

Decision

Reconsidering the “Bias” in “The Correspondence Bias”

Drew Walker, Kevin A. Smith, and Edward Vul

Online First Publication, April 4, 2022. <http://dx.doi.org/10.1037/dec0000180>

CITATION

Walker, D., Smith, K. A., & Vul, E. (2022, April 4). Reconsidering the “Bias” in “The Correspondence Bias”. *Decision*. Advance online publication. <http://dx.doi.org/10.1037/dec0000180>

Reconsidering the “Bias” in “The Correspondence Bias”

Drew Walker¹, Kevin A. Smith², and Edward Vul³

¹ Department of Cognitive Science, University of California at San Diego

² Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

³ Department of Psychology, University of California at San Diego

We do not directly observe the internal qualities of others so we must infer them from behavior. Although classic attribution theories agree that we consider situational pressures when estimating such internal qualities, one of the best-known results in psychology is that we are prone to a *correspondence bias*: That we draw inferences from behavior, even when we know that the situation has constrained the action. Dozens of theoretical accounts have sought to explain this result, with the most famous being the proposal that we commit a *fundamental attribution error*: We are systematically biased to underappreciate the influence of external factors and thus overattribute behavior to disposition. Although there remains disagreement about why we attribute constrained behavior to disposition, most researchers agree that this tendency is in fact an *error*. We propose that the social judgments made in classic attitude attribution studies have been widely interpreted as reasoning errors only because they have been compared to an inappropriate benchmark, predicated on the assumption of deterministic dispositions and situations. Building from earlier probabilistic accounts, we review classic results that demonstrate that social inferences are consistent with unbiased probabilistic attribution of the influence of situations and dispositions in an uncertain world.


Keywords: correspondence bias, fundamental attribution error, social inference, probabilistic inference

A museum patron drops \$5 into the “pay what you can” donation jar; how generous is she? What if a stern museum docent was monitoring the

donation jar? In general, we cannot directly see internal qualities like “generosity” and so they must be inferred from behavior, but behavior is also influenced by external circumstances, such as a watchful docent. Thus, attributing behavior to internal qualities or external situations is an underdetermined problem.

Since the 1960s, an extensive literature has argued that there are systematic discrepancies between the inferences about internal qualities that people *should* make and the inferences they *do* make. Specifically, it has been argued that we tend to blame a person’s behavior on internal qualities and neglect the influence of outside pressures. That is, when we witness someone make a donation, we are prone to think that she is generous, *not* to conclude that the watchful docent is imposing pressure to donate. In a classic demonstration of this phenomenon, Jones and Harris (1967) asked university students to read an essay, ostensibly written by a classmate, which either favored or opposed Fidel Castro. Even

Drew Walker  <https://orcid.org/0000-0001-7565-7355>

Kevin A. Smith  <https://orcid.org/0000-0001-5009-0460>

Drew Walker played a lead role in data curation, investigation, project administration, writing of original draft, and writing of review and editing and an equal role in conceptualization, formal analysis, methodology, and visualization. Kevin A. Smith played a supporting role in conceptualization, formal analysis, methodology, and writing of review and editing. Edward Vul played a supporting role in methodology and writing of review and editing and an equal role in conceptualization, formal analysis, and visualization.

The analyses presented in this article were not preregistered. The materials have not been made available on a permanent third-party archive; requests for the material can be sent via email to the lead author at dehoffma@ucsd.edu.

Correspondence concerning this article should be addressed to Drew Walker, Department of Cognitive Science, University of California at San Diego, 9500 Gilman Drive La Jolla, CA 92093, United States. Email: dehoffma@ucsd.edu

when told that the author was assigned their position by an instructor, readers still thought that the author actually held the view expressed in the essay. Jones and Harris concluded that this behavior deviated from a “rational analysis”: We ought to believe that behavior corresponds to attitudes only when authors were free to choose, but when the instructions compelled the author’s action, observers ought not infer internal qualities based on behavior. The core result has been replicated in numerous experiments using a variety of social situations (e.g., Fein et al., 1990; Gilbert & Jones, 1986; Gilbert & Osborne, 1989; Gilbert et al., 1988, 1992; Johnson et al., 1984; Jones et al., 1971; Miller et al., 1981, 1990; Reeder & Spores, 1983; Reeder et al., 1989, 2004; Snyder & Frankel, 1976). This tendency to blame observed behavior on actors’ dispositions has become known as the *correspondence bias (CB)* and has been called “a candidate for the most robust and repeatable finding in social psychology” (Jones, 1990).

Theoretical Accounts of the CB

Many theoretical accounts of the CB have been proposed. Jones (1979) attributed it to anchoring and adjustment (Tversky & Kahneman, 1974)—the behavior provides the anchor on which observers judge disposition, and they insufficiently adjust for situation pressure. Other variations of this account propose that trait attribution is spontaneous, while incorporating situational influences to correct these judgments is effortful and often not performed (e.g., Trope, 1986; Trope & Gaunt, 1999; Uleman, 1999). Gilbert and Malone (1995) proposed additional, complementary factors, wherein observers lack situational awareness, have unrealistic expectations of behavior, or misinterpret observed behavior. Gawronski (2004) argued that the failure to adjust dispositional attribution based on situation arises from systematic misapplication of calibrated causal theories.

In general, the presence of a “bias” does not necessarily reflect illogical thinking. For example, if you are waiting for an important phone call, shifting your detection criteria to overrespond to the sensation of a vibration in your pocket will lead to many false alarms, but this systematic tendency to overcheck is logical given your goal. Likewise, if it is more important to draw causal inferences about behavior, this systematic tendency would not necessarily be illogical. In line

with this, others have proposed that CB is beneficial and adaptive (Andrews, 2001; Vonk, 1999).

Still others have rejected the situation–disposition distinction altogether: For instance, Sabini et al. (2001) argue that situational pressure can often be reframed as disposition influence (e.g., in Jones & Harris, 1967, one could ask: Which disposition did the situation activate, the need to please the experimenter, or the need to express their belief about Castro?). Likewise, Malle (1999) argues that observers rely on actors’ intent to make social judgments, and so using the traditional situation–disposition distinction to understand attitude attribution is a misguided approach.

These accounts of the CB notwithstanding, the most well-known account for why we draw correspondent inferences from constrained behavior is that we are prone to a *fundamental attribution error (FAE)*, which is the idea that observers have incorrect causal theories about behavior; we think situation exerts a weak causal influence, but disposition exerts a strong causal influence (Ross, 1977). Over the decades, Ross and other theorists have maintained this view that we are “lay dispositionalists” (Ross & Nisbett, 2011), wired to neglect situational pressure and overweight the influence of stable internal qualities.

Deterministic Assumptions About Normative Behavior

There are many accounts competing to explain why people make dispositional inferences when “a logical analysis suggests they should not” (Gilbert & Malone, 1995). Here, we argue that the error lies in the “logical analysis” to which people are compared. The normative account on which the CB is based presumes that situational pressures determine human behavior, much like a power-outage guarantees that a light bulb will not turn on. With such deterministic influences in place, attribution to other possible causes is indeed unwarranted: If a light bulb fails to turn on during a power outage, you ought not conclude that the bulb is bad. Likewise, a pro-Castro essay written at gunpoint in a Cuban prison ought not indicate that the author is a Castro supporter. Such assumptions are consistent with presuming a “multiple sufficient causes” schema (Kelley, 1973), under which knowing the presence of one cause entails complete discounting of all other causes. Although Kelley (1973) proposed that “the role of a given cause in producing a given effect is discounted if

other plausible causes are present” he was clear that complete discounting ought not be expected under other plausible causal schemas (such as “multiple *necessary* causes”), despite this, much of the social attribution literature has presumed that full discounting is normative (McClure, 1998), yielding an implicit assumption that situations are either uninfluential, or sufficient, deterministic causes of behavior.

In reality, we rarely encounter situations that are deterministic. Even when society takes great care to make behavior as constrained by the situation as possible (e.g., locking someone in jail), these situations are still not *completely* deterministic (people still escape from jail). Situations, instead, interact with internal qualities to produce behavior: Some people would not make an optional donation even when a docent was watching, while others would donate even without a witness. Likewise, situations like those used in attitude attribution tasks are also far from deterministic, for example, when Sherman (1980) instructed university students to write an essay supporting a controversial school policy, less than 70% of students complied. Thus, even when a situation is presumed to be influential, it is not reasonable for people to assume that it will completely determine behavior.

Probabilistic Assumptions About Normative Behavior

A less popular, but more realistic, assumption about the nature of situational causes would be that they are *probabilistic* rather than *deterministic*. Indeed, many theorists have formalized how we might assume probabilistic influences produce outcomes, and how we might use Bayesian inference in causal attribution (e.g., Fernbach & Erb, 2013; Liefgreen et al., 2018; Pearl, 1988). Using this approach to describe social attribution was first proposed by Ajzen and Fishbein (1975) and later by Morris and Larrick (1995).

Morris and Larrick (1995) show that consistent with Kelley’s view, if we witness an action, and then learn about the presence of a situation that was sufficient to cause the outcome (e.g., a deterministic event; one that would by itself cause the action 100% of the time), it is indeed inappropriate to use that action as evidence for an additional potential cause, such as the person’s belief or disposition. But critically, they also modify Kelley’s (1972) causal schema

framework to accommodate situations that are not sufficient (not deterministic; would cause the action less than 100% of the time) and present a compelling mathematical argument that it is consistent with optimal probabilistic reasoning for observers to estimate a higher probability that the actor possesses an action-consistent disposition in these cases. Further, they replicate the original Castro study but additionally collecting participants’ subjective probabilities about events (e.g., the probability of writing an essay if one were pro-Castro but not instructed, or if one were anti-Castro but forced) and demonstrate that participants’ dispositional attributions are internally consistent given their assumptions about the sufficiency of instructions to compel a pro-Castro essay. In other words, in cases where the subject believes the situation is not entirely constraining, the CB pattern of results could be consistent with optimal probabilistic inference.

Despite these results, the probabilistic explanation of social attribution continues to be overshadowed in popular literature (although some social cognition theorists have adopted the probabilistic perspective; Hilton, 2017). The prevailing attitude, summarized by Langdrige and Butt (2004), is that “there is no unifying theory to account for the extensive catalog of experimental work [on situational discounting]” and the FAE and CB continue to be cited to contextualize a range of human behaviors, from social and political events (Haney & Zimbardo, 2009) to lie detection (O’Sullivan, 2003) to perspective taking (Hooper et al., 2015). Indeed, despite being published the same year as Morris and Larrick (1995), Gilbert and Malone’s (1995) review that asserts attributions violate “logical analysis” has been cited eight times as often.

