1 When You Think It's Bad, It's Worse Than You Think: Psychological Bias and the Ethics of Negative Character Assessments

Hagop Sarkissian

We often find ourselves making judgments in the absence of complete information. This happens across a range of domains, including judgments about persons. Is this person praiseworthy or blameworthy? Honest or deceitful? Trustworthy or suspect? In such times of uncertainty, where evidence might be interpreted one way or another, one might consider giving the person in question the benefit of the doubt—that is, to suspend negative judgments for the time being and assume that, in the fullness of time and with increasing evidence and familiarity, the more favorable judgment will prove the correct one. The admonition to give others the benefit of the doubt is most straightforwardly understood as asking that we look upon others favorably and extend them a kindness in the face of countervailing considerations. But how far should one go in complying with it?

In a much-discussed article, Susan Wolf argues that extending to others the benefit of the doubt is an unwavering disposition of the moral saint, who must "be patient, considerate, even-tempered, hospitable, charitable in thought as well as in deed" and "must have and cultivate those qualities which are apt to allow him to treat others as justly and kindly as possible."[1] However, having stated these qualities, Wolf claims there is a substantial tension between being a moral saint and having certain other qualities of character that we would otherwise consider part of an enjoyable, well-lived life. For example, "a cynical or sarcastic wit … requires that one take an attitude of resignation and pessimism toward the flaws and vices to be found in the world," which is an attitude the moral saint cannot adopt.[2] Instead, the moral saint will do "whatever is morally necessary to secure good outcomes." Indeed, the moral saint "will be very reluctant to make negative judgments of other people," even in the face of countervailing considerations.[3] Insofar as cynicism or sarcasm would be inimical to this goal, a moral saint would shun such qualities of character.

> A moral saint … has reason to take an attitude in opposition to [cynicism and sarcasm]—he should try to look for the best in people, give them the benefit of the doubt as long as possible, try to improve regrettable situations as long as there is any hope of success … . A moral saint will have to be very, very nice.[4]

In sum, the moral saint is committed to doing "whatever is morally necessary" in order to make the world a better place. This will regularly require that the moral saint give others a pass, let bygones be bygones, and extend to them the benefit of the doubt—notwithstanding the fact that there might be apparent reason to judge them morally suspect.

The moral saint, so described, might seem naive or overly trusting. She might even seem to lack a basic sense of justice, oblivious to the fact that the world does contain bad actions and bad characters, that cynicism is warranted at times, and that sarcasm can be an appropriate reaction to social (even moral) transgressions. Refusing to make negative judgments of others and routinely giving them the benefit of the doubt for "as long as possible" might seem to many neither heroic nor laudable but instead silly or misguided, apt for misapplication or even exploitation. If this is what morality demands, one might reasonably conclude that it demands too much. Indeed, Wolf ends up concluding that such considerations speak to the importance of intuitions (as opposed to principles) in helping us strike a balance between the moral and the amoral in our personal value orientations.[5]

In this chapter, I will argue that one need not be committed to the saintly goal of securing favorable outcomes at any cost in order to have a standing commitment to give others the benefit of the doubt and refrain from negative character assessments. Moreover, one need not rely exclusively upon some form of intuitionism to guide one in this regard. Instead, there are compelling reasons to abide by these commitments once we focus on the nature of negative character judgments themselves, and the mechanisms that give rise to them. Are they reliable? What is their epistemic status? In what follows, I will outline a number of reasons that should weaken our confidence in our ability to accurately judge others and find them worthy of condemnation. And I believe that some of the most fruitful resources for thinking about these issues come from an entirely different moral tradition than the ones that Wolf assays, a tradition which has a distinct perspective on the prompts of human behavior (both good and bad) and, therefore, a distinct way of evaluating the nature of character judgments.

1 Reasons to Give the Benefit: The *Analects* of Confucius

My main resource for thinking about these issues is the *Analects*, a collection of conversations among Confucius (551?–479? BCE), his advisees, and his colleagues, collected over the decades (fifth to third centuries BCE) following Confucius' death. As with Wolf's discussion of moral saints, the *Analects* addresses such topics as whether one should be reluctant to make negative judgments of others, or whether one ought to give others the benefit of the doubt. However, it does so from a particular perspective—namely, one of a highly interconnected social world. According to this perspective, how any single person acts in any social occasion hinges greatly on

the behavior of the other individuals at hand.[6] Hence, whenever one wishes to explain or understand another's behavior—that is, whenever one were to judge it in some way—one would look beyond a person's motivations, goals, or traits of character. These would not, in the first instance, be the focus of judgment. Instead, one would examine a range of other considerations that, while external to the person's private mental life, would nonetheless be part of the context of the behavior in question and hence part of the explanatory account.

For example, one might consider who else was present at the time, what was said and in what tone of voice, how the individuals were related to one another, what roles they were occupying, and what expectations apply to them in these roles. The early Confucians (most especially Confucius and the third-century-BCE thinker Xunzi) viewed behavior as highly interconnected, prompted and shaped by one's social and environmental contexts in subtle yet sure fashion. It would seldom be appropriate to discount or overlook such factors in accounting for the person's behavior, as they might carry great explanatory weight. Imagine, for example, judging why a person is quiet at the dinner table. One might advert to various personality traits to explain this particular bit of behavior. For example, one might conclude that the person is shy, introverted, or diffident. Such traits might, indeed, be appropriate explanations for the token behavior in question. By contrast, one might advert to various aspects of the person's situational context. For example, one might note that the person is a junior member of the family, and that such members are not expected to lead discussion at the dinner table. One might also note that the person is in the presence of a teacher or mentor, and similarly advert to conventions about speaking improperly or out of turn.

