Title
Biases in Social Perception Arise from Rational Inference

Author
Hoffman, Drew E.

Publication Date
2017

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Biases in Social Perception Arise from Rational Inference

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Psychology

by

Drew Ellen Hoffman

Committee in charge:

Professor Edward Vul, Chair
Professor Nicholas Christenfeld
Professor Craig McKenzie
Professor Dana Nelkin
Professor John Wixted

2017

The Dissertation of Drew Ellen Hoffman is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2017

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

This dissertation, and my graduate career, have been a collective effort and there are so many people who I am deeply indebted to for making these things possible and worthwhile.

First I would like to thank my advisor, Ed Vul. Ed has contributed greatly to this dissertation, as well as to my growth as a scientist and a person. He continually challenges me to think more clearly and precisely, and to approach problems in a principled and rigorous way. His feedback in all domains is thoughtful and honest, and inevitably improves my approach to any problem. His insatiable curiosity about human behavior, and ideas more broadly, is contagious and I feel truly privileged to have had the opportunity to be his student. Thank you for always pushing me to jump onto bigger boxes.

I am also extremely appreciative of my dissertation committee and the thoughtful feedback and direction they've provided me over the years. I would like to thank Nicko Christenfeld, whose profound and original insights are enviable, and have contributed much to this dissertation. I am thankful to have had his guidance and creative perspective on this and other work, and his mentorship through undergraduate and graduate school. I also appreciate Craig McKenzie's helpful comments, particular his ideas regarding details of experimental design, and also appreciate the conversations I have had with Dana Nelkin, which prodded me to think more deeply about how I think about how people think. I would also like to thank John Wixted for the many words of wisdom he has shared with me over the years, some of which I take to heart (e.g. writing well, like any skill, takes practice) and some of which I doltishly ignore (e.g. never say in an email what you would not want printed on the cover of the New York Times).

I would also like to thank my lab mates Kevin Smith, who made extensive contributions to this dissertation, and Tim Lew. Without Kevin and Tim I would not know how to make a computer do stuff, understand most xkcd punch lines, hold Karl Popper in high esteem, or know that the cake is lie. I have never enjoyed being the dumbest person in the room so much. Thanks also to Rob. St. Louis for his incisive feedback, and Kristin Donnelly, whose passion for psychology research reinvigorated mine when I needed it most.

Finally, I would like to thank my family. The length of this volume would double if I listed all of the ways you have supported me through this process, so I will just say thank you: Ryan Walker, for holding my world together during this process. Ellen Murphy and Drew Hoffman for showing me the virtue of hard work. Thanks also to: Hillary Hoffman, Reza Asgari, Laurie, Tom, Even, Joel and Michelle Walker, Jessica Wi, Larry McCarty and Laura Simons. I would also like take this time to remind these people of a little known international card game convention which states: in the case of a tie, or other dispute, the player holding the highest earned degree wins.

Chapter 1, in full, is a reprint of the material as it appears in *Psychological Science*. Walker, Drew E.; Vul, Ed. 2013. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is currently being revised for publication of the material. Walker, Drew E.; Smith, Kevin; Vul, Edward. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is currently being prepared for submission for publication of the material. Walker, Drew E.; Vul, Edward. The dissertation author was the primary investigator and author of this pager.

VITA

EDUCATION

Ph.D. in Psychology 2017
University of California, San Diego

M.A. in Psychology 2015
University of California, San Diego

B.A. in Psychology 2011
University at California, San Diego

- Honors

PUBLICATIONS

Walker, D, Smith, K. A., & Vul, E. (2015) The "Fundamental Attribution Error" is Rational in an uncertain world. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*

Walker, D., Bajic, D., Mickes, L., Kwak, J., & Rickard, T. C. (2014). Specificity of children's arithmetic learning. *Journal of experimental child psychology*,*122*, 62-74.

Walker, D. & Vul, E. (2014). Hierarchical Encoding Makes Individuals in a Group Seem More Attractive. *Psychological Science*, *25*(1), 230-235.