In this article, we hope to amplify the probabilistic explanation of social attributions and expand on Morris and Larrick’s (1995) proposal in two key ways that make this explanation more general. First, Morris and Larrick treated dispositions, situations, and actions as dichotomous: People are pro-Castro or not, instructors assign a position or do not, and an essay is pro-Castro or anti-Castro. But this is not how the world works in general, as there can be full-throated and waffling supporters, assignments can count for half the course grade or be worth paltry extra credit, and essays can weakly or strongly support Castro (Jones et al., 1971). We propose a model based

on similar logic to Morris and Larrick's but that can account for these graded attributes.

Second, probabilistic proposals have previously focused on the CB and demonstrated that not fully discounting situational factors can be normative (e.g., Ajzen & Fishbein, 1975; Ajzen, 1977; Ajzen et al., 1979; Morris & Larrick, 1995). However, Quattrone (1982) found a curious mirror to the CB, in which people seemingly fail to discount what they know about another person's disposition and overattribute behavior to the situation. This behavior is fundamentally inconsistent with the explanation that we are "lay dispositionalists" (Ross & Nisbett, 2011) and has required theorists to propose an additional set of heuristics and biases to explain this behavior (Gilbert et al., 1988; Quattrone, 1982). Here, we show that in a probabilistic framework, this dispositional discounting when judging the situation is simply the mirror of situational discounting when judging disposition and requires no additional explanatory mechanisms beyond probabilistic inference.

We therefore begin by explaining our model for probabilistic social inference and demonstrate how an unbiased agent should make inferences about dispositions in nonconstraining situations. This model builds upon the framework that Morris and Larrick (1995) proposed to explain an apparent CB in the attribution literature and extends it to explain five classic results in the CB literature in which causes are not binary. The first two results allow us to demonstrate this framework's applicability in classic CB situations: How people attribute dispositions in the presence of explanatory situational pressures (Jones & Harris, 1967) or how prior beliefs influence dispositional attribution (Jones et al., 1971). With the third result, we show how this framework naturally captures the mirror of the CB: How people attribute the influence of situations when others' dispositions are known (Quattrone, 1982). Finally, we extend our model to explain how people make social attributions in the presence of more graded information, such as when actions can be weak or strong (Jones et al., 1971). Together, this demonstrates that rather than being based on a myriad of biases and heuristics, many instances of social discounting can be explained by a unified framework of Bayesian inference.

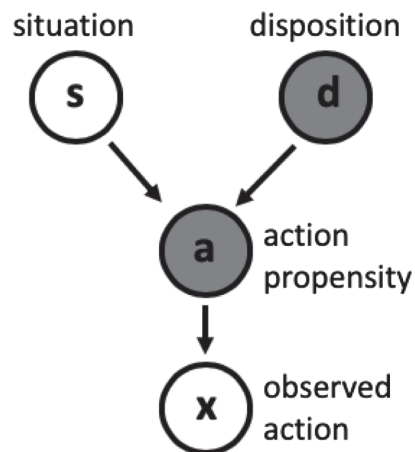
Probabilistic Social Attribution

Given the uncertainty inherent in reasoning about the causes of others' behavior, we cannot

expect people to make errorless social inferences. As Morris and Larrick (1995) have shown, the relevant question for assessing whether we make such inferences with systematic flaws is not whether we make errors at all, but rather whether these errors are inconsistent with those an unbiased observer would make. Thus, we must compare human social attribution to an unbiased observer operating with the same information.

How, for instance, would an unbiased observer infer the generosity of a museum patron who donates while a watchful docent is present? We have suggested that the influence of the situation and the influence of the person's disposition will combine to yield the probability of taking a specific action. This can be expressed as a simple four-node graphical model (Figure 1; Pearl, 1988): The probability of making a donation will be a function of the situation (*docent present/not present*) and the individual's disposition (how generous the person is). An observer can only see whether the donation was made and whether the docent was present; this observer must make an inference using this information to determine how generous a patron that donated is.

Figure 1
Graphical Model of How Situation and Unknown Disposition Combine to Produce Action



Note. This is a graphical model of an action arising from the combination of two classes of causes. White circles indicate observed variables, while gray circles indicate variables that must be inferred. Situation and disposition both influence the propensity for an action (the probability that an action will occur). Various attribution experiments amount to conditioning on (observing) two of the three nodes (usually situation and action) and inquiring about a third (disposition).

In this scenario, we treat behavior as a simple, binary action: The museum patron either leaves a donation or does not. We can express the binary action x (donate or not) as a draw from a Bernoulli distribution with some latent probability of donating (a), which we may think of as the individual's propensity to take the action in this situation. Situation (s) and disposition (d) will both influence this probability. For simplicity, and convention, we will assume that s and d are additive in log odds such as in (Equation 1):

$$\log \log \left(\frac{a}{1-a} \right) = s + d. \quad ((1))$$

This formulation, in which situation and disposition can take on graded values, extends the capabilities of Morris and Larrick's (1995) model, allowing for causes to have an influence that is continuous rather than dichotomous (*present/absent*). Thus, the probability of donating is given by the inverse logit (logistic) transform of (Equation 2):

$$a = \text{logit}^{-1}(s + d) = \frac{1}{1 + \exp^{-(s+d)}} \quad ((2))$$

Situational influence (s) and dispositional influence (d) can take on real values from negative infinity to positive infinity: Positive numbers reflect influences that encourage a behavior (*donating*) and negative numbers discourage the behavior (*not donating*). The log odds of an individual donating is thus the sum of the situational and dispositional influences expressed in this manner. Under this model, the probability of making a donation ($x = 1$) is a , and the probability of not making a donation ($x = 0$) is $1 - a$, as shown in (Equation 3):

$$p(x|a, \theta) = \{a, \quad \text{if } x = 1 \\ 1 - a, \quad \text{if } x = 0 \quad ((3))$$

where θ has no influence (it is a place holder for more sophisticated likelihood functions that could capture *graded* action strengths, described later).

This formulation¹ offers an intuitive interpretation of "situation strength" (s) and "disposition strength" (d) as our expectation about how people will act. A person with a disposition of $d = 0$ is equally likely to take the chosen action or not in an

unconstrained situation (e.g., a person with this disposition and no pressure from a docent would donate 50% of the time). People with positive disposition strengths will be more likely than chance to take the chosen action in an unconstrained situation, e.g., when there is no pressure from a docent a person with this disposition $d = 1$ would donate 73% of the time: $\log(a/(1-a)) = 1$; $a = 0.73$), and people with negative disposition scores will be less likely (e.g., $d = -1$) yields a 27% chance of donation for a person when no docent is present. The situation strength reflects how much the situation changes these probabilities. A nonconstraining situation has situation strength $s = 0$, and so the probability of taking an action (e.g., donating) relies only on the actor's disposition. Positive situation strengths represent conditions that encourage taking the chosen action (e.g., a watchful docent), while negative situation strengths represent conditions that discourage donating (e.g., learning people often steal from the donation jar). So, for example, if you expect 73% of people to donate without any pressure (average disposition of $d = 1$), but 92% of people donate when the docent is watching, this would indicate that the situational influence of the docent is $s = 1.5$.²

Although in most attitude attribution experiments, action can be considered to be dichotomous (either one action is taken or its alternative), in the real-world action is rarely dichotomous, but instead can take on fine gradations. When we are dealing with binary actions (*donating* or *not donating*), situation (s) and disposition (d) combine to influence the probability (a) that one of two outcomes occurs. However, if we want to consider the *intensity* of the action, we can expand the donation situation and imagine that the museum patron found an envelope containing \$20 in single bills before encountering the donation jar and can ask how much of this money does she donate? This allows us to treat the action not as binary but as a continuous variable, to capture the intuition that donating \$20 means something quite different than donating \$1. Thus, if we have only dichotomous information about an action,

¹ This framework shares similarities with item-response theory, a statistical approach to assess how responses to a set of questions is related to latent person variable (e.g., van der Linden & Hambleton, 2013).

² This follows from solving the equation: $\log(0.92/(1-0.92)) = s + 1$.

we link action propensity (a) to the action via a Bernoulli likelihood (yielding binary actions 0 or 1). But if we have graded information about action intensity, we can describe it as falling anywhere on the interval of [0 to 1], and we formulate the likelihood linking action propensity (a) to action (x) as a Beta distribution, such as in (Equation 4):

$$p(a, \theta) = \beta(x|a\theta, (1-a)\theta) \quad ((4))$$

where θ is the concentration parameter, indicating dispersion around the central tendency of a . This formulation yields action strengths ranging from *strongly negative* (0), *through neutral* (.5) to *very positive*. It is sufficient for capturing the data in the classic FAE studies where action intensity is reported on a bounded interval (usually a Likert scale).

So far, we have only explained how situation and disposition might combine to determine the probability that a specific action did or did not occur. But in most cases, people know the situation (or at least have a pretty good guess), observe an action, and must infer the *disposition*. Reasoning backwards, to infer the disposition, requires inverting the causal model, by relying on the rules of conditional probability and our prior expectations about the distribution of dispositions in the world: $P(d)$ —for example, how many people are more generous or stingy than average? Given this prior distribution on dispositions, and the observation of an action x in a situation of some strength s , we can calculate the posterior probability of the generosity of the actor using Bayes rule, such as in (Equation 5):

$$p(s, x) = \frac{p(d, s)p(d)}{\int_d p(d', s)p(d')}. \quad ((5))$$

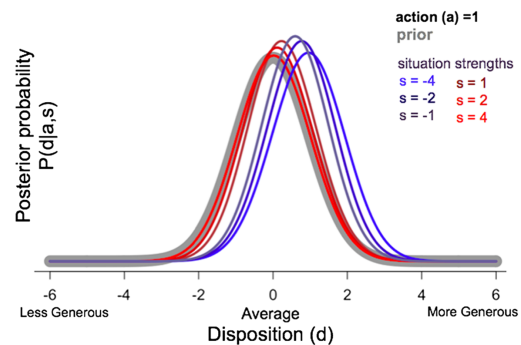
This calculation yields a posterior probability distribution over the disposition that the actor might hold: That is, which dispositions are likely given the situation strength we assumed, and the action we observed. Throughout our analyses, we will compare the judgments that people make about dispositions with the expected value of this posterior over dispositions.