Many passages of the *Analects* are centrally occupied with how one might be affecting or influencing the behavior in question oneself. A running theme throughout the text is cultivating one's own moral influence on others (one's *de* 德, or effective moral influence). In order to do so, one must mind one's comportment and monitor its effect on others. The following passage reflects this preoccupation:

> There are three things in our *dao* that a gentleman values most: by altering his own demeanor he avoids violence and arrogance; by rectifying his own countenance he welcomes trustworthiness; through his own words and tone of voice he avoids vulgarity and impropriety. (8.4)[7]

Here we see a direct connection between features of one's scrutable self and the behavior of others. The *Analects* maintains that the *junzi* 君子 (nobleman / exemplary person) can understand behavior in others by attending to aspects of his own comportment. Thus, when accounting for the behavior of others, one's own behavior would often be part of the explanatory context and hence part of the explanatory account. Searching for explanations of others' behavior without accounting for one's own influence on it would be incomplete at best, wholly misguided at worst.

A related point concerns how one ought to react when one comes across disagreeable personalities or behaviors. In such instances—when one has friction with others, or experiences frustration with them (or worse)—one is typically directed to look at oneself when trying to explain such troubles.

Master Zeng said, "Every day I examine myself on three counts: in my dealings with others, have I in any way failed to be dutiful? In my interactions with friends and associates, have I in any way failed to be trustworthy? Finally, have I in any way failed to repeatedly put into practice what I teach?" (1.4)

Master Zeng strives to become an exemplary moral person—a *junzi*, or person of noble bearing. When he inspects his behavior he does not simply compare it to certain prescribed rules of conduct. Rather, he attends to the way his own behavior may be affecting others. For Master Zeng, moral failure consists not in failing to mimic formal ritual ideals (an important motivation for early Confucian practitioners) but rather in failing to successfully influence his environments. Given these aims, focusing on others to explain why interactions with them are less than optimal would not only be inaccurate and incomplete; it would also be unproductive.

The Master said, "Do not be concerned that you lack an official position, but rather concern yourself with the means by which you might become established. Do not be concerned that no one has heard of you, but rather strive to become a person worthy of being known." (4.14; cf. 14.30)

The Master said, "When you see someone who is worthy, concentrate upon becoming his equal; when you see someone who is unworthy, use this as an opportunity to look within." (4.17)

The *junzi* is distressed by his own inability, rather than the failure of others to recognize him. (15.20)

Since individuals are loci of influence, affecting those with whom they interact, working on *oneself* is a way to influence how *others* behave.

Attacking your own bad qualities, not those of others—is this not the way to redress badness? (12.21)

There is a distinct pattern in these passages that concerns how the moral exemplar is supposed to react when dealing with recalcitrant, disagreeable, or otherwise bad individuals—that is, when one has reason and opportunity to make negative judgments of others. The pattern is one of caution and restraint.

Zigong was given to criticizing others. The Master said [sarcastically], "How worthy he is! As for myself, I hardly devote enough time to this." (14.29)

Zigong asked, "Does the *junzi* despise anyone?" The Master replied, "Yes. He despises those who pronounce the bad points of others." (17.24) [8]

Similar to the description of the moral saint above, here we also find that the *junzi* frowns upon voicing criticisms and making negative evaluations of others. Yet distinct

from the moral saint's aim of trying to maximize the chance of things going well, the reasons given here are largely epistemic, and have to do with whether such judgments are accurate or well supported. Consider the following commentary from the *Record of the Three Kingdoms* (third century CE) on this general theme in the text.

> Criticism and praise are the source of hatred and love, and the turning point of disaster and prosperity. Therefore the sage is very careful about them … . Even with the *de* of a sage, Confucius was reluctant to criticize others—how reluctant should someone of moderate *de* be to carelessly criticize and praise?[9]

Here we see a couple of reasons adduced for this reluctance to judge others (whether negative or positive). Part of the reason for caution stems from the possible fallout from such assessments. Insofar as one's own assessments might inform those of others, or shape those of others, one must be cautious in being loose with them. More important, perhaps, being quick to judge risks moral hubris. Focusing on others' bad qualities shields one from the more important task of self-scrutiny. Blaming others is easy; admitting one's own deficiencies is difficult. This finds poignant expression in a comment by Wu Kangzhai 吳康齋(1392–1469): "If I focus my attention on criticizing others, then my efforts with regard to examining myself will be lax. One cannot but be on guard against this fault!"[10] So long as one is preoccupied with pointing out the flaws in others, one avoids this more arduous task. It's as though we have natural tendencies that blind us to our own causal role in influencing unfavorable outcomes, and compel us to pin the blame on others and their shortcomings.

