, satisfaction of any DSM criteria for a personality disorder, age at the time of the index offense, separation from either parent (except by death) by the age of sixteen, failure on prior conditional release, nonviolent offense history score (using the Cormier-Lang scale), unmarried status (or equivalent), meeting DSM criteria for schizophrenia, most serious victim injury (from the index offense), alcohol abuse score, and any female victim in the index offense (Quinsey et al. 1998). Many of these categories are independently known to interact richly with social behavior. It is not as though the diagnostic problem of deciding whether this person is likely to commit a similarly violent crime is being determined by facts known to be ontologically unrelated to or isolated from social behavior, such as the psychic's interpretation of tarot cards.

Now let's turn our attention to grounded SPRs. Many good examples of grounded SPRs come from medicine. In the case of determining the extent of prostate cancer, for example, there is a four-variable SPR that takes into account patient age, PSA (prostate specific antigen) test value, the biopsy Gleason score (arrived at from a pathologist's assessment of tissue samples), and the observable properties of the magnetic resonance image. Each variable makes an incremental improvement in determining the patient's prognosis. But we understand very well why three of those variables help to reliably predict the target property. We don't understand much about what mechanisms account for age being a good predictor.

Recall that we said that there was an exception to the general failure of strategies of selective defection. Grounded SPRs provide that exception. Experts can sometimes improve on the reliability of SPRs by adopting a strategy of selective defection (Swets, Dawes, and Monahan 2000). But notice that the improved reliability comes about because the expert can apply her well-supported theoretical knowledge to a problem. When someone is in possession of a theory that has proven to be reliable and that theory suggests defecting from an SPR (particularly when the expert's judgment relies on a cue not used by the SPR), then a strategy of selective defection can be an excellent one.

Even when an expert is able to outperform an SPR because of her superior theoretical knowledge, there are two notes of caution. First, there is every reason to believe that a new SPR can be developed that takes the expert's knowledge into account and that the refined SPR will be more reliable than the expert. One way to think about this is that when an expert is able to defeat the best available SPR, this situation is typically temporary: There is likely another SPR that can take into account the extra theoretical knowledge being employed by the expert and that is at least as reliable as the expert. The second note of caution is that even in domains with grounded SPRs, selective defection is not *always* a good strategy. The reasoner who has adopted the selective defection strategy needs to be able to apply the relevant theoretical understanding well enough to reliably defect from the SPR. And this will not always be easy to do. Even when the reasoner knows what variables to look at, he might still have a hard time weighing and integrating different lines of information (see section 3, above).

What about the (unfortunately) more common ungrounded SPRs, such as the Goldberg Rule, the VRAG, and the F minus F Rule? For most of the variables that make up these rules, there is no well-confirmed theory that explains their incremental validity, even if we feel we can tell a good story about why each variable contributes to the accuracy of prediction. Broken leg problems are particularly acute when it comes to ungrounded SPRs. Since we don't know why, specifically, the SPR is reliable, we are naturally diffident about applying the SPR to cases which seem to us to have some relevantly different property. For example, as we have noted, the VRAG was originally developed for violent Canadian psychiatric patients. But in order to prove its worth, it was tested on other populations and shown to be robust. A reasoning rule, particularly an ungrounded rule, that is not tested on a wide variety of different subpopulations is suspect.

Once we know that an ungrounded rule is robustly more reliable than unaided human judgment, the selective defection strategy is deeply suspect. As far as we know, VRAG has not been tested on violent criminals in India. So suppose we were asked to make judgments of violent recidivism for violent criminals in India, and suppose we didn't have the time or resources to test VRAG on the relevant population. Would it be reasonable to use VRAG in this situation? Let's be clear about what the issue is. The issue is *not* whether VRAG in the new setting is as reliable as VRAG in the original setting (where it has been tested and found successful). *The issue is whether VRAG in the new setting is better than our unaided human judgment in the new setting.* Let's consider this issue in a bit of detail.