What Makes a Situation Informative About Disposition?

Under the probabilistic attribution framework, an observer should always infer that there was

some influence of disposition on an observed action. However, the amount that is learned about the actor's disposition will depend on the strength of the situation. That is, according to this framework, if you see someone leave a donation at a free museum you should always infer some degree of generosity, but *how much* generosity will depend on how much the situation encouraged or discouraged this behavior. Figure 2 shows how the posterior distribution about the actor's disposition changes from the prior after observing a donation under different situation strengths. When there is strong pressure against donating, we learn a lot about how generous the person is—they must have been very generous in order to overcome strong pressure *not* to donate. But even when the situation encourages a donation, we still infer more generosity when we

Figure 2
The Posterior Distribution Over Disposition



Note. The posterior distribution over disposition after observing a positive action (e.g., donating; $x = 1$), under different assumed situation strengths. The gray distribution represents the prior distribution of generosity (the generosity that would be inferred if nothing else was known about the situation or action). When the situation strongly discourages a donation (e.g., $s = -4$) but someone donates anyway a lot of generosity is inferred because it must have taken a lot of generosity to overcome the pressure of the situation. When the situation only weakly discourages the action ($s = -1$) but someone donates anyway less generosity is inferred because less generosity is needed to overcome the pressure of the situation. However, even when the situation slightly ($s = 1$) or moderately ($s = 2$) encourages a donation some generosity beyond the prior is still inferred. However, when the situation nearly forces a donation ($s = 4$) seeing someone donate yields essentially no information about the actor's disposition since nearly everyone would act this way regardless of their disposition. Thus, as situations become extremely strong, such that the situation alone can determine behavior, probabilistic attribution will yield results consistent with deterministic, "logical" attribution: No inferences about disposition will be made. See the online article for the color version of this figure.

observe a donation compared to the prior. It is only when the pressure to donate is so strong that nearly no one would refuse that the inferred generosity becomes indistinguishable from the prior.

Let us imagine we are now in a world with somewhat more stingy museum patrons than before, where only 50% of people would donate when no docent is watching, here represented by the prior $d \sim N(0, 1)$. Now you observe someone leave a donation ($s = 0$). Based on Equation 5 above, you should infer that the visitor is somewhat more generous than average, $E[ds = 0, x = 1] = 0.39$, Figure 3 point a.

But what if there is strong pressure against donating? For instance, if the donation jar is missing ($s = -3$), yet the person leaves a donation anyway, you should infer even more strongly that they are generous, $E[ds = -3, x = 1] = .75$; Figure 3 point b. If the action occurred despite pressure against the action, it must have been motivated by disposition, and strong dispositional inferences are made.

Conversely, if the situation strongly encourages the action, e.g., the docent lays on a guilt trip, telling the patron the museum is in financial trouble and needs her help to stay open ($s = 3$), the unbiased observer will still infer something about the patron's disposition, $E[ds = 3, x = 1] = .08$, Figure 3 point c, since there are some people who still would not leave a donation in that situation. As long as situations are not deterministic, the ideal observer should make *some* dispositional attribution, but the strength of that attribution should be modulated by situation strength.³

Applying Probabilistic Attribution to Classic Attitude Attribution Results

For the remainder of this article, we ask how well the probabilistic attribution framework captures human inferences in classic social attribution experiments. We limit our discussion to the classic studies that have been directly interpreted as evidence for a bias.⁴ We first consider Jones and Harris's (1967) seminal essay paradigm in which participants attribute a disposition to an essay author even when that author was forced to take that position. Because the CB literature is extensive and often uses modifications of this task that are tangential to predictions of probabilistic social attribution (e.g., Miller et al., 1981), we

only focus on versions of the classic task that allow us to test novel predictions beyond what could be demonstrated with this classic study. Specifically, we model three tasks: (a) a task where people have preexisting beliefs about the author's disposition (Jones et al., 1971), (b) a puzzling result for the CB hypothesis: When the classic paradigm is inverted—when people are asked to infer the strength of the *situation* after reading an essay written by an author with a *known disposition*—people overattribute to situations (Quattrone, 1982), (c) a task where the *strength* of the essay's argument is manipulated (Jones et al., 1971) to demonstrate how the model may be applied to nonbinary actions.

In all cases, we find a qualitative match between human behavior and probabilistic social attribution. For ease of comparing probabilistic inference with human judgments, for all of the studies we consider, we scale the inferred log odds quantities to a bounded scale via the inverse logit transform and scale the results reported from the original studies by the upper and lower bound of the scales used.⁵

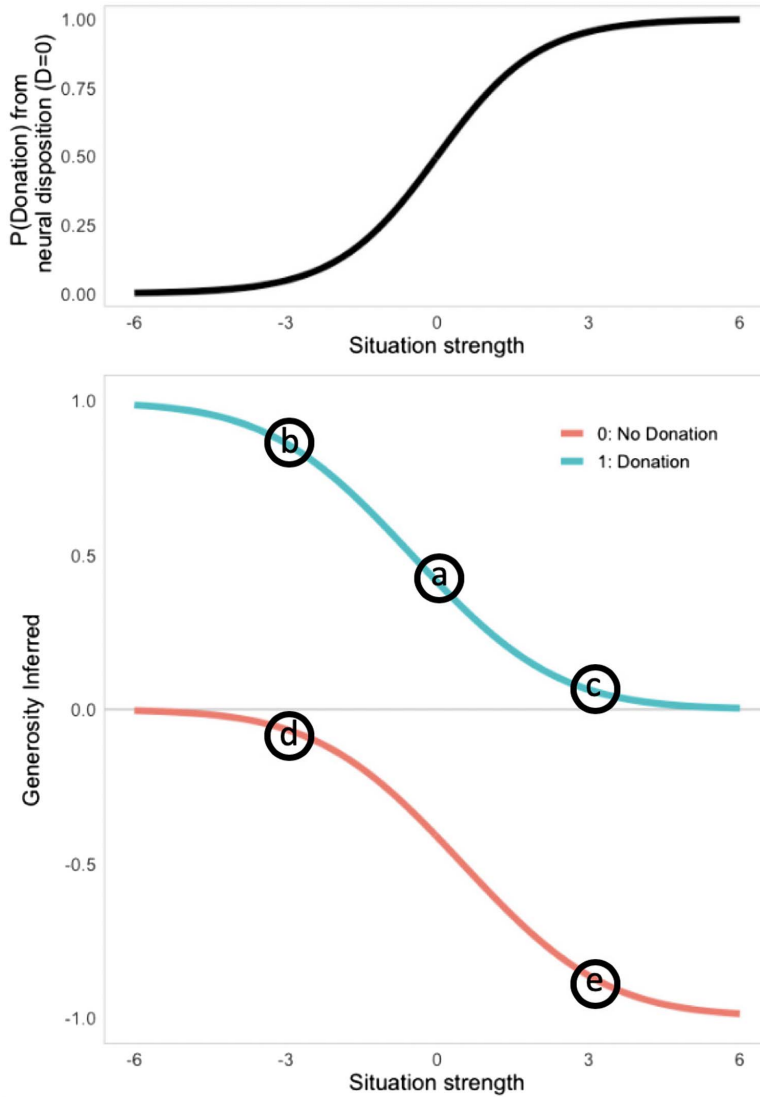
To generate predictions from a probabilistic attribution model for these classic studies, we must assume (a) what are the relevant situation strengths, given that they are not deterministic and (b) what are the relevant priors on attitudes or dispositions? Throughout the subsequent results we make an assumption that situation strengths are fairly strong ($s = 2$, yielding roughly 88% compliance from neutral participants). This assumption is conservative given that, on average,

³ A tool to explore how inferences change under different parameter values when actions are binary can be found at <https://edvul.shinyapps.io/fae-binary-demo/>. A tool for continuous actions can be found at <https://edvul.shinyapps.io/fae-continuous-demo/>.

⁴ Here, we do not, for example, consider process models such as studies that show that "cognitively busy" individuals make stronger inferences about disposition (e.g., Trop & Alfieri, 1997). Such work takes the CB as an *assumption*, but provides no explicit evidence per se for bias, and is thus beyond the scope of this analysis. We also do not address the role-conferred advantage paradigm (e.g., Ross, Amabile, et al., 1977) which requires accounting for the behavior of multiple actors simultaneously and is thus beyond the scope of this instantiation of our model.

⁵ For example, if the original experiment asked for attitude reports on a 1–6 Likert scale, we would transform the reported values y to the $[0, 1]$ interval via: $(y-1)/(6-1)$. For the ideal observer, we obtain a point estimate as the posterior mean log odds of disposition (or situation), then map it to the $[0, 1]$ interval via the logistic transform $1/(1 + \exp(-d))$.

Figure 3
Inferences of the Social Attribution Model Depending on Situation Strength and If Action Occurred



Note. The strength and direction of the inferred disposition by a probabilistic social attribution model depends on the situational pressure in combination with whether the action occurred. For example, if a donation occurred (blue) when there was a lot of pressure to donate, the ideal observer makes weaker inferences about the actors' generosity (point a), than in cases where there is no pressure to donate (point b). However, when there is stronger situational pressure that discourages donating and the person donates anyway, the ideal observer infers the actor is more generous than average (point b). Symmetrically, not donating (red) when there is strong pressure to donate (point e) suggests the person is far below average on generosity, compared to when the person does not donate but there was also strong pressure against donating (point d). The ideal observer will always infer a disposition consistent with the observed action (though this will become vanishingly small as situations become nearly deterministic; even more extreme than points c and d, respectively). See the online article for the color version of this figure.

participants in Morris and Larrick (1995) assumed that instructions to write an essay would result in 85% compliance, and Sherman (1980) found about 67% of students actually complied with the instruction to write an essay on a controversial topic. We also assume that priors on attitudes are fairly diffuse. We assume the prior distribution of anti-Castro attitudes is normally distributed with $M = 0$ and SD of 2.5. We will use this prior distribution for all subsequent results that do not explicitly manipulate prior expectations. The qualitative behavior of probabilistic attribution is not particularly sensitive to these assumptions: So long as situations are not so strong as to be deterministic, and priors are not so narrow as to preclude updating from an observed action, probabilistic attribution predicts incomplete discounting of behavior in the presence of situational influences (see Figure 3). Quantitative matches to human behavior, however, are sensitive to these details: Shifting or scaling the priors will shift and scale the posterior; increasing or decreasing situation strengths will increase or decrease discounting. Moreover, demonstrating quantitative matches to human behavior requires specifying how subjects mapped subjective estimates onto the Likert scales used in the original studies. Thus, we will not undertake the under constrained task of fitting parameters. Instead, our goal is to show that the *structure* of the probabilistic inference model, rather than the specific parameters used, yields the qualitative behavior observed throughout a broad range of classic results in the FAE literature.