We find in these passages a commitment to one value of the moral saint—being very reluctant to pass negative judgments on others. Yet the reasons supporting this commitment go beyond that of trying as much as possible to make the world a better place. Apart from any such motivations, the *Analects* suggests that one take a broader perspective on the prompts of the behavior itself. It suggests that we shift our attention away from the person and instead look at the context of the behavior—including oneself insofar as one is part of that context. What is most remarkable about these injunctions to resist the impulse to blame others and instead look at oneself is that doing so goes against a well-documented tendency to do the contrary—a tendency that has been investigated for decades in experimental psychology.

2 Reasons for Giving the Benefit: Experimental Psychology

The actor/observer asymmetry has long enjoyed a status as one of the bedrock findings of social psychology, revealing something deep about our ways of explaining social behavior. As its name implies, the actor/observer asymmetry posits a difference between how actors explain their own behavior and how those observing them explain the very same behavior: actors invoke *situational* or *external* characteristics,

whereas observers invoke *personal* or *internal* characteristics. One way to capture the difference between these types of explanation is to consider the following sets of questions one might ask to explain the behavior in question[11]:

Personal or *internal* questions (asked when observers explain others' behavior): A person's personality, character, attitude, mood, style, intentions, thoughts, desires, and so on—how important were these in causing the behavior in question? To what extent can the behavior be attributed to the person's abilities, intentions, and effort?
Situational or *external* questions (asked when actors explain their own behavior): Situational context, the effect of other persons, the nature of the task at hand, the demands of one's position, environmental factors—how important were these in causing the behavior in question? To what extent can the behavior be attributed to situational variables or just dumb luck?

On the face of it, the asymmetry in explanation seems plausible. After all, why shouldn't we expect actors and observers to explain one another's behavior differently? Yet a meta-analysis by Bertram Malle has revealed very little support for the asymmetry as a *general* pattern of explanation.[12] Indeed, it emerges only when a handful of variables are in play. The strongest among these is when the behavior has a certain obvious *valence*—i.e., when it is seen as positive (successes; skilled activity; generosity; other socially desirable behaviors) or negative (failures; mishaps; aggression; other socially undesirable behaviors).[13] In short, we explain others' *negative* behavior as arising from *personal* or *internal* variables, and others' *positive* behavior as resulting from *situational* or *external* variables. However, when we explain our own behavior, this pattern is reversed (table 1.1).

The asymmetry is obviously self-serving: we disown our own failures, yet refuse to allow others to do so; we take credit for our successes, yet pin others' successes on things external to them. Could this, in fact, be an accurate assessment of what causes good and bad behavior? On the face of it, the self-serving nature of this tendency should provide us some *prima facie* reason to doubt the veracity of our explanations of others' negative behavior. But before jumping to this conclusion, it might help to try to understand why we have such a marked asymmetry in the first place. Here, I'll focus on two different accounts.

The first one explains this tendency as a by-product of evolutionary pressures faced by our ancestors.[14] Our ancestors struggled in competitive environments with limited

Table 1.1

	Positive behavior	Negative behavior
Actor's explanation	Personal or internal	Situational or external
Observer's explanation	Situational or external	Personal or internal

resources, and faced many threats and dangers; mortality rates were much higher than today, and life expectancy shorter. In such environments, where one is not guaranteed access to resources necessary for survival, reacting quickly and decisively at signs of perceived threat would be advantageous and fitness-enhancing. Potential threats to one's survival (such as competitive conspecifics who might endanger one's well-being) have an urgency and must be addressed immediately lest one bear terrible consequences. Keeping track of such individuals and adopting an intentional stance toward them (that is, attributing their threats as personal or internal) would be one way to ensure preparation against any potentially threatening behavior. Put another way, threats against the self signal that quick and decisive action is necessary, so evolutionary advantages would accrue to those who assumed that potential threats were stable and not ephemeral, and to those who acted quickly and automatically in response to them.

This explanation is, of course, highly speculative. Even so, let's grant for the sake of discussion that such an explanation is plausible. Can it justify the present asymmetry? Notice that for a behavioral tendency to be fitness enhancing it need not track truth. That is, it need not have been the case that all potential threats in our evolutionary past were intentional (or otherwise products of a person's motives or desires) in order for a tendency to consider them so to enhance fitness. False positives—treating threats as intentional when they weren't—might prove costly in terms of lost opportunities at forging cooperative relationships, but these would likely be outweighed by the costs of false negatives—where failing to react to a *true* threat might risk the very survival of the individual. Put another way, a tendency to attach negative behavior to the intentions of persons and then to track such persons over time would prove fitness-enhancing even if the tendency would routinely misfire. More important, perhaps, our present environments do not resemble those of our ancestors. We are no longer in a Hobbesian state in which personal security can be assured only through personal diligence. We have the entire apparatus of the modern state and its various policing institutions to help secure our persons and ensure an environment of predictable social interactions. This changes the cost-benefit structure of false positives as opposed to false negatives, making it unclear whether, strictly from a selfish perspective, pinning others' negative behavior to their character traits is beneficial.