Walker, D., Mickes, L., Bajic, D., Nailon, R., & Rickard, T. C. (2013). A Test of Two Methods of Arithmetic Fluency Training and Implications for Educational Practice, *Journal of Applied Research in Memory and Cognition*. *2*(1), 25-32

Mickes, L., **Walker, D**., Parris, J. L., Mankoff, R., & Christenfeld, N. (2011). Who's funny: Gender stereotypes, humor production, and memory bias*. Psychonomic Bulletin & Review*, *19*(1), *108-112*

ABSTRACT OF THE DISSERTATION

Biases in Social Perception Arise from Rational Inference

by

Drew Ellen Hoffman

Doctor of Philosophy in Psychology

University of California, San Diego

Professor Edward Vul, Chair

The social information available to us at any given moment is, at best, ambiguous. Yet, remarkably, we are able to efficiently resolve this ambiguity and successfully navigate the social world. In this dissertation I use a rational inference framework to understand how we form rich, and largely accurate, social perceptions given this uncertain and underconstrained information. Our perceptions, of course, do not always perfectly align with reality, but – contrary to the classic perspective in social psychology – this is not evidence that we are irrational. In this dissertation I show how social biases can arise not as a *failure* of rationality, but as a *consequence* of making optimal use of statistical structure in the world. In Chapter 1, I demonstrate that our visual system's strategy to circumvent resource limitations by capitalizing on redundancy in visual scenes can result in a bias to perceive faces in a crowd as more attractive. In Chapters 2 and 3, I show that two of the most well-known social biases – The Fundamental Attribution Error and Role-conferred Advantage – are not actually evidence of irrational reasoning.

Although in these paradigms observers seem “bias” to systematically make attributions that are in a direction consistent with observed behavior, these judgments fall naturally out of optimal probabilistic inference.

GENERAL INTRODUCTION

There is a long and powerful tradition in psychology to view human social reasoning as irrational and full of biases (e.g., Ross & Nisbett, 2011). And although biases are indeed perceptions that differ from *reality*, the traditional view is that these inaccuracies are caused by reasoning that deviates from *rationality*, and are indicative of information-processing failures. But errors in our impressions and perceptions can arise even when we are using information in a reasonable or even optimal way, simply because we are working with incomplete or uncertain information.

Social reasoning usually involves underconstrained problems, wherein the available information is insufficient to identify a single answer, so even an ideal observer using all available information must exercise some 'inductive bias' (Griffiths, et al. 2010) to come up with an answer, and must thus produce some systematic errors. Solving unconstrained inference problems is not just an issue when we are trying to figure out the source of complex social behavior, but is present in even the most basic interactions we have with our environment. For example, we determine the color of an object based on the wavelength of light that hits our retina, but because the spectrum we sense is determined not just by the reflectance of the objects' surface, but also the spectrum of the light that is illuminating the object. It is impossible to deduce the state of the world based on only the information directly available to us, since multiple combinations of causal factors could yield the same observed experience, yet somehow we are able to come up with internally consistent percepts that are generally well calibrated to the world. Bayesian probabilistic inference has been used as a framework to understand how humans might solve such underdetermined problems by using 'inductive biases' that reflect accumulated knowledge about the statistical structure in the world — for example using prior probabilities about surfaces and illuminants – to make reasonable guess about what is the most likely source of the data we are experiencing. In this dissertation I adopt this rational inference framework to understand how we form rich, and largely accurate, social perceptions given incomplete and ambiguous information, and show how social biases can arise not as a *failure* of rationality, but as a *consequence* of making optimal use of statistical structure in the world.

In Chapter 1 I demonstrate how the optimal use of statistical structure in visual scenes can result in a bias to perceive people to be more attractive than they are. Our visual system is bombarded by more information than we can effectively process – whether we are looking at leaves on a tree or faces in crowd, we are often unable to accurately encode all of the individual elements presented to us. But a lot of this information is redundant, and our visual system takes advantage of this redundancy by automatically extracting a summary representation of all the similar elements (Ariely, 2001, Chong & Treisman, 2003). When we have perceptual uncertainty about an individual item we perceive it to be biased toward the group average, which, in aggregate, makes our judgment better (Brady & Alvarez, 2011). I show that this useful tendency to perceive items as bias toward the ensemble average has the side effect of systematically distorting faces to be perceived as more attractive than they are. When we see faces in a group our visual system automatically forms an average representation, and the individual faces are perceived to be biased toward this group average. Since average faces tend to be more attractive (Langlois & Roggman, 1990), this biasing of individual faces toward a group average causes the "cheerleader effect": a face in a group looks better than the same face alone.