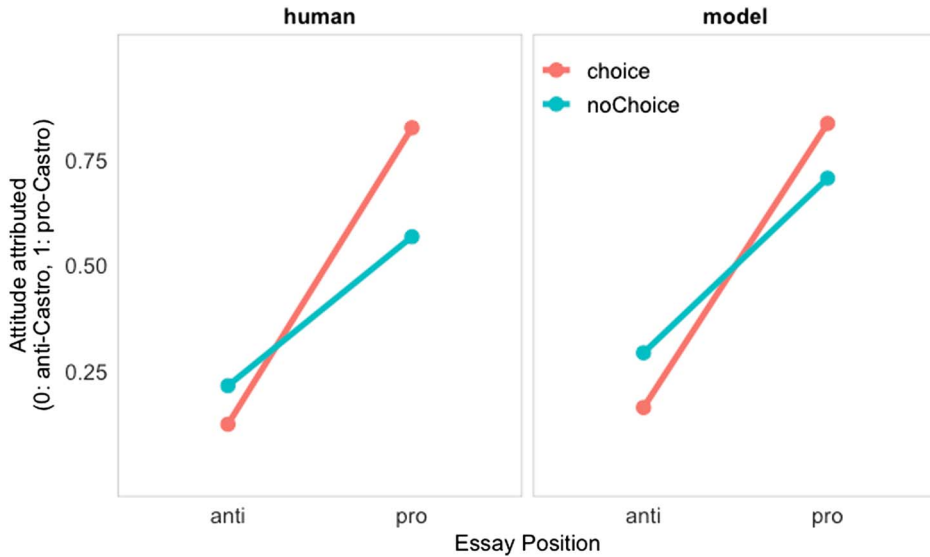
Inferred Attitude When Action Is Encouraged by the Situation (Jones & Harris, 1967)

In the classic CB experiment, university students read a pro- or anti-Castro essay and were told that the essay was written by a classmate either assigned or free to choose their position. Observers then answered ten 7-point Likert scale questions about their perception of the author's attitude; these 10 responses were summed to yield an *anti-Castro belief* (10) to *pro-Castro belief* (70) scale. Jones and Harris reasoned that a chosen position should reveal the authors' attitude to observers, but an assigned position should not be informative. As predicted, in the free-choice condition, observers judged the author's attitude to correspond to their essay position.

However, they still inferred a corresponding attitude (albeit more weakly) in the assignment condition, original data replotted in Figure 4, after scaling the 10–70 point scale to the interval [0, 1]. Jones and Harris (1967) concluded that people behave illogically, inferring something of the writer's attitude when the situation pressure should fully explain the behavior.

But what inferences should we expect from an unbiased observer who *did not* believe that instructions to write a particular essay are completely deterministic? Given the observation of either a pro- or anti-Castro essay (a binary action), and some assumption about the influence of instruction (situation strength) what might the actor's attitude about Castro be (disposition)? From the logic captured in Figure 3, we would expect that such an observer would infer *some* attitude that is consistent with the essay even when the position had been assigned. If the instruction to write a pro-Castro essay does not completely determine behavior, then those with vehemently anti-Castro views might still write an anti-Castro essay; therefore, seeing a pro-Castro essay still tells us *something* about the author's attitude, namely that the person does not dislike Castro enough to resist writing a pro-Castro essay when asked to.

To what degree an unbiased observer infers a dispositional cause depends on the observers' assumptions about how compelling the situation is. To formalize this, we must specify the "situation strength" of choosing one's position and that of being assigned a position, as well as the prior distribution about Castro attitudes. Changing the mean of this distribution yields a roughly uniform shift in inferred attitude across all four conditions, while changing the variance increases the magnitude of the inferred attitude (see Figure 5). We assume that the free-choice condition imposes no influence on the position that a writer would take ($s = 0$); such that a neutral person (not average, but split between positions; $d = 0$) who chooses what to write would be equally likely to produce a pro- or anti-Castro essay. Further, we assume that the assignment to write a pro- or anti-Castro essay has situation strengths that would compel a perfectly neutral person to write the assigned essay 88% of the time ($s = 2$ and -2 , respectively). We will again use these situation strengths for all subsequent studies. Since the situation strength determines how much "discounting" one does when inferring attitude, stronger (further from 0)

Figure 4*Inferred Attitude When Action Is Encouraged by the Situation (Jones & Harris, 1967)*

Note. Inferred attitude as a function of essay position, and whether this position was chosen or assigned. *Left:* Observers inferred the essay was indicative of the author’s attitude in both the choice and no-choice condition (Jones & Harris, 1967). Stronger inferences occurred when the position was chosen (pink line), and weaker when assigned (blue line). *Right:* An ideal observer also infers that the essay is indicative of the author’s true attitude but is more informative when the position was chosen versus assigned (blue line). See the online article for the color version of this figure.

situation strengths produce less of a FAE (yield no correspondent inferences when situation strengths are nearly deterministic), and weaker situation strengths (closer to 0) yield less discounting and larger correspondent inferences. Finally, we obtained ratings on a [0, 1] scale from the posterior belief about disposition via a logistic transform.⁶

Under these assumptions, an ideal observer infers the same pattern of dispositions as people do: In the “choice” conditions, the ideal observer treats the attitude expressed in the essay as very informative and infers that the author’s true attitude roughly mirrors what was expressed in the essay. In the “no-choice” condition, both humans and the ideal observer treat the behavior as informative (though less so) and make correspondingly weaker dispositional attributions (Figure 4, panel B).

It is critical to note that the *qualitative* pattern of results (weaker, but nonzero, attribution of attitudes in the no-choice condition) holds regardless of the specific assumptions we make about prior distributions over disposition or situation strength in the no-choice condition. As shown in Figure 5, decreasing the variance of the prior on

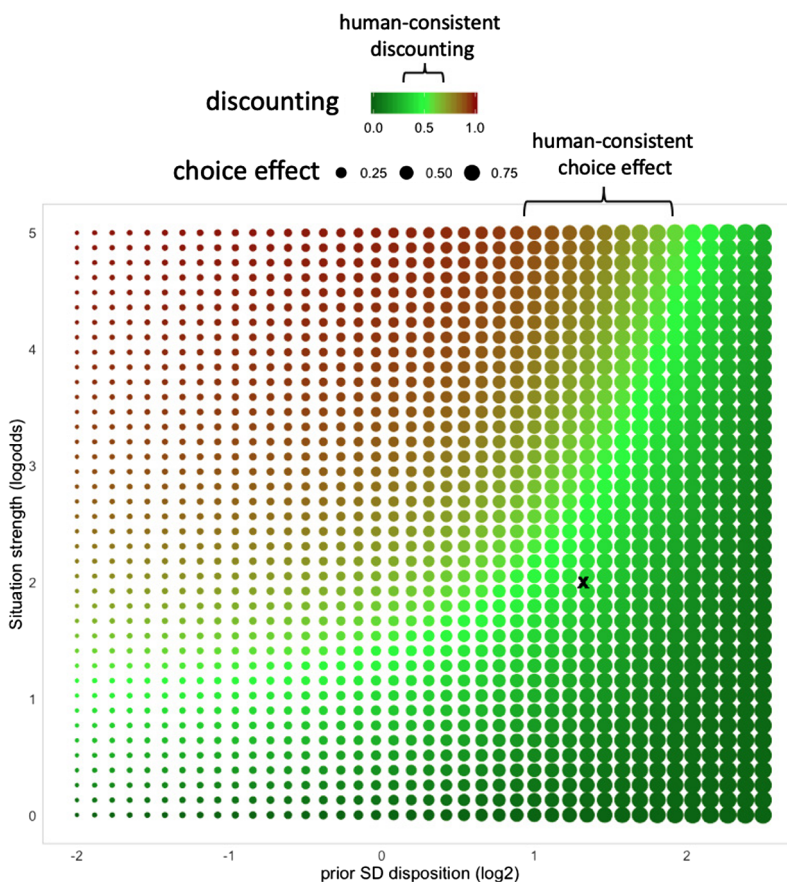
disposition shrinks the magnitude of the inferred attitudes in both choice and no-choice conditions, while altering situation strengths changes the amount of discounting in the no-choice condition. For this model to *not* yield a CB (e.g., to discount nearly completely in the no-choice condition), situation strengths need to exceed the variability in disposition by a factor of about four (yielding greater than 98% compliance).

The Influence of Preconceptions on Inferred Attitude (Jones et al., 1971)

Real-life social situations typically contain much more context that we use to flavor our inferences about other people, compared to what was provided in the previous experiment. For example, in the USA, if you meet someone at a National Rifle Association rally, that person is more likely to be politically conservative, compared to someone you meet at a vegan potluck. Jones et al. (1971) investigated how prior

⁶ Reported pro-Castro attitude = $10 + 60/(1 + \exp(-E[\Delta a, s]))$.

Figure 5
Sensitivity to Model Parameters



Note. Sensitivity to model parameters. Here, we show a wide range of possible parameters we might use in the model. For each possible prior standard deviation of disposition (x , log base-2 scale) and situation strength (y), we show the magnitude of the pro-anti effect in the choice condition (size of dots), and the amount of discounting in the no-choice condition (magnitude of the pro-anti effect in the no-choice condition as a proportion of the effect in the choice condition; color). The key result—partial, but incomplete discounting of behavior in the presence of situational pressures (k holds for a wide range of prior standard deviations on disposition (x) and situation strengths (y)). For small prior dispersions, the overall size of the pro-anti effect (in both choice and no-choice conditions) is much smaller than human behavior. For particularly strong situations (with nearly deterministic rates of compliance, log odds > 4), discounting is nearly complete. For weak situations, discounting is even smaller than observed for human behavior. See the online article for the color version of this figure.

expectation and the intensity of an action affect attribution. Subjects first read fake responses to a political questionnaire designed to alter their expectations about how generally conservative the essay author was, and thus how likely they are to support/oppose marijuana legalization. They read an essay they believed was freely chosen, or assigned, which either favored or opposed

legalization, and then estimated the author's attitude about legalization on a 6-point Likert scale (1 = *strongly anti-legalization* to 6 = *strongly pro-legalization*).