A second type of explanation suggests that this tendency is the product of a naive theory of social behavior that individuals tacitly maintain—a theory that need not be influenced by evolutionary pressures.[15] According to this naive folk theory, individuals are continuously compelled by others to behave in socially desirable ways—that is, to be helpful, accommodating, and cooperative. These pressures are sufficient to explain why most individuals behave in ways that are generally helpful or benign; they do so owing to the continual demands of social existence. Why, then, do individuals act

Table 1.2

	Negative behavior	Positive behavior
Good people	Not capable	Capable
Bad people	Capable	Capable

contrary to accepted social norms and practices? The naive theory of social behavior maintains that it must be because either (a) they have a standing intention, desire, or motive to do so (indicating a person's bad character) or (b) they are constituted in such a way that they are incapable of acting otherwise. Both of these latter explanations refer to a person's character: if someone acts badly or poorly, it must be because of who he or she is. After all, since there are obvious costs that accrue to an individual for acting in a negative fashion, no individual would do so without intending to. As a consequence, negative behavior occurs less frequently but is intentional, and therefore more diagnostic.[16] Put another way, most individuals have good reason to act in socially desirable ways, and this is sufficient to explain why both good individuals and bad individuals will exhibit socially desirable behavior. By contrast, negative behavior can only stem from disreputable characters. Hence, inferences from good behavior to good character traits will be risky, and inferences from bad behavior to good character traits will be erroneous. Bad behavior comes from bad individuals (table 1.2).

This type of explanation might also enjoy an initial degree of plausibility. After all, the social pressures invoked seem real enough, and it seems reasonable to think that people would want to avoid the costs they would incur from contravening widely held norms. However, the explanation doesn't withstand critical scrutiny. For example, many instances of norm transgression or negative behavior might be accidental or unintended. In fact, if individuals have a standing motivation to act in socially approved ways, it seems just as likely to infer that any deviation must be the result of accident as opposed to intent. (Of course, being prone to accidents may reveal something about a person's character, but the asymmetry is not limited to instances of repeated observation.) More important, moral life is not free of conflict, and good people will often have to choose between several competing moral demands that cannot all be practically met. Failing to meet *all* of one's moral demands may lead to norm violations in some areas, yet it would be unfair to conclude that someone has a bad character or acted from bad motives simply because he cannot satisfy all of his demands.[17] It seems that one ought, instead, to take into account the situational constraints that may be impinging on the behavior in question.

Difficulties accounting for bad behavior are compounded when we judge unfamiliar persons—when we take their poor behavior on any particular occasion as indicative

of their character or their motivations generally—just because there is so little evidence to judge the person's character. People act differently when presented with different prompts and when placed within different contexts, and so drawing conclusions about someone's character in any situation on the basis of the previously observed behavior of others in similar situations is necessarily tenuous. Unfortunately, it is in our interactions with strangers that we are particularly vulnerable to making and maintaining such negative character evaluations, and injunctions to withhold judgment and give others the benefit of the doubt are particularly vital in such interactions.

When we meet others we form impressions of them, and those impressions tend to stick.[18] This tendency for first impressions to persevere motivates numerous social practices, such as grooming before a first date or rehearsing before an important presentation. It can be unfair, of course, to judge or evaluate persons on the basis of their behavior on any particular occasion, as the behavior may not be representative. Nonetheless, first impressions are easy to form and difficult to overcome.

Indeed, first impressions are remarkable predictors of the overall trajectory of interpersonal relationships. A study by Michael Sunnafrank and Artemio Ramirez suggests that we decide within minutes what sort of relationship we'll come to have with someone.[19] For the study, participants (college freshmen) were paired on the first day of class with another student, of the same sex, whom they didn't know. The participant was told to introduce himself or herself to the other individual and to talk to that person for either three, six, or ten minutes, then was asked to list the things the two individual had in common, to assess the overall quality of the interaction, and to estimate what sort of relationship was likely to develop: "nodding acquaintance," "casual acquaintance," "acquaintance," "close acquaintance," "friend," or "close friend." After nine weeks, the participants were contacted and asked to describe their current relationships with the partners. The best predictor of relationship status turned out to be how positive the initial interaction was (all things considered), which turned out to be far more important than common interests or likeability in predicting the relationship's trajectory. Although this highlights the importance of positive first impressions, it also gives us reason to discount negative ones, lest we close off the possibility of forging positive relationships in the future.

Alas, when it comes to initial impressions of others, we also find a pronounced asymmetry between negative impressions and positive impressions.[20] Negative impressions are taken as more diagnostic of individuals than positive information; negative behavior is taken as indicative of a person's character, whereas positive behavior is not.[21] Negative behavior is more easily remembered than positive behavior.[22] We remember negative behavior more accurately than positive behavior, and we are more confident about such memories.[23] Similarly, we take less time to arrive at negative judgments,[24] and we require considerably less evidence and information to ascribe

negative traits to individuals than to ascribe positive traits; in fact, the less favorable the character trait, the less evidence we need to believe in it.[25]

This asymmetry is likely related to the greater certainty we feel about our negative assessments of others relative to our positive assessments.[26] When we experience uncertainty we tend to be more systematic and careful when processing any relevant information we encounter; conversely, when we have a sense of certainty in our judgments we tend to process information in a more superficial fashion. Hence, those judgments that we tend to be certain about—namely, judgments resulting from negative behavior—tend to coincide with very shallow information processing, whereas those judgments we tend to be uncertain about—namely, judgments resulting from positive behavior—tend to coincide with more rigorous information processing. Finally, we tend not to look for alternative explanations of negative behavior once we have concluded that the behavior in question is representative of a bad character trait.[27]

In view of all the asymmetrical tendencies noted above, we have reason—above and beyond the moral saint's desire that everything go well—to doubt the veracity of our negative assessments of others—especially people not familiar to us. The suggestion here is not that we should always suspect them, or that all our assessments are equally susceptible to bias in every instance. Nonetheless, owing to the tendency of negative information to be weighed more heavily, to persevere longer, to be taken as more representative, and to be shielded from disconfirmation, we have good reason to doubt the negative assessments we make of others.