Importantly, although this "cheerleader effect" is indeed a perception that systematically deviates from reality, it is not the result of an information-processing failure. Quite the contrary, ensemble coding is a sophisticated processing strategy that allows us to overcomes cognitive resource limitations by making optimal use of available information. It makes fast and efficient visual processing possible, and even aids real-time social interactions, such as allowing us to quickly discern the collective direction of a crowds' gaze (Sweeny & Whitney, 2014) the direction that a crowd is moving (Sweeny et al., 2013), and a groups' average emotion and gender (Haberman & Whitney, 2009). The "cheerleader effect" is merely a byproduct of a globally efficient strategy.

In Chapter 2 and 3 I argue that just as we rely on structure in a visual scene to resolve uncertainty in our perceptions, we similarly rely on accumulated social experiences to resolve ambiguity when making more abstract social attributions, such as inferring disposition. And just as globally useful reliance on ensemble information can result in perceptual biases, an optimal social inference process can result in two of the most well-known, and widely accepted, biases in social psychology: the fundamental attribution error (FAE) and the role-conferred advantage.

According to the fundamental attribution error, we do not properly consider the pressure of the situation, and are prone to think others' behavior is caused by their disposition (e.g. Gilbert & Malone, 1995). For instance, participants who read an essay that argues in favor of marijuana legalization, and are told that the author was assigned this position, still think that the author is personally in favor of legalization. The classic, and widely uncontested, account argues that it is unreasonable for people to attribute a behavior to disposition when they know that it was caused by the situation; consequently, studies documenting attribution to disposition in such cases have long been interpreted as evidence that people underappreciate the power of situations, and over-attribute to dispositions.

In Chapter 2 and 3 I challenge this view, and show that this extensive literature does not actually provide evidence that social reasoning is systematically flawed. In Chapter 2 I show – using classic data and three new experiments-- that the logic used by classic theorists is sound only for situations known to be deterministically strong (i.e. the person wrote the essay with a gun to their head), but crumbles in the vast majority of situations that are not deterministic (an assignment from a course instructor would not compel every person to endorse a position at odds with their beliefs). Unlike deterministic situations, which completely explain the behavior, thus entirely "explaining away" the role of disposition, *probabilistic* situations that exert some – but not complete – influence on behavior should only *partially* explain away dispositional attributions. When inferring the situational or dispositional causes of behavior in such under-constrained, probabilistic environments, inferring some role of disposition is the correct inference to make.

In Chapter 3 I investigate a phenomenon that is conceptually similar to the FAE, known as the *role-conferred advantage*: the idea that we don't properly account for the benefits that social roles confer when making attributions. In a series of five experiments

I demonstrate that observers (a) are not insensitive to the advantages conferred and disadvantages imposed by social roles, (b) show that observers are better calibrated to the social-roles used in this paradigm than has been argued, and (c) also make sensible use of other information available to them. In short: the way we reason about others' internal qualities is not irrational, but a natural consequence of optimal probabilistic inference.

Conclusion

When navigating the social world we have only ambiguous or incomplete information available to us at any given moment, and it is therefore impossible to deduce the true state of the world. Our inferential infrastructure solves these problem by taking advantage of the statistical structure present in our social world—we rely on our knowledge of how the world work *in general* to help resolve our uncertainty in an optimal way. Although our social perception can deviate from reality, this is not –at least in the cases I address-- because our social reasoning process is irrational or systematically flawed. In fact, systematic deviations from reality – perceiving faces as more attractive in crowd, or inferring internal-qualities to in a direction consistent with behavior—actually arise as a consequence of rational inference.

Reference

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological science*, *12*(2), 157-162.

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological science*, *22*(3), 384-392.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision research*, *43*(4), 393-404.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological bulletin*, *117*(1), 21.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences*, *14*(8), 357-364.

Haberman, J., & Whitney, D. (2009). Seeing the mean: ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(3), 718.