Jones et al. (1971) found that when the essay was consistent with expectations (an anti-legalization essay from a conservative or vice versa), readers estimated that author's attitude

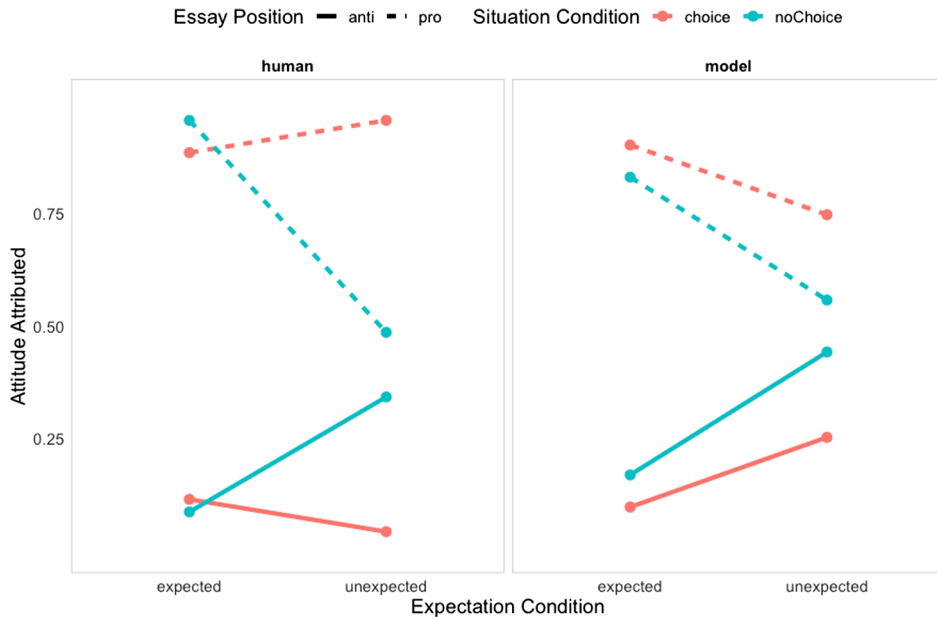
was consistent with both the opinion expressed and the prior expectation, regardless of whether the person was assigned or chose the position. However, when the essay was inconsistent with expectations (e.g., a pro-legalization essay from a conservative), attributions differed depending on whether the essay position was assigned or freely chosen: Observers inferred an attitude more consistent with the essay position when they believed the author chose his position, compared to if they thought it had been assigned (Figure 6, left panel).

What inferences should we expect from probabilistic social attribution under the assumptions that (a) the instructions to take a certain essay position are not deterministic and (b) expectations about the author change based on the questionnaire? We would again expect inferred attitudes to be consistent with the observed action (as discussed previously in Figure 3), with the strength of this inference modulated by situation strength (weaker inference under the no-choice

condition); and we would expect that these inferences would yield deviations away from the expected attitude in the general population. Insofar as the essay is consistent with expected attitude, it will give us little cause to update our beliefs about the author, but when the essay is inconsistent, it may either reflect situational pressures overcoming expected dispositions or an error in our assumptions about the dispositions (or a combination thereof). Thus, the unexpected free-choice essay should give us the most reason to change our beliefs about authors' disposition, and this change in beliefs will be weaker in the no-choice condition—when the situation is known to have influenced the essay. In short, the probabilistic attribution model is expected to yield the same qualitative pattern of behavior as observed in humans.

To formalize these predictions, we again specify the “situation strength” of being assigned a position and of having free choice. For

Figure 6
The Influence of Preconceptions on Inferred Attitude (Jones et al., 1971)



Note. When the essay was consistent with expectations, observers (left) attributed a corresponding attitude to the author, regardless of whether this position was assigned or chosen. However, when the essay deviated from expectation observers took the essay as more diagnostic of the author's attitude when the author chose the position, rather than having it assigned. Likewise, the probabilistic attribution observer (right) infers an attitude maximally consistent with the action when the action and prior belief are consistent, regardless of situational pressures, but an action inconsistent with expected attitudes is more diagnostic of true attitude when there is not situational pressures. See the online article for the color version of this figure.

consistency, we retain the same situation strength as in the previous scenario: $s = 2$ for instructions to write an anti-legalization essay and $s = -2$ for a pro-legalization essay (these correspond to situations that influence 80% of neutral people to write an essay in the instructed position). Again, we assume that in the choice condition, the situation strength is 0—exerting no influence on the essay position. Finally, we assume equally strong expectations from the questionnaire manipulation: A person who is portrayed as conservative has an expected disposition $d \sim N(1, 2.5)$, and an ostensibly liberal author is assumed to have a disposition $d \sim N(-1, 2.5)$; this means that they would, on average write anti- and pro-legalization essays (respectively) 80% of the time when given the choice of which position to take. Again, we scaled posterior beliefs about disposition to the 1–6 point Likert scale: Reported attitude = $1 + 5/(1 + \exp(-E[d|a, s]))$.

Probabilistic social attribution with these prior expectations about authors' beliefs infers the same authors' attitudes as human subjects. When the essay position was expected, both humans and probabilistic attribution infer that the author held the expressed belief, regardless of whether the essay was freely chosen or assigned. However, when the essay direction was unexpected both humans and the ideal observer infer a stronger attitude in the direction of the essay when it was freely chosen compared to when it was assigned.

Again, it is critical to note that this qualitative pattern of inferences is robust to variation in parameters, and that the priors and situation strengths used here are the same as those used in the previous experiment.

Inverting the CB: Inferring Situation When Attitude Is Known (Quattrone, 1982)

The CB hypothesis posits that people overattribute behavior to disposition. Thus, under this account, we would not expect people to infer situational influences when a known disposition can account for the observed action. However, a curious finding suggests just the opposite: When people know an actor's disposition, they are more likely to "overattribute" the actor's action to situational pressures. This result is inconsistent with the conventional interpretation of the CB; however, the symmetry of overattribution to

disposition can be captured naturally in a probabilistic framing (e.g., see Figure 7 vs. Figure 4).

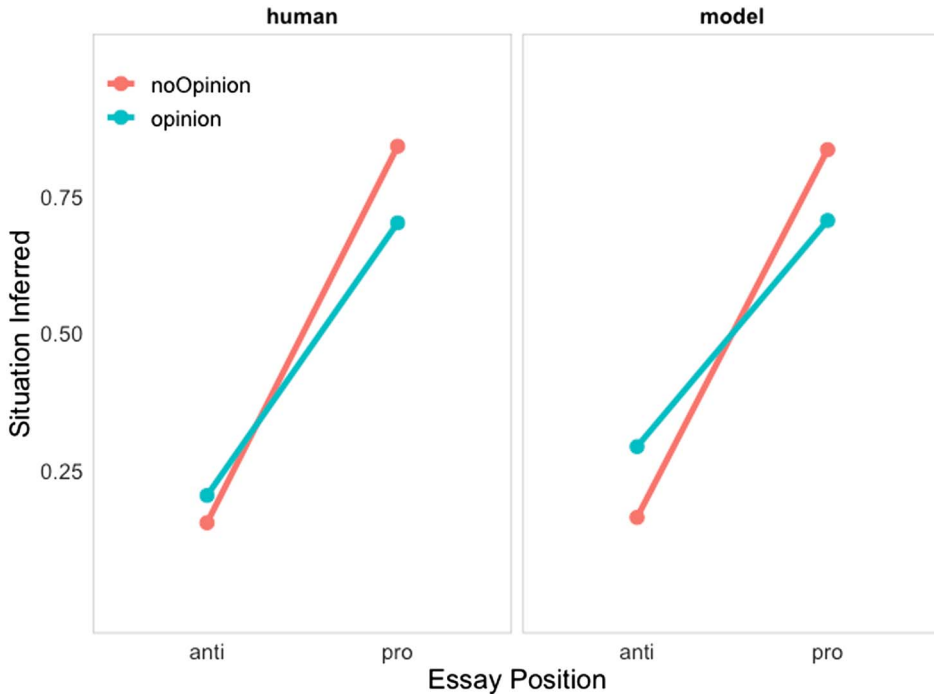
To manipulate disposition strength and measure inferred situational pressure, Quattrone (1982) had observers read an essay favoring or opposing the legalization of marijuana, but instead of telling readers that the essay position was chosen or assigned, they were told the author had either a neutral opinion about legalization, or one consistent with the attitude expressed in their essay, under the guise that the researchers were interested in potential effects of experimenter pressure (Quattrone, 1982). Subjects were next asked to estimate the likely situational pressure on a 30-point Likert scale ($-15 = \textit{pressure to oppose}$, $15 = \textit{pressure to favor}$). Even when subjects were told that the author held a pro-legalization view, they estimated that there was pressure to write a pro-legalization essay, and vice versa (original data replotted in Figure 7, panel A). This finding is the opposite of the classic explanation of the CB and calls into question the theoretical accounts that claim that we have an inclination to overattribute behavior to dispositions and not attribute enough to situations (e.g., Gilbert et al., 1988; Taylor & Fiske, 1978).

A probabilistic attribution account, however, naturally predicts this pattern of results. When someone behaves in a way that is motivated by their known disposition, it is still reasonable to infer that the situation was also motivating the action, given that probabilistic dispositions do not completely determine behavior. Assuming the same generative process as explained previously (Figure 1), inferring the unknown situation strength given a known disposition is symmetric inferring the disposition given a known situation (Figure 8). Knowing the disposition and what action the agent chose, but having a prior distribution over types of situations people encounter, we can use Bayes' formula to derive a posterior probability of the impact of the situation, such as in (Equation 6):

$$P(a, d) = \frac{P(s, d)P(s)}{\sum_s P(s, d)P(s)}. \quad ((6))$$

This framework provides mirrored inferences to the framework used to reason about disposition: Probabilistic attribution should always estimate that the situation had *some* influence in favor of the observed action. Just as we would infer that

Figure 7
Inverting the Correspondence Bias (Quattrone, 1982)



Note. Inferred situation as a function of the essay and the known attitude. *Left:* Subjects inferred that the position expressed in an essay indicated the situation was motivating behavior, both when they thought the author had a preexisting attitude and when they did not. The situation inferred was stronger when they thought the author had no preexisting opinion. *Right:* The ideal observer also infers that the situation was pressuring the essay position, but more so when the author had no existing opinion. See the online article for the color version of this figure.