The doubt has two inter-related components. The first stems from the factors just mentioned—the asymmetrical weighting, perseverance, representativeness, and obstinacy of negative assessments versus positive ones. Since there seems to be no good reason to accept these effects as tracking truth, we should be willing to doubt them. Relatedly, a second reason stems from our systematic failure to search for other explanations of the token behavior in question—explanations that don't emphasize a person's character but instead look to situational, contextual, or accidental features. We may not be in error if we include character explanations in our understanding, yet we will often be in error if we take them to be exhaustive. We should try to expand our perspectives as observers and to explain the behavior of others as we would explain our own—that is, from the observed person's own perspective. In other words, when thinking badly of others, we should consider that we might be falling victim to a psychological tendency that prevents us from seeing their behavior in a more complete and accurate light. In doing so, we can leave open possibilities for constructive engagement and cooperation where they would otherwise be cut off.

3 Reasons to Give the Benefit: Game Theory

An analogue to the strategy of giving the benefit of the doubt in order to open up possibilities for engagement and cooperation can be found in game theory. In Robert Axelrod's famous tournament, players were pitted against one another in repeated encounters based on the classic Prisoners' Dilemma, in which each player has an opportunity to either defect or cooperate with the other. If the players cooperate, each receives a modest payoff; if both of them defect, neither receives a payoff; if one defects but the other cooperates, the defector gets an even greater payoff than he would have gotten if he had cooperated, while the cooperator is assessed a penalty. In view of these outcomes, it is rational to defect no matter what your opponent does: at best you get the highest reward, at worst nothing, whereas cooperating for a modest payout risks a considerable penalty. But if everyone always defects, no person receives any payoffs. That's the dilemma.

In Axelrod's tournament, a very simple strategy called Tit for Tat emerged victorious in the face of far more sophisticated strategies. The Tit for Tat strategy had only two rules:

1. When you first meet another player, cooperate.
2. Thereafter, choose the response that the other player chose when last encountered.

This strategy proved remarkably effective, besting several more complicated strategies. However, it had a significant flaw; it had no tolerance for noise or error. When noise is added to an iterated Prisoners' Dilemma tournament in the form of errors or misunderstanding, Tit for Tat strategies can become trapped in a long string of retaliatory defections, thereby depressing their score. Such noise may come in one of two forms: a co-player might either send the wrong signal (also known as misimplementation or "trembling hand") or might send the right signal yet be misinterpreted by other players (misperception or "noisy channels"). "Faulty transmission of strategy choices (noise) severely undercuts the effectiveness of reciprocating strategies"[28] such as Tit for Tat. In one and the same tournament, Tit for Tat can go from the winning strategy to sixth place if strategies are randomly set to misfire 10 percent of the time (i.e., defecting where one would otherwise cooperate or vice versa).[29] The reason is easy enough to grasp: Upon encountering a defector, a Tit for Tat strategy will reciprocate with defection, which will result in an extended series of mutual obstructions (that is, both players defect). In such situations, Tit for Tat strategies must rely on the co-player to initiate a cooperative move; without such initiative from the co-player, the Tit for Tat player will continue to defect indefinitely, even if the original defection of the co-player was an unintended result of noise. Indeed, Tit for Tat is particularly

vulnerable against *itself* in noisy environments—a single miscue can result in an extended mutual obstruction, and only another miscue will be capable of triggering a new series of cooperation.

The obvious way to break the vicious cycle of retaliation is to requite defection with cooperation from time to time—to let bygones be bygones. Such strategies are generally known as Generous Tit for Tat. Consider, for example, a variant called Tit for Two Tats, or Forgiving Tit for Tat, which will wait for two defections in a row before retaliating with defection. In a mixed environment in which there are many strategies at play, Tit for Two Tats works just as well as Tit for Tat, and in some instances outperforms it. Indeed, in biologically relevant evolutionary games interactions can be twisted away from defection and toward cooperation by the introduction of such strategies, which are more tolerant of noise. Adding Generous Tit for Tat "greatly increases the overall level of cooperation and can lead to prolonged periods of steady cooperation."[30]

Our own social environments most resemble "noisy" games. It is not uncommon to misinterpret others' signals or to fail to convey our own intentions clearly. Wires get crossed, identities are mistaken, and unwarranted assumptions are made. Sadly, such miscues are often taken to be highly diagnostic of character and purpose, weighted accordingly, and thus reciprocated by real-life "defection"—our tendency to have negative impressions harden into obstinate beliefs. The social/moral game (as it were) can be decided quickly and ruthlessly. Opportunities for negotiation and moral advancement can be nipped in the bud as a result of bad first moves. Yet if negative assessments should no more admit to personal or internal explanations than to positive ones, it is important to foster this habit of giving others the benefit of the doubt and allowing fruitful, constructive, and productive relationships to unfold.