Langlois, J. H., & Roggman, L. A. (1990). Attractive faces are only average. *Psychological science*, *1*(2), 115-121.

Ross, L., & Nisbett, R. E. (2011). *The person and the situation: Perspectives of social psychology*. Pinter & Martin Publishers.

Sweeny, T. D., Haroz, S., & Whitney, D. (2013). Perceiving group behavior: Sensitive ensemble coding mechanisms for biological motion of human crowds. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(2), 329.

Sweeny, Timothy D., and David Whitney. "Perceiving crowd attention ensemble perception of a crowd's gaze." *Psychological science* (2014): 0956797614544510.

CHAPTER 1

Hierarchical encoding makes individuals in a group seem more attractive

Walker, D.E. and Vul, E.

As it appears in

Psychological Science

2013

Volume 25, Issue 1, Page 230-23

Psychological Science OnlineFirst, published on October 25, 2013 as doi:10.1177/0956797613497969

Research Report

Psychological Science
XX(X) 1–6

Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797613497969
pss.sagepub.com
SAGE

Hierarchical Encoding Makes Individuals in a Group Seem More Attractive

Drew Walker and Edward Vul
University of California, San Diego

Abstract

In the research reported here, we found evidence of the *cheerleader effect*—people seem more attractive in a group than in isolation. We propose that this effect arises via an interplay of three cognitive phenomena: (a) The visual system automatically computes ensemble representations of faces presented in a group, (b) individual members of the group are biased toward this ensemble average, and (c) average faces are attractive. Taken together, these phenomena suggest that individual faces will seem more attractive when presented in a group because they will appear more similar to the average group face, which is more attractive than group members' individual faces. We tested this hypothesis in five experiments in which subjects rated the attractiveness of faces presented either alone or in a group with the same gender. Our results were consistent with the cheerleader effect.

Keywords
visual perception, face perception

Received 1/18/13; Revision accepted 6/19/13

In the seventh episode of the fourth season of *How I Met Your Mother*, the character Barney Stinson postulates the *cheerleader effect*: that people seem more attractive in a group than when considered individually (Rashid & Fryman, 2008). As proposed, this effect is not simply that a member of the cheerleading squad, for instance, is more attractive than a person sitting alone in the bleachers (which could be due to factors such as objective attractiveness, altered demeanor, or social signaling), but rather that any given cheerleader will seem more attractive when seen as part of the squad than in isolation.

We propose that the cheerleader effect occurs at a perceptual level, arising from the interplay between ensemble coding in the visual system and properties of average faces. The visual system automatically computes summary representations of ensembles of objects, such as the average size of an array of dots (Ariely, 2001; Chong & Treisman, 2003), the average orientation of an array of gratings (Parks, Lund, Angelucci, Solomon, & Morgan, 2001), and even the average emotional expression of a group of faces (Haberman & Whitney, 2009). Not only does the summary that is formed influence observers' perception of the group as a whole, but it also biases their percepts of individual items to be more like the group average (Brady & Alvarez, 2011). Thus, we expected individual faces seen in a group to appear to be more similar to the average of the group than when seen alone. Moreover, the average of a number of faces tends to be perceived as more attractive than the individual faces it comprises (Langlois & Roggman, 1990). Thus, the bias of individual elements toward the ensemble average, when applied to faces, will yield a perception of individual faces as being more attractive than they would otherwise be perceived to be. In other words, the biasing effect of ensemble coding should produce a cheerleader effect. We tested this prediction in five experiments.

Subjects

Subjects were undergraduate students from the University of California, San Diego, and received partial course credit. There were 25 subjects in Experiment 1 (4 men, 21 women), 18 in Experiment 2 (6 men, 12 women), 20 in

Corresponding Author:
Drew Walker, University of California, San Diego, Department of Psychology, 9500 Gilman Dr., La Jolla, CA 92093-0109
E-mail: dehoffma@ucsd.edu

Experiment 3 (3 men, 17 women), 37 in Experiment 4 (13 men, 24 women), and 39 in Experiment 5 (10 men, 29 women).

Experiments 1 and 2

Subjects rated the attractiveness of female faces in Experiment 1 and male faces in Experiment 2. Faces were presented in a group photograph and in isolated portraits cropped from the group photos.