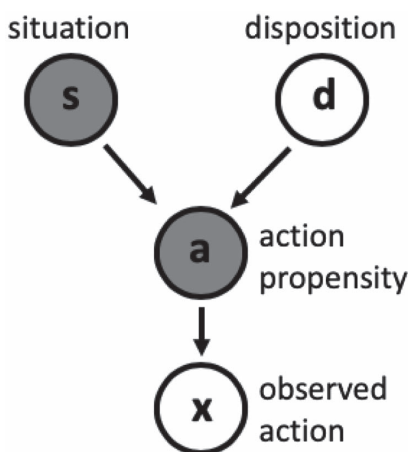
a museum patron who gives a donation is somewhat generous even when a docent is watching, we should also infer that when a generous friend donates, the docent is likely exerting some pressure on her. And just as an action is more informative of an actor's disposition in situations exerting weak (or contra-action) influences, the action is more informative of situations when dispositions are weak (or oppose the observed action). Again, so long the actor's disposition does not compel them to act identically in all situations, it is reasonable to infer that the situation had some influence.

In the Quattrone (1982) task, the ideal observer model again observes a binary action (either a pro- or anti-legalization essay) but now knows the author's attitude and must infer the strength of the situation. We formalize the "no opinion" attitude as a disposition strength of $d = 0$ (equally likely to write a pro- or anti-legalization essay under no situational pressure), and the "existing opinion"

condition has a disposition strength of $d = 1$ and -1 (for pro- and anti-legalization essays, respectively; meaning that these people would write essays consistent with their opinions 73% of the time when given the choice). Just as before, we used a logistic transformation and rescaled the expected posterior situation strength to place it on the same scale as Quattrone (1982).

Again, probabilistic social attributions are consistent with humans, and in this case, both are inconsistent with the classic CB account (Figure 8). Readers estimate that the experimental situation influenced authors toward the position expressed by the essay, both when the author purportedly had no prior opinion and when the authors were reported to have a prior opinion consistent with the essay position (albeit to a smaller degree). Again, we note that the qualitative pattern of inferences is robust to parameter variation (see Appendix A). Thus, the probabilistic attribution model can capture both the CB effect, as well as

Figure 8
Graphical Model of How Unknown Situation and Known Disposition Combine to Produce Action



Note. Graphical model shows that situation and disposition influence the probability that an action will occur. Instead of conditioning on (observing) situation and the action, and inquiring about situation, here we condition on disposition and action and inquire about the strength of the situation.

the inverse CB effect, where human behavior is directly opposite to the predictions of the CB hypothesis.

The Influence of Action Intensity on Inferred Attitude (Jones et al., 1971)

Since real-world behaviors are not easily classified as dichotomous (*donate/do not donate*), but rather fall along a continuum of extremeness (e.g., donating a lot or a little), it is useful to consider how people treat varying action strengths. To test this Jones et al. (1971; Experiment 2) used four essays, varying both the direction (pro- or anti-marijuana legalization) and extremeness: Strongly anti-legalization, weakly anti-legalization, weakly pro-legalization, and strongly pro-legalization. Observers each read one essay which they believed was chosen or assigned and estimated the author's attitude a 6-point Likert scale.

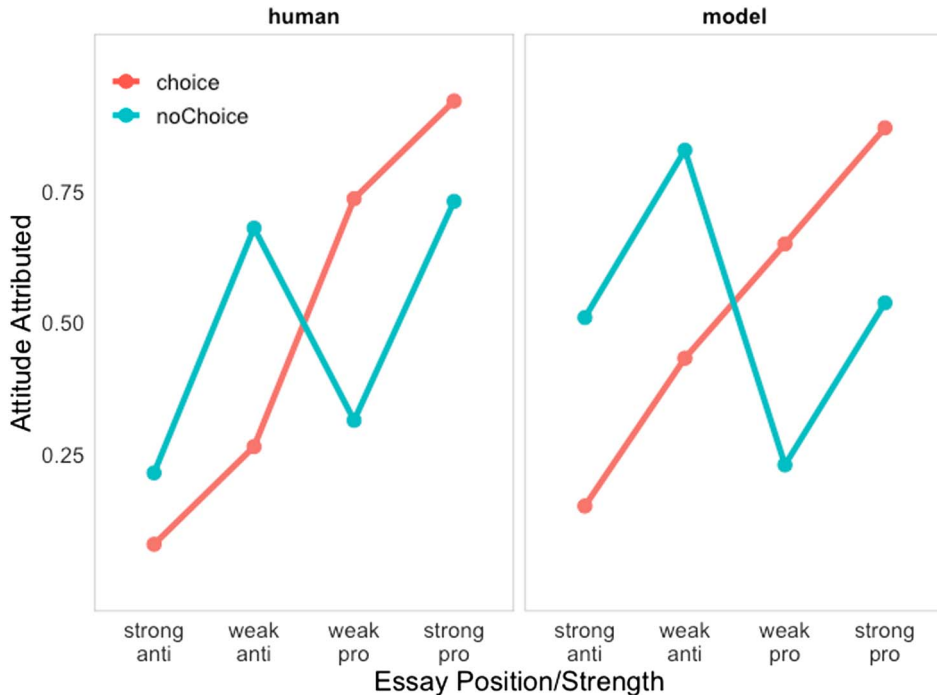
Subjects thought that the author's attitude scaled with the expressed position in a freely chosen essay: Inferred legalization attitude changed monotonically from a strong anti essay, to weak anti essay, to weak pro essay, to strong pro essay. However, when they thought it had been assigned, this monotonic pattern was

dramatically disrupted: A weak anti essay was interpreted as indicating a more pro-legalization attitude than a weak pro essay. That is, when someone was assigned to write a pro-legalization essay, but made a weak argument, readers inferred that they were actually against legalization, and vice versa (Figure 7, left panel). This negative inference from a positive action is analogous to the inferences we might make from a weak letter of recommendation: Since letter writers often feel obligated to write a letter upon request, it is not the letter itself, but the strength of the letter that provides us with the most information.

Again, while the qualitative pattern of results is expected under the probabilistic attribution model regardless of parameter details, we will evaluate the predictions using as many of the same parameter values we had used in the previous demonstrations. We used the same situation strengths as previously ($s = 1.38$ and -1.38 ; i.e., 80% of people who are neutral about legalization would be compelled to write in the direction they were told), and we assumed prior beliefs about attitude to be centered on neutral, $d \sim N(0, 1)$.

Since in this experiment the actions were explicitly nonbinary, we switched to the beta-likelihood function (Equation 4), which relates action propensity to a continuous scale from 0 to 1 (0, again is a concentration parameter—how much actions are assumed to vary around .5; it was arbitrarily set to 10). Accordingly, we re-scaled the four essay strength ratings from the 1 to 10 scale, they were rated on by Jones et al. (1971) onto this 0 (*strongly favor legalization*) to 1 (*strongly against legalization*) scale. For example, an essay rated as a 6 on the 10-point scale would correspond to an action strength of 0.55, which can be interpreted as a percentile: Roughly 55% of essays one can imagine would be more strongly against legalization, and 45% would be more strongly in favor of legalization.

Similar to the inferences people make, when the situation exerts no pressure, probabilistic attribution infers authors' attitudes that scale with essay strength (Figure 9, right panel). However, when the position was externally motivated, a weak essay is taken as evidence that the authors' true attitude is actually the opposite: A situation that encourages a pro-legalization essay would yield a relatively strong essay (in the 80th percentile) from a neutral person; seeing an essay that is weaker than that expectation (in the 60th percentile)

Figure 9*The Influence of Action Intensity on Inferred Attitude (Jones et al., 1971)*

Note. Both human subjects (left) and the probabilistic attribution model (right) infer that a strong essay reflects the author's attitude (and more so for a freely chosen essay). However, a weak essay only indicates a consistent attitude if the essay position was freely chosen; a weak essay taking a position assigned by the experimenter indicates that the author is actually likely to oppose the essay position. See the online article for the color version of this figure.

implies that disposition pushed the author in the direction opposite from the situational pressure.

Extensions and Limitations

The general logic of probabilistic inference is well suited to capture a wide variety of other interesting results in social attribution literature not addressed here, even findings that seem, at first blush, inconsistent with optimal discounting. For example, Snyder and Frankel (1976) demonstrated that despite the same observable behavior, observers infer that a woman in a situation motivating anxious behavior (talking about sex) has a more anxious disposition than the same behavior when observers think she is talking about politics. However, as Trope (1986) later argued, behavior does not exist in isolation; we use situational context to interpret behavior: In an early stage we go through a behavior identification

procedure, and only then do we engage in the dispositional inference process. An expanded probabilistic attribution framework would provide a natural mechanism to capture these types of findings: Situation and disposition combine to probabilistically produce an internal state (e.g., anxiety vs. boredom), which in turn produces a visible but ambiguous action (e.g., fidgeting). We use observed action and known situations to make an inference about the value of the internal state. Given the same observed physical behavior (e.g., fidgeting), a stronger known situation (e.g., sex vs. politics as a topic) would result in stronger inferred internal state (e.g., anxiety vs. boredom). Thus, when we think the person is discussing sex, we interpret her ambiguous behavior as anxious, and thus she subjectively displays anxiety only in the sex condition, leading to a stronger dispositional attribution of anxiety, despite any discounting we might do for the situation.

Another set of interesting empirical findings that have puzzled researchers is that observers make stronger social inferences when an actors' performance and situation strength covary (e.g., the rank order of performer ability and rank order of task difficulty perfectly covary), compared to when there is variation in only one causal factor (observers evaluating performers witness no evidence of task difficulty), a result which these authors have interpreted as inconsistent with Kelley's (1967, 1973) covariation model of social inference (Fiedler et al., 1999). Although the current formalization of our probabilistic social inference is not designed to formally model such accounts, the finding that two variables that covary will be mutually reinforcing, is a natural consequence of probabilistic reasoning as joint inference over situation and ability (it is also the foundation item-response theory, to which a probabilistic reasoning framework shares similarities; e.g., van der Linden & Hambleton, 2013).