4 Reasons to Withhold the Benefit: *Analects*

In the theory of games, as in real life, being forgiving has benefits and costs. For example, Tit for Two Tats is at a disadvantage when faced with very aggressive strategies, which exploit them not once but twice before being punished in turn. In environments where individuals routinely exploit forgiving natures, it would be imprudent to forgive others' transgressions. This underscores the importance of both giving the benefit of the doubt *and* drawing accurate assessments of others—even if they are unfavorable. Without the latter virtue, one can be exposed to moral vulnerability.

If I am correct in claiming that giving others the benefit of the doubt is an important theme in the *Analects*—that it contains injunctions to look beyond internal or personal characteristics when explaining behavior, that it enjoins us to see others as like ourselves—then it would be putting its adherents at risk of being exploited by

morally unscrupulous individuals. Indeed, this issue is broached in a number of places in the text.

Zai Wo asked, "If someone were to lie to a *ren* 仁 [humane] person, saying "A man has just fallen into a well!"—would he go ahead and jump in after him [to try and save him]? The Master said, "Why would he do that? The *junzi* can be enticed but not trapped; he can be tricked but not duped." (6.26)

The Master said, "Is a man not superior who, without anticipating attempts at deception or presuming acts of bad faith, is nonetheless the first to perceive them?" (14.31)

Edward Slingerland's selection of commentary on this passage merits lengthy citation:

The gentleman is trusting of others, and expects the best of them. As *Dai's Record* says, "The gentleman does not anticipate badness from others, nor does he suspect others of untrustworthiness." Li Chong sees this open attitude as the key to the Gentleman's ability to educate others: "If you perceive an act of untrustworthiness in the beginning and then necessarily expect untrustworthiness in the future, this indicates an impairment of the merit of patient forbearance, and also blocks the road to repentance and change." Nonetheless, the gentleman is not a fool, and is the first to perceive when his trust has been misplaced.[31]

Being capable of properly judging others and of despising them when that is necessary would be important for anyone for whom being cooperative, deferential, mindful, and conscientious are important commitments. Those pursuing the Confucian *dao* 道 would be prone to exploitation when surrounded by individuals seeking power, position, fame, and wealth, as was the case in Warring States China (475–221 BCE). In such environments, it would be imperative to identify those truly worthy of hatred (even while erring on the side of false negatives). Indeed, as E. Bruce Brooks and A. Taeko Brooks note, despising (*wu* 惡) is a classic virtue, appearing in the earliest stratum of the *Analects*.[32]

Only the *ren* can truly love others, and truly despise them. (4.3)

The Master said, "I have not seen a person who loved *ren* or despised what was not *ren*. He who loved *ren* would esteem nothing above it. He who hated what is not *ren* would be *ren* himself, since he would not allow anything that is not *ren* to be associated with his person." (4.6)

Nonetheless, it remains true that the text recommends a general attitude of favorableness toward others. After all, if expecting the worst from others can make them act poorly,[33] then expecting well from them, thinking favorably of them, might do the opposite.

The Master said, "The *junzi* helps others fulfill their attractive qualities rather than their unappealing ones. The petty person does the opposite." (12.16)

Admittedly, it seems difficult to figure out just how the *junzi* will balance the injunction to be favorable to others and give them the benefit of the doubt with the equally important injunction to properly judge some of them as being morally despicable. Yet no matter how the *junzi* might balance these injunctions, we should keep in mind that fighting the tendency to blame or resent others is a losing proposition unless the person's behavior changes within a reasonable length of time. In other words, giving others the benefit of the doubt is a strategy with a limited shelf life; the cognitively demanding act of staving off blame and resentment can be expected to last only so long. The injunction to give others the benefit of the doubt is, after all, a strategy to redress a standing psychological bias, and will prove effective only when others provide evidence of the transitory or contingent nature of their initial disagreeable behavior.

We find much of this summarized in a noteworthy passage in the writings of Mencius, a Confucian thinker of the fourth century BCE:

> Suppose someone were to be harsh in their treatment of me. A *junzi* would, in such a case, invariably examine himself, thinking "I wasn't benevolent; I lacked propriety. How else could such a thing have come about?" But if, after examining himself, he discovers he had been benevolent, he had acted with propriety, and yet the person *still* treats him harshly, then the *junzi* will again invariably examine himself, thinking "I must have lacked commitment." But if he discovers that he was, in fact, committed, and the person *still* treats him harshly, only then would the *junzi* say, "I suppose he is the incorrigible one."[34]

Here we find the epistemic considerations adduced above expressed most directly. The *junzi* has encountered disagreeable conduct directed toward him. His first impulse is to see how he might have engendered the conduct himself: Was he indiscrete or unkind? Did he lack patience or resolve? Here he is merely trying to come to a proper or complete understanding of what may have caused the person to act in such a fashion. Only after arriving at a more definite understanding of the prompts of the behavior—after concluding that it is unlikely to have been the result of some contingent prompt—is the *junzi* satisfied with blaming the person. (Here we find an analogue to the Tit for Two Tats strategy: Pause not once but twice before retaliating with a defection—in this case, with a negative character assessment.)