Method

For each experiment, we found 100 group photographs and cropped them to frame the faces of three people of the same gender. We then cropped each individual face to create three portrait images from each group photo. In both experiments, subjects rated the 300 unique faces twice, once in the group photo and once in an isolated portrait. Ratings were made by moving a mouse to set a marker on a continuous scale from *unattractive* to *attractive* (the rating scale and example stimuli are shown in Fig. 1). The order of images and whether a face appeared first in a group or as a portrait was random.

Fig. 1. Rating scale and example stimuli used in Experiments 1 and 3. Subjects rated the attractiveness of 300 faces twice, once in a group photo (top; the arrow indicated which face was to be rated) and once in an isolated portrait (bottom). Attractiveness was rated using a mouse to set a marker along a continuous scale. Stimuli were presented in color in the actual experiments.

On group trials, the three faces in the image were rated individually in a random order. Subjects saw the group photo for 1 s, after which an arrow appeared for 1 s below one of the faces (randomly chosen). Then the group image disappeared, and subjects made a rating. The group photo then reappeared for 1s, and the next face was cued for 1 s. This process repeated once more so that all three faces in the image were rated. On portrait trials, the cropped single-person image appeared for 2 s, disappeared, and then subjects made their rating.

Results

In our analysis, we aimed to measure the cheerleader effect, the advantage in perceived attractiveness granted a face when it is seen in a group rather than alone, while factoring out the variation in how individual subjects used our rating scale and variations in how attractive they found the different faces to be. To factor out individual differences in rating-scale use, we converted the raw rating given by a subject for each image in each condition (group and portrait) into a within-subjects z score by subtracting the mean rating and dividing the result by the standard deviation of the 600 ratings made by the subject. To factor out the effect of the attractiveness of specific faces, we then subtracted each subject's standardized rating of a face presented as a portrait from his or her standardized rating of that same face presented in a group. The resulting difference in z scores corresponded to the number of standard deviations higher that a given image was rated in a group than when isolated in a portrait. Using these difference scores, we assessed the average cheerleader-effect size (z-score difference) for each subject, as well as the average effect size across subjects (Fig. 2).

Although there was considerable between-subjects variation in effect sizes, subjects on average rated female faces in a group as being 5.5% of a standard deviation more attractive than those same faces in isolation (Experiment 1), $t(24) = 2.53$, $p = .018$. This cheerleader effect also held (with surprising consistency in effect size) for male faces: There was an average advantage of 5.6% of a standard deviation for faces in a group (Experiment 2), $t(17) = 2.52$, $p = .022$.

Experiment 3

In Experiments 1 and 2, each face in the group condition was presented uncued three times for 1 s each (a total of 3 s) and presented cued for 1 s, which suggests that any one face was on average attended for 2 s total. Thus, average time spent attending to any one face in the group condition was equivalent to the 2-s presentation in the portrait condition. However, any one trial of the group