The current instantiation of probabilistic social reasoning model, however, is not able to accommodate every social attribution process that has been proposed. For example, the abnormal conditions focus mode of causal attribution (Hilton & Slugoski, 1986) delineates an account of how natural language provides different information about how much consensus there is between individuals, how distinctive an action is in a given situation, and how consistent behavior is across situations, which moderates causal attributions. There is no straightforward way to explicitly align our account with these types of verbal frameworks. That said, the goal of the present article is to illustrate that the judgments observers make in the CB does not contradict a sound inference process and certainly does not preclude the possibility that there are other accounts equipped to show that this classic data can be produced from a reasonable reasoning process.

Discussion

Our results show that human attribution of behavior to situational and dispositional causes—which has long been considered systematically biased to overestimate dispositional influences—is consistent with sensible inferences if situations

and dispositions are thought to influence behavior probabilistically rather than deterministically. This probabilistic social attribution model yields the patterns of behavior classically interpreted as evidence of the CB and can capture how such behavior varies due to prior expectations about attitudes, as well as varied and ambiguous action strengths. Furthermore, the probabilistic attribution model explains a pattern of "overattribution" to situations, which is the opposite of the predictions of the FAE hypothesis. Classic experiments on social attribution, which have been interpreted as evidence of a systematic error, seem instead to yield robust evidence that human social attribution reflects reasonable inferences in a world where neither situations nor attitudes are sufficient to fully determine behavior. It is, however, worth noting that the interference pattern in the classic data would also occur if instead of making no errors, observers make two: They first assume incorrectly that the situation would deterministically compel behavior (misjudge situation strength), and despite this, fail to properly explain away the contribution of disposition (an error in the reasoning process). We offer a simpler explanation that observers make a reasonable estimate of the situation and then reason rationally given that estimate; the two-error account is less parsimonious and therefore it is that account that requires more evidence to be compelling. Furthermore, it is of course possible that people misestimate certain classes of social constraints which result in inaccurate inferences about disposition, but these biases would not be fundamental errors in the social reasoning process, but instead specific errors of interpreting particular situations.

Our intention in this work is to broadly demonstrate that behavior traditionally interpreted as a bias in the classic CB literature is also consistent with an unbiased probabilistic reasoning framework. Our intention here is not to suggest that this *specific* instantiation of probabilistic inference is necessarily the precise way in which humans use information to reason socially. It is critical to note that a "rational" model might be rendered consistent with human judgments merely by adding a biased prior about the strength of situations (i.e., by supposing that situational constraints are systematically underestimated); however, this would amount to merely reframing the CB

in probabilistic jargon. For instance, Jennings (2010) assumed that reasoning about dispositions could be explained by Bayesian inference using a biased prior and showed that people's attributions could still be internally consistent. Our account does not rely on such a strategy, which is perhaps most clearly illustrated in the fact that we can capture a situation in which people behave inconsistently with the typical explanation of the CB: Overattributing behavior to the situation when disposition is known (Quattrone, 1982; Experiment 2). This result, and the rest of those we present, do not arise from simply building in miscalibrated expectations about situations, but rather arise from the structure of probabilistic causal inference.

An appealing property of the social attribution model is that it allows for parametric manipulation of perceived situation strength and makes precise quantitative predictions about what human inferences *should* be if reasoning in accordance with an ideal observer. For example, the bottom panel of Figure 3 shows the inferences of an ideal observer for a wide range of situation strengths (nearly deterministically motivating an action, to nearly deterministically discouraging the action). One could gather empirically grounded judgments of situation strength for a variety of situations (e.g., "what proportion of people would write the assigned essay?") to use to produce precise quantitative predictions and confirm qualitative predictions. For example, if observers are acting in accordance with unbiased probabilistic inference, their inferences should follow the pattern of inferences shown in Figure 3, and this pattern should hold regardless of whether the situation or disposition is the unknown variable. Further, in the extreme cases in which the situation (or disposition) is perceived to be deterministically strong, the behavior should give the observer no information about the actor's disposition (or situation) and nothing should be attributed to the unknown variable. Walker and Vul (2021) did something similar using a game situation. They asked one group of participants to estimate how compelling different levels of a game were (e.g., "what proportion of people would make/miss a coin into a shot glass versus a kiddie pool from 5 feet away?"). They then asked a separate group to make disposition (skill) judgments based on an actor making or missing

a coin into various containers, and using the situation strength parameters determined previously, showed that observers made a situation inferences consistent with the predictions of the unbiased social inference model. These judgments are not ground truth as we do not actually know how many people of average skill would make a coin into a particular container, but they go beyond showing the inferences are internally consistent with one individual, average independent assessment of the strength of a situation.

That said, the critical contribution of the present article is to show that the causal structure employed in the probabilistic social attribution model is sufficient to capture the effects of the social attribution literature. The qualitative match between our model and human behaviors is not sensitive to variation in parameters: All sensible parameter values (please see Appendices A and B for elaboration) would yield the qualitative effects in the classic CB studies. In short, our work suggests that results from decades of attribution experiments, which have been classically interpreted as evidence that our social inferences are fundamentally flawed, might instead be the natural outcome of reasoning about a complex and uncertain world.

References

- Ajzen, I. (1977). Intuitive theories of events and the effects of base-rate information on prediction. *Journal of Personality and Social Psychology*, 35(5), 303–314. <https://doi.org/10.1037/0022-3514.35.5.303>
- Ajzen, I., Datto, C. A., & Blyth, D. P. (1979). Consistency and bias in the attribution of attitudes. *Journal of Personality and Social Psychology*, 37(10), 1871–1876. <https://doi.org/10.1037/0022-3514.37.10.1871>
- Ajzen, I., & Fishbein, M. (1975). A Bayesian analysis of attribution processes. *Psychological Bulletin*, 82(2), 261–277. <https://doi.org/10.1037/h0076477>
- Andrews, P. W. (2001). The psychology of social chess and the evolution of attribution mechanisms: Explaining the fundamental attribution error. *Evolution and Human Behavior*, 22(1), 11–29. [https://doi.org/10.1016/S1090-5138\(00\)00059-3](https://doi.org/10.1016/S1090-5138(00)00059-3)
- Fein, S., Hilton, J. L., & Miller, D. T. (1990). Suspicion of ulterior motivation and the correspondence bias. *Journal of Personality and Social Psychology*, 58(5), 753–764. <https://doi.org/10.1037/0022-3514.58.5.753>
- Fernbach, P. M., & Erb, C. D. (2013). A quantitative causal model theory of conditional reasoning.

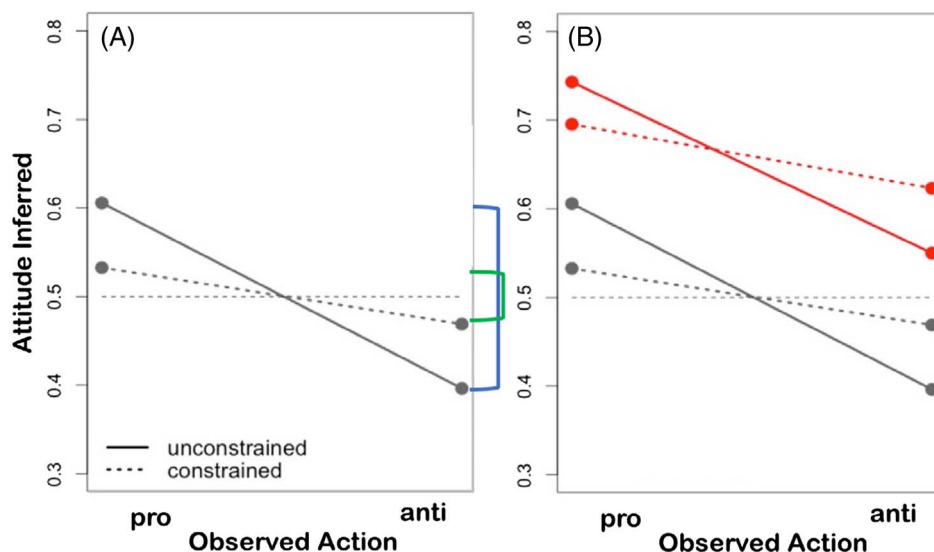
- Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(5), 1327–1343. <https://doi.org/10.1037/a0031851>
- Fiedler, K., Walther, E., & Nickel, S. (1999). Covariation-based attribution: On the ability to assess multiple covariates of an effect. *Personality and Social Psychology Bulletin*, 25(5), 609–624. <https://doi.org/10.1177/0146167299025005006>
- Gawronski, B. (2004). Theory-based bias correction in dispositional inference: The fundamental attribution error is dead, long live the correspondence bias. *European Review of Social Psychology*, 15(1), 183–217. <https://doi.org/10.1080/10463280440000026>
- Gilbert, D. T., & Jones, E. E. (1986). Perceiver-induced constraint: Interpretations of self-generated reality. *Journal of Personality and Social Psychology*, 50(2), 269–280. <https://doi.org/10.1037/0022-3514.50.2.269>
- Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, 117(1), 21–38. <https://doi.org/10.1037/0033-2909.117.1.21>
- Gilbert, D. T., McNulty, S. E., Giuliano, T. A., & Benson, J. E. (1992). Blurry words and fuzzy deeds: The attribution of obscure behavior. *Journal of Personality and Social Psychology*, 62(1), 18–25. <https://doi.org/10.1037/0022-3514.62.1.18>
- Gilbert, D. T., & Osborne, R. E. (1989). Thinking backward: Some curable and incurable consequences of cognitive busyness. *Journal of Personality and Social Psychology*, 57(6), 940–949. <https://doi.org/10.1037/0022-3514.57.6.940>
- Gilbert, D. T., Pelham, B. W., & Krull, D. S. (1988). On cognitive busyness: When person perceivers meet persons perceived. *Journal of Personality and Social Psychology*, 54(5), 733–740. <https://doi.org/10.1037/0022-3514.54.5.733>
- Haney, C., & Zimbardo, P. G. (2009). Persistent dispositionalism in interactionist clothing: Fundamental attribution error in explaining prison abuse. *Personality and Social Psychology Bulletin*, 35(6), 807–814. <https://doi.org/10.1177/0146167208322864>
- Hilton, D. (2017). Social attribution and explanation. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning*. Oxford University Press.
- Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, 93(1), 75–88. <https://doi.org/10.1037/0033-295X.93.1.75>
- Hooper, N., Erdogan, A., Keen, G., Lawton, K., & McHugh, L. (2015). Perspective taking reduces the fundamental attribution error. *Journal of Contextual Behavioral Science*, 4(2), 69–72. <https://doi.org/10.1016/j.jcbs.2015.02.002>
- Jennings, K. E. (2010). Determining the internal consistency of attitude attributions. In *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 978–984). Cognitive Science Society.
- Johnson, J. T., Jemmott, J. B., III, & Pettigrew, T. F. (1984). Causal attribution and dispositional inference: Evidence of inconsistent judgments. *Journal of Experimental Social Psychology*, 20(6), 567–585. [https://doi.org/10.1016/0022-1031\(84\)90044-1](https://doi.org/10.1016/0022-1031(84)90044-1)
- Jones, E. E. (1979). The rocky road from acts to dispositions. *American Psychologist*, 34(2), 107–117. <https://doi.org/10.1037/0003-066X.34.2.107>
- Jones, E. E. (1990). *Interpersonal perception*. Freeman.
- Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, 3(1), 1–24. [https://doi.org/10.1016/0022-1031\(67\)90034-0](https://doi.org/10.1016/0022-1031(67)90034-0)
- Jones, E. E., Worchel, S., Goethals, G. R., & Grumet, J. F. (1971). Prior expectancy and behavioral extremity as determinants of attitude attribution. *Journal of Experimental Social Psychology*, 7(1), 59–80. [https://doi.org/10.1016/0022-1031\(71\)90055-2](https://doi.org/10.1016/0022-1031(71)90055-2)
- Kelley, H. H. (1967). Attribution theory in social psychology. In *Nebraska symposium on motivation*. University of Nebraska Press.
- Kelley, H. H. (1972). Attribution in social interaction. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 151–174). General Learning Press.
- Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, 28(2), 107–128. <https://doi.org/10.1037/h0034225>
- Langdrige, D., & Butt, T. (2004). The fundamental attribution error: A phenomenological critique. *British Journal of Social Psychology*, 43(3), 357–369. <https://doi.org/10.1348/0144666042037962>
- Liefgreen, A., Tesic, M., & Lagnado, D. A. (2018). Explaining away: Significance of priors, diagnostic reasoning and structural complexity. *Proceedings of the annual meeting of the cognitive science society* (Vol. 41). Cognitive Science Society
- Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review*, 3(1), 23–48. https://doi.org/10.1207/s15327957pspr0301_2
- McClure, J. (1998). Discounting causes of behavior: Are two reasons better than one?. *Journal of Personality and Social Psychology*, 74(1), 7–20. <https://doi.org/10.1037/0022-3514.74.1.7>
- Miller, A. G., Ashton, W. A., & Mishal, M. (1990). Beliefs concerning the features of constrained behavior: A basis for the fundamental attribution error. *Journal of Personality and Social Psychology*, 59(4), 635–650. <https://doi.org/10.1037/0022-3514.59.4.635>
- Miller, A. G., Jones, E. E., & Hinkle, S. (1981). A robust attribution error in the personality domain. *Journal of Experimental Social Psychology*, 17(6), 587–600. [https://doi.org/10.1016/0022-1031\(81\)90041-X](https://doi.org/10.1016/0022-1031(81)90041-X)