Conclusion

At the outset of the chapter, I noted Susan Wolf's argument that morality can demand too much, and that there may be personal, amoral ideals of character that have valid claims among our personal aspirations. On this view, moral perfection cannot be the sole or primary ideal that structures our lives; we have reason to aspire to ideals that are amoral. For Wolf, this means that "we have reason to want people to live lives

that are not morally perfect," and that "any plausible moral theory must make use of some conception of supererogation" to mark off moral demands that are optional and discretionary from those that are obligatory.[35] Presumably, this would include marking off the moral demand to give others the benefit of the doubt. Although this demand can be motivated by a number of considerations, none of them point to its being a strict duty. Instead, one will need to consider the particular contexts of any token negative judgment to determine whether it is, all things considered, something one ought to give credence to or something one ought to doubt. There is room for discretion here, so one can take on board Wolf's suggestion that intuitions will be necessary to the process.

Nonetheless, and in contrast with other supererogatory acts, I have argued that giving others the benefit of the doubt is motivated by strong epistemic reasons, and that we should question the veracity of our negative assessments of others—especially when we are unfamiliar with the individuals involved. Hence, though this particular virtue may not be obligatory, it warrants standing concern beyond any desire to maximize the chances that things will go well. Instead, it can be motivated by a desire to treat others fairly, to be accurate in one's assessments, and to avoid the costs associated with closing others off because of cognitive processes that are likely to be biased or erroneous. And although experimental psychology provides evidence as to the biased nature of this particular range of judgments, it remains a discipline that trades largely in descriptive facts as opposed to prescriptive norms. For the latter, it is fruitful to look to a tradition—Confucianism—that has, from its outset, taken a perspective on social life that recognizes the precarious nature of drawing such judgments, and which has rich normative resources structured around this perspective.

Giving others the benefit of the doubt may not be easy. On any realistic assessment of moral life, we must admit that, as we navigate the social world, there will be endless opportunities for friction with others to arise. Even if one is conscientious about one's own behavior and mindful of being respectful of others, these will never safeguard one from finding others disagreeable or difficult. Moreover, doubting such judgments can require going against what others have said about an individual and flagging the information as tentative and needing confirmation, as the *Analects* is well aware.

> The Master said, "It doesn't matter if the multitude hates someone; you must still examine the person and judge for yourself. It doesn't matter if the multitude loves someone; you must still examine the person and judge for yourself." (15.28)

At other times, it will require overcoming first-person observations and evidence. Yet adopting such a stance may be a winning strategy both in the theory of games and in the game of life. And though serious moral tolerance and accommodation may not always be in the offing, and though certain individuals may not seem to warrant the

benefit of the doubt, a disposition to giving one can be propitious to (and sometimes necessary for) accommodation and cooperation to emerge as live options.

Acknowledgments

My thanks to Brian Bruya for helpful suggestions on previous drafts, and to Owen Flanagan and David Wong for helpful discussions.

Notes

1. Susan Wolf, "Moral Saints," 421.

2. Ibid., 422.

3. Ibid., 421.

4. Ibid., 422.

5. Ibid., 439.

6. For an argument justifying an interconnected perspective in current moral psychology, see Hagop Sarkissian, "Minor Tweaks, Major Payoffs." For an important elucidation of this theme in early Confucian thought more generally, incorporating insights from newly unearthed texts, see Mark Csikszentmihalyi, *Material Virtue*, especially 178–192.

7. All translations are my own, using the Chinese text in D. C. Lau, *Confucius: The Analects*.

8. The passage continues: "He despises those who remain below while criticizing those above; he despises those who are bold but lack courtesy, daring yet violent."

9. Edward G. Slingerland, *Confucius Analects*, 166.

10. Ibid.

11. Adapted from Bertram F. Malle, "The Actor-Observer Asymmetry in Attribution," 896.

12. Ibid.

13. In much of what follows, I focus on the moral dimensions.

14. See, for example, Felicia Pratto and Oliver P. John, "Automatic Vigilance."

15. Oscar Ybarra, "Naive Causal Understanding of Valenced Behaviors and Its Implications for Social Information Processing."

16. See, for example, Susan T. Fiske, "Attention and Weight in Person Perception."

17. For further discussion of how competing moral demands, as well as the agent's attitude toward any transgression she may have to commit in choosing between such demands, are taken into account in our folk psychology, see Mark Phelan and Hagop Sarkissian, "Is the 'Trade-Off Hypothesis' Worth Trading For?"

18. This section shares parallels with portions of Sarkissian, "Minor Tweaks, Major Payoffs."

19. Michael Sunnafrank and Artemio Ramirez Jr., "At First Sight."

20. Fiske, "Attention and Weight in Person Perception."

21. Glenn D. Reeder and Marilynn B. Brewer, "A Schematic Model of Dispositional Attribution in Interpersonal Perception."

22. Jeffrey W. Sherman and Leigh A. Frost, "On the Encoding of Stereotype-Relevant Information under Cognitive Load."

23. Donal E. Carlston, "The Recall and Use of Traits and Events in Social Inference Processes."

24. John H. Lingle and Thomas. M. Ostrom, "Retrieval Selectivity in Memory-Based Impression Judgments."

25. Myron Rothbart and Bernadette Park, "On the Confirmability and Disconfirmability of Trait Concepts."

26. Carlston, "The Recall and Use of Traits and Events in Social Inference Processes"; Vincent Y. Yzerbyt and Jacques-Philippe Leyens, "Requesting Information to Form an Impression."