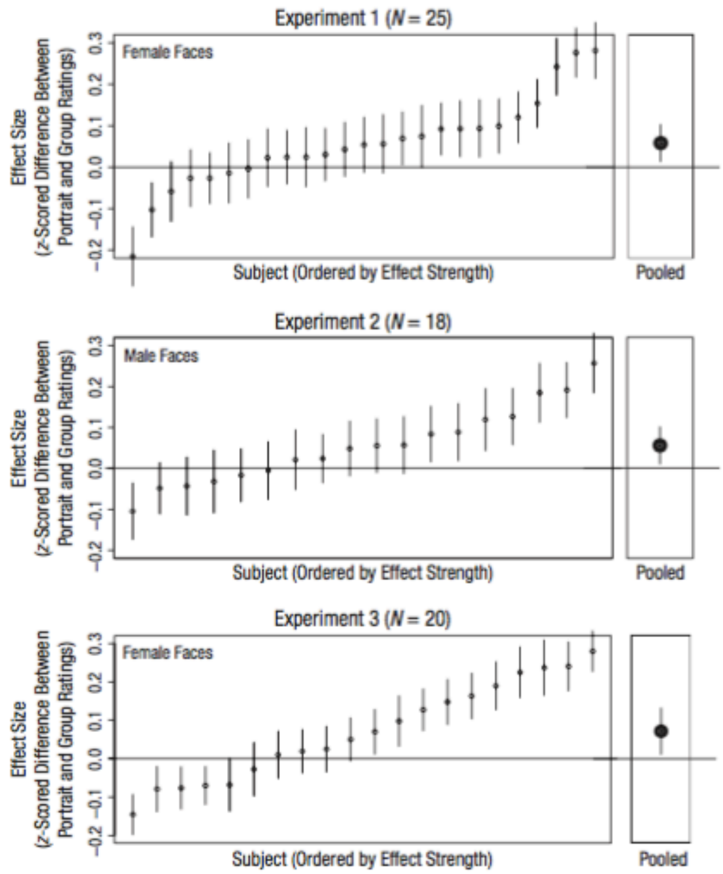


Fig. 2. Results for Experiments 1, 2, and 3: standardized size of the cheerleader effect for ratings of faces, separately for individual subjects (left panels) and pooled across subjects (right panels). To calculate effect sizes, we converted the raw rating given by a subject for each image in each condition (group and portrait) into a within-subjects *z* score by subtracting the mean rating and dividing the result by the standard deviation of the 600 ratings made by the subject. For each subject, we then subtracted this *z* score for the faces in the portrait condition from the *z* score for the faces in the group condition. This difference yielded each subject's effect size: the number of standard deviations higher that a given face was rated when seen in a group than when seen isolated in a portrait. Error bars for individual data show ±1 *SEM*, and error bars for pooled data show 95% confidence intervals.

condition consisted of an uncued group of three faces for 1 s, and a cued face for 1 s, which meant that 1.33 s was spent attending to that face. In this sense, the expected time spent attending to a face in one group trial was shorter than in a portrait trial. It is plausible that this difference drove the effect in Experiments 1 and 2 because faces shown for shorter durations are rated as more attractive than faces shown for longer durations (Willis & Todorov, 2006). Although Willis and Todorov found an advantage of shorter duration only for presentations briefer than 500 ms (and ours were all longer than 1 s), we wanted to replicate our results from Experiments 1 and 2 by equating the presentation time of one portrait trial to one presentation in the group trial. We did so in Experiment 3 by presenting the portrait images for just 1.33 s (otherwise, the design, stimuli, and method of data analysis were the same as in Experiment 1). With this modified timing, we replicated the cheerleader effect from Experiments 1 and 2: When the presentation duration of portrait images was shortened, faces were rated 6.8% of a standard deviation more attractive when presented in a group than when presented alone, $t(19) = 2.50$, $p = .022$ (Fig. 2).

Experiment 4

In Experiments 1 through 3, all of the faces had originally been photographed together in a real-life social context. Perhaps group images were rated as more attractive than single images not because of ensemble coding of the group but because the coherent context disambiguated facial expressions or other image idiosyncrasies (just as videos of an individual are rated as more flattering than the static photos that comprise them; Post, Haberman, Iwaki, & Whitney, 2012). In Experiment 4, we sought to rule out this class of explanations by presenting an array composed of multiple portrait faces that had been photographed separately. In addition to this control, we also tested for effects of group size: Increasing group size should yield a more precise average face that should not only be rated as more attractive (Langlois & Roggman, 1990) but should also exert a greater bias on the perceived attractiveness of individual faces (given a probabilistic combination of individuals and the ensemble; Brady & Alvarez, 2011).

Method

We randomly chose 77 unique faces from the stimuli used in Experiment 1. Each was presented once in each of four conditions: alone and as part of a group of 4, 9, and 16 other faces. The flanker faces in the group conditions were randomly chosen from the 223 remaining faces used in Experiment 1 (target faces were never used as flankers). Faces were presented in a square grid (1 × 1 for 1 face, 2 × 2 for 4 faces, 3 × 3 for 9, and 4 × 4 for 16; Fig. 3a). Each grid appeared for 2 s, and then a box appeared around the target face for 1 s. The faces then disappeared, and subjects made a rating as in the previous experiments. In the portrait condition, the face was presented alone in the center of the computer screen for 2 s before subjects made a rating.