- Morris, M. W., & Larrick, R. P. (1995). When one cause casts doubt on another: A normative analysis of discounting in causal attribution. *Psychological Review*, *102*(2), 331–355. <https://doi.org/10.1037/0033-295X.102.2.331>
- O’Sullivan, M. (2003). The fundamental attribution error in detecting deception: The boy-who-cried-wolf effect. *Personality and Social Psychology Bulletin*, *29*(10), 1316–1327. <https://doi.org/10.1177/0146167203254610>
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.
- Quattrone, G. A. (1982). Overattribution and unit formation: When behavior engulfs the person. *Journal of Personality and Social Psychology*, *42*(4), 593–607. <https://doi.org/10.1037/0022-3514.42.4.593>
- Reeder, G. D., Fletcher, G. J., & Furman, K. (1989). The role of observers’ expectations in attitude attribution. *Journal of Experimental Social Psychology*, *25*(2), 168–188. [https://doi.org/10.1016/0022-1031\(89\)90011-5](https://doi.org/10.1016/0022-1031(89)90011-5)
- Reeder, G. D., & Spores, J. M. (1983). The attribution of morality. *Journal of Personality and Social Psychology*, *44*(4), 736. <https://doi.org/10.1037/0022-3514.44.4.736>
- Reeder, G. D., Vonk, R., Ronk, M. J., Ham, J., & Lawrence, M. (2004). Dispositional attribution: Multiple inferences about motive-related traits. *Journal of personality and social psychology*, *86*(4), 530. <https://doi.org/10.1037/0022-3514.86.4.530>
- Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In *Advances in experimental social psychology* (Vol. 10, pp. 173–220). Academic Press.
- Ross, L., & Nisbett, R. E. (2011). *The person and the situation: Perspectives of social psychology*. Pinter & Martin Publishers.
- Ross, L. D., Amabile, T. M., & Steinmetz, J. L. (1977). Social roles, social control, and biases in social-perception processes. *Journal of Personality and Social Psychology*, *35*(7), 485–494. <https://doi.org/10.1037/0022-3514.35.7.485>
- Sabini, J., Siepmann, M., & Stein, J. (2001). The really fundamental attribution error in social psychological research. *Psychological Inquiry*, *12*(1), 1–15. <https://doi.org/10.1080/10478400802615744>
- Sherman, S. J. (1980). On the self-erasing nature of errors of prediction. *Journal of Personality and Social Psychology*, *39*(2), 211–221. <https://doi.org/10.1037/0022-3514.39.2.211>
- Snyder, M. L., & Frankel, A. (1976). Observer bias: A stringent test of behavior engulfing the field. *Journal of Personality and Social Psychology*, *34*(5), 857–864. <https://doi.org/10.1037/0022-3514.34.5.857>
- Taylor, S. E., & Fiske, S. T. (1978). Saliency, attention, and attribution: Top of the head phenomena. *Advances in Experimental Social Psychology*, *11*, 249–288. [https://doi.org/10.1016/S0065-2601\(08\)60009-X](https://doi.org/10.1016/S0065-2601(08)60009-X)
- Trope, Y. (1986). Identification and inferential processes in dispositional attribution. *Psychological Review*, *93*(3), 239–257. <https://doi.org/10.1037/0033-295X.93.3.239>
- Trope, Y., & Alfieri, T. (1997). Effortfulness and flexibility of dispositional judgment processes. *Journal of Personality and Social Psychology*, *73*(4), 662. <https://doi.org/10.1037/0022-3514.73.4.662>
- Trope, Y., & Gaunt, R. (1999). A dual-process model of overconfident attributional inferences. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 161–178). The Guilford Press.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>
- Uleman, J. S. (1999). Spontaneous versus intentional inferences in impression formation. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 141–160). Guilford
- van der Linden, W. J., & Hambleton, R. K. (Eds.). (2013). *Handbook of modern item response theory*. Springer Science & Business Media.
- Vonk, R. (1999). Effects of outcome dependency on correspondence bias. *Personality and Social Psychology Bulletin*, *25*(3), 382–389. <https://doi.org/10.1177/0146167299025003009>
- Walker, D., & Vul, E. (2021). Blame the player and the game. *Proceedings of the annual meeting of the cognitive science society* (Vol. 43). Cognitive Science Society.

(Appendices follow)

Appendix A

The Correspondence Bias pattern arises under a large range of parameter values



Note. See the online article for the color version of this figure.

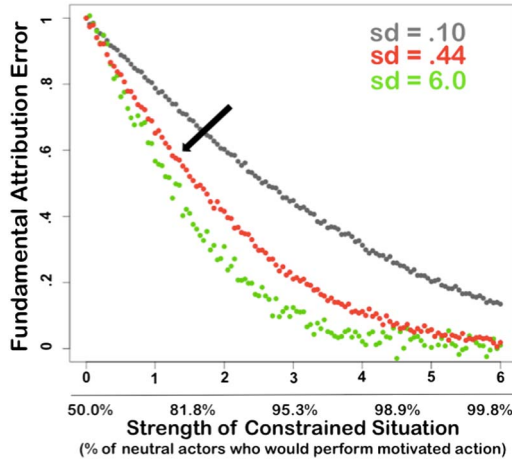
The correspondence bias error is not a peculiar phenomenon that arises from probabilistic attribution under a narrow set of parameter values. Instead, it is inherent to the structure of probabilistic attribution, and only under extreme (and unrealistic) parameter values would probabilistic attribution not exhibit this pattern. To illustrate behavior in classic tasks using a probabilistic observer, we needed to make assumptions about parameters not collected in the original studies: (a) observer's belief about situation strength, (b) observers' belief about the central tendency of peoples' dispositions, and (c) variability of dispositions in the world. When a binary action (e.g., a pro or anti essay) is observed, the degree of correspondence bias (CBs) can be thought of as

the difference between the attitudes inferred from the two distinct actions (difference between the points on the dotted gray indicated by the green bracket in panel A), normalized by dividing by the difference in the attitudes inferred from these same actions in the unconstrained situation (difference in the points on the solid gray line indicated by the blue bracket). This ratio represents the proportion of the difference in the inferred attitude that the two actions produce even when the situation is constraining. Note that changing the assumption about the average attitude in the population has no effect on the CB, it simply shifts all of the inferences, leaving the ratio representing the degree of CB intact (red lines in panel B).

(Appendices continue)

Appendix B

A Correspondence Bias pattern arises for a wide range of perceived situations strengths and variability in disposition



Note. See the online article for the color version of this figure.

Although belief about the central tendency of the disposition does not change the degree of the CB, the perceived strength of the constraining situation and the perceived dispersion of attitudes in the world do have an effect on the degree of CB. Critically, however, the probabilistic social attribution model produces the pattern of inferences indicative of a nonnegligible CB for a wide range of parameters values. Until the pressure of the constrained situation is interpreted to be nearly deterministically strong (a situation in which over 98% of neutral people would be compelled to

perform the action), we observe a pattern of inference consistent with a CB. This pattern holds regardless of whether the variability of dispositions is assumed to be *very low* (gray points) or *very high* (green points). The black arrow indicates the parameter values we adopted in our models of the classic empirical studies.

Received December 23, 2020

Revision received February 18, 2022

Accepted February 18, 2022 ■