27. Ybarra, "When First Impressions Don't Last."

28. Robert Axelrod and Douglas Dion, "The Further Evolution of Cooperation."

29. Christian Donninger, "Is It Always Efficient to Be Nice?"

30. Martin Nowak and Karl Sigmund, "Chaos and the Evolution of Cooperation."

31. Slingerland, *Confucius Analects*, 166.

32. E. Bruce Brooks and A. Taeko Brooks, *The Original Analects*.

33. See, for example, Mark Chen and John A. Bargh, "Nonconscious Behavioral Confirmation Processes."

34. *Mencius* 4B:28.

35. Wolf, "Moral Saints," 438.

Works Cited

Axelrod, Robert, and Douglas Dion. "The Further Evolution of Cooperation." *Science* 242, no. 4884 (1988): 1385–1390.

Brooks, E. Bruce, and A. Taeko Brooks. *The Original Analects: Sayings of Confucius and His Successors*. Columbia University Press, 1998.

Carlston, Donal E. "The Recall and Use of Traits and Events in Social Inference Processes." *Journal of Experimental Social Psychology* 16, no. 4 (1980): 303–328.

Chen, Mark, and John A. Bargh. "Nonconscious Behavioral Confirmation Processes: The Self-Fulfilling Consequences of Automatic Stereotype Activation." *Journal of Experimental Social Psychology* 33, no. 5 (1997): 541–560.

Csikszentmihalyi, Mark. *Material Virtue: Ethics and the Body in Early China*. Brill, 2004.

Donninger, Christian. "Is It Always Efficient to Be Nice? A Computer Simulation of Axelrod's Computer Tournament." In *Paradoxical Effects of Social Behavior: Essays in Honor of Anatol Rapoport*, ed. Andreas Diekmann and Peter Mitter. Physica-Verlag, 1986.

Fiske, Susan T. "Attention and Weight in Person Perception: The Impact of Negative and Extreme Behavior." *Journal of Personality and Social Psychology* 38, no. 6 (1980): 889–906.

Guglielmo, Steve, and Bertram F. Malle. "Can Unintended Side Effects Be Intentional? Resolving a Controversy over Intentionality and Morality." *Personality and Social Psychology Bulletin* 36, no. 12 (2008): 1635–1647.

Lau, D. C. *Confucius: The Analects*. Chinese University Press, 1992.

Lingle, John H., and Thomas. M. Ostrom. "Retrieval Selectivity in Memory-Based Impression Judgments." *Journal of Personality and Social Psychology* 37 (2) (1979): 180–194.

Malle, Bertram F. "The Actor-Observer Asymmetry in Attribution: A (Surprising) Meta-Analysis." *Psychological Bulletin* 132, no. 6 (2006): 895–919.

Nowak, Martin, and Karl Sigmund. "Chaos and the Evolution of Cooperation." *Proceedings of the National Academy of Sciences* 90, no. 11 (1993): 5091–5094.

Phelan, Mark, and Hagop Sarkissian. "Is the 'Trade-Off Hypothesis' Worth Trading For?" *Mind and Language* 24, no. 2 (2009): 164–180.

Pratto, Felicia, and Oliver P. John. "Automatic Vigilance: The Attention-Grabbing Power of Negative Social Information." *Journal of Personality and Social Psychology* 61, no. 3 (1991): 380–391.

Reeder, Glenn D., and Marilynn B. Brewer. "A Schematic Model of Dispositional Attribution in Interpersonal Perception." *Psychological Review* 86, no. 1 (1979): 61–79.

Rothbart, Myron, and Bernadette Park. "On the Confirmability and Disconfirmability of Trait Concepts." *Journal of Personality and Social Psychology* 50, no. 1 (1986): 131–142.

Sarkissian, Hagop. "Minor Tweaks, Major Payoffs: The Problems and Promise of Situationism in Moral Philosophy." *Philosophers' Imprint* 10, no. 9 (2010): 1–15.

Sherman, Jeffrey W., and Leigh A. Frost. "On the Encoding of Stereotype-Relevant Information under Cognitive Load." *Personality and Social Psychology Bulletin* 26, no. 1 (2000): 26–34.

Slingerland, Edward G. *Confucius Analects: With Selections from Traditional Commentaries*. Hackett, 2003.

Sunnafrank, Michael, and Artemio Ramirez Jr. "At First Sight: Persistent Relational Effects of Get-Acquainted Conversations." *Journal of Social and Personal Relationships* 21, no. 3 (2004): 361–379.

Wolf, Susan. "Moral Saints." *Journal of Philosophy* 79, no. 8 (1982): 419–439.

Ybarra, Oscar. "Naive Causal Understanding of Valenced Behaviors and Its Implications for Social Information Processing." *Psychological Bulletin* 128, no. 3 (2002): 421–441.

Ybarra, Oscar. "When First Impressions Don't Last: The Role of Isolation and Adaptation Processes in the Revision of Evaluative Impressions." *Social Cognition* 19, no. 5 (2001): 491–520.

Yzerbyt, Vincent Y., and Jacques-Philippe Leyens. "Requesting Information to Form an Impression: The Influence of Valence and Confirmatory Status." *Journal of Experimental Social Psychology* 27, no. 4 (1991): 337–356.