When trying to make judgments about a new situation in which we aren't sure about the reliability of our reasoning strategies, we are clearly in a rather poor epistemic position. It is useful to keep in mind that this is not the sort of situation in which *any* strategy is likely to be particularly reliable. But our unaided human judgments often possess a characteristic that ungrounded SPRs don't—a deep confidence in their correctness. When we consider whether to employ an SPR (like VRAG) or our unaided human judgment to a new situation, it will often seem more reasonable to employ our judgment than the SPR. But notice, we typically don't know why either of them is as reliable as it is in the known cases. So we are not deciding on the basis of a well-grounded theory that the new situation has properties that make our judgment more reliable than the SPR. Instead, we're probably assuming that our reasoning faculties are capable of adapting to the new situation (whereas the SPR isn't), and so our faculties are likely to be more reliable. But on what grounds do we make such an assumption? After all, in a wide variety of situations analogous to the new one (recall, we're assuming the SPR is robustly more reliable than human experts), the SPR is more reliable than the expert. Why should we think that the expert is going to do better than the SPR in a quite defective epistemic situation? Perhaps neither of them will do any better than chance; but surely the best bet is that the strategy that has proven itself to be more reliable in analogous situations is going to be more reliable in the new situation.

Our tendency to defect from a lovely SPR is related to our tendency to plump for causal stories. Consider a disturbing example of a catchy story being accepted as causal fact. For too long, infantile autism was thought to be caused by maternal rejection. The evidence? Parents of autistic children could readily recall episodes in which they had not been accepting of their child (Dawes 2001, 136). It is easy to piece together a story about how

maternal rejection would lead to the characteristic social, emotional, and communication troubles associated with autism. But it is beyond appalling that such weak evidence could have been used to justify the view that mothers were causally responsible for their children's autism. As this case makes clear, stories are cheap. But even some of the most inaccurate stories are irresistible. When we tell a story, we begin to feel we understand. And when we think we understand, we begin to think we know when to defect from an SPR. Our unconstrained facility in generating stories, and our arrogance in accepting them, causes us to defect from far more accurate predictive rules. Consider another story. There are more ''muscle car'' purchases in the southeastern U.S. than in any other region. What explains this southeastern taste for Mustangs, Camaros, and Firebirds? Elements of an explanation immediately spring to mind. No doubt the Daytona and Winston-Salem stock car races influence local tastes. And (perhaps making a bit of a leap here), there's a good ol' boy hot-rod culture in the area—isn't there? As we fit these images into a more or less coherent assemblage, centered on a stereotype of rural poverty, poor education, and green bean casseroles, a gratifying sense of understanding washes over us. We become confident that we have hit upon an explanation. But as it turns out, the typical muscle-car purchaser also enjoys wok cooking and oat-bran cereal, uses fax machines, and buys flowers for special events (Weiss 1994, 62). Is the stereotype that motivates the story easily integrated with delectation of wok-prepared cuisine and floral sensibilities? It is hard to see how. Our ''explanation'' is really just a folksy story, creatively cobbled lore of familiar anecdotal cast. It is also dead wrong, and the sense of understanding it conveys, however comforting, is counterfeit. And yet it is hard to shake the story. Especially when it is fortified with apparently confirming evidence: The demographic map for muscle-car purchases looks very much like the demographic map for rates of response to junk mail. Those queried who aren't too shy sum it up very simply: It's what you'd expect from trailer trash (Weiss 1994).

As we have already admitted, sometimes reasoners should defect from SPRs, even ungrounded ones. One of our colleagues in psychology has developed an SPR for predicting recidivism for people convicted of child sexual abuse. When asked about the broken leg problem, the psychologist admitted that one should always correct the rule if it doesn't predict a zero chance of recidivism for dead people. There are very well-grounded causal hypotheses for why this sort of situation would call for defection. But in absence of a situation in which we have documented reasons (not merely easy causal stories) to believe that the ''broken leg'' property (e.g., death) is

a powerful predictor of the target property (e.g., crime), defection is usually a bad idea. The best advice is probably that one should typically resist defecting well beyond what intuitively seems reasonable. As Paul Meehl has said, we should defect from a well-tested SPR when the "situation is as clear as a broken leg; otherwise, very, *very* seldom" (1957, 273).

4.3. Three caveats on defection

In light of the documented failure of selective defection strategies, we have suggested that overriding an SPR is a good idea only in very unusual circumstances. But we offer three caveats. First, for particularly significant problems in a new domain, it will often make sense to test the SPR against expert prediction on the new cases before making judgments. There is an attitude (and often explicit prescriptions) of caution when applying instruments or techniques to new domains, particularly high-risk domains. This attitude is evident in gene therapy and cloning. But when it's not possible to carefully determine which tool is better on the new domain, a conservative attitude to defection is warranted—particularly for domains without grounded SPRs. As we've already argued, in those domains, defection to human judgment is generally unreliable.

Second, it is important to keep SPRs current—especially those that tend to handle especially significant problems. The parts of the natural and social world to which SPRs are applied are dynamic. If SPRs detect people's dispositions, then we should attend to any of the social or psychological trends that change people's relevant behavioral dispositions. Many of these conditions change over time: Crime initiatives in law enforcement, federal housing subsidies, emergency health care policies, and yes, even people's knowledge that statistical prediction rules, and more broadly actuarial methods, are being used to categorize them in various ways (see Hacking 1999). In order to ensure that the SPRs perform with optimal accuracy, SPRs must be regularly updated with fresh outcome information. In fact, it will often be more important to keep an SPR current than it will be to put effort into determining the conditions under which it is best to defect from it.

And third, after defecting from an SPR on the grounds of a broken leg problem, it is important to go back to the SPR next time (unless there is another such problem). Applying successful SPRs is an epistemically excellent tendency to cultivate. Defecting from an SPR frustrates and undermines the formation of such positive habits. If defecting from an SPR undermines our long-term commitment to using it, then defection is

a risky proposal, even when one is faced with a genuine broken leg problem. Ideally, we should take the proven exceptions and build them into a better SPR, if this can be done simply enough that people can use it.

5. Conclusion

Two central lessons of Ameliorative Psychology are that when it comes to social judgment, (a) proper unit weight models outperform humans in terms of reliability and (b) improper unit weight models (of which the Goldberg Rule and the F minus F rule are examples) often perform nearly as well as proper models and therefore better than humans. So why the resistance to these findings? We suspect that part of the reason people resist this "practical conclusion" is that the SPR results are noxious to our conception of ourselves as good reasoners. Further, they undermine our hope—so evident in the a priorism of so much contemporary epistemology—that we can be experts at recognizing good reasoning without massive empirical aid. (The SPR results do not, of course, suggest that we are naturally atrocious at recognizing good reasoning. It just suggests that we aren't *experts*; we aren't so good that we couldn't learn a lot from Ameliorative Psychology.) Once our dreams of native epistemological expertise are dashed, we can no longer take seriously the idea that we should attempt to build a theory of good reasoning without attending to empirical matters.

The fact that people are slaves to the temptation of broken legs suggests a deep problem with the methods of Standard Analytic Epistemology. SAE makes our considered epistemic judgments the final arbiters of matters epistemic. But it is precisely these epistemic judgments that so often fall to the temptation of broken legs. We have seen this countless times in discussions with philosophers. When confronted with 50-years worth of evidence suggesting that short, unstructured interviews are worse than useless, we are now accustomed to philosophers dismissing these findings ultimately because, well, they just don't fit in with their considered judgments. Now the defender of SAE might reply that there is no principled reason why SAE is committed to excessive defection—for the evidence here presented can now help to guide our judgment. Our reply is that, after 50 years, it hasn't. Avoiding defection isn't a matter of simply knowing the threat; it is a matter of avoiding it in the first place. And we can't avoid it if we have a philosophy that presses our faces into temptation's fleshy cargo.