Results

As in the previous experiments, we z-scored ratings within a given subject to factor out between-subjects variation in scale usage. To factor out variability in the actual attractiveness of a given face, we subtracted the average (across subjects and conditions) standardized rating given to each face from each rating of that face. This gave us a measure of the effect of each presentation condition. Figure 3b shows the average standardized ratings in each condition across all subjects. There was a significant effect of group size on attractiveness ratings, $F(3, 144) = 11.74$, $p < .001$: Consistent with a cheerleader effect, results showed that faces were rated as less attractive when presented alone than when presented in a group of 4, $t(36) = 3.23$, a group of 9, $t(36) = 4.25$, or a group of 16, $t(36) = 4.0$. However, attractiveness ratings were not different for faces rated in groups of 4, 9, or 16. These results suggest that it is not the coherent context of group photos but rather the presence of additional faces that drives the cheerleader effect.

Experiment 5

The influence of group membership on individual members may be greater when there is more uncertainty about the individual elements in a scene; this is because the average is less sensitive to the increased uncertainty than the individual elements are. Following this logic, we blurred the faces in Experiment 5 to see whether the cheerleader effect would be increased when uncertainty was increased.

Method

We randomly selected 50 group images from those used in Experiment 1 and blurred them by convolving them with a Gaussian filter with a standard deviation of 4 pixels. Subjects rated the three faces in each of those 50 images (150 unique faces) four times each: in unblurred group and portrait conditions and in blurred group and portrait conditions (example stimuli are shown in Fig. 4a). Other than the addition of the blurring factor, methods were identical to those used in Experiment 1.

Results

As in Experiment 4, we isolated the effect of condition by z-scoring ratings within subjects and subtracting the across-subjects average ratings for each face. Figure 4b shows the average standardized ratings in each condition. As in our other experiments, faces were rated as more attractive when seen in groups than when seen alone, $F(1, 152) = 9.0$, $p < .01$, and subjects rated blurred images as more attractive than unblurred images, $F(1, 152) = 17.91$, $p < .001$. However, although the cheerleader effect was bigger in the blurred condition than in the unblurred condition (7.3% vs. 5.9% of a standard deviation), the interaction between image clarity and

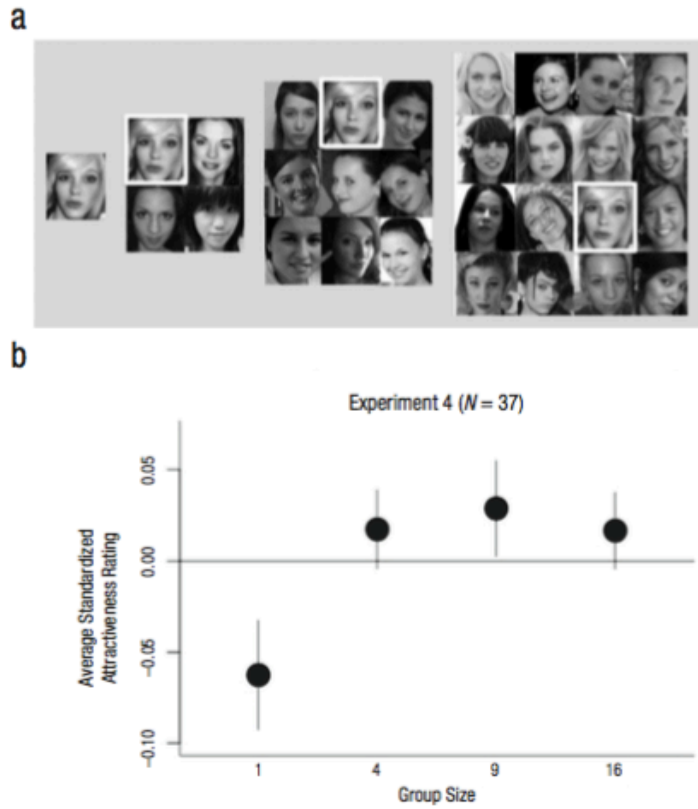


Fig. 3. Example stimuli (a) and results (b) from Experiment 4. Subjects rated the attractiveness of 77 faces four times: alone and in a group of 4, 9, and 16 other faces. Each group appeared for 2 s, and then a box appeared around the target face for 1 s. In the portrait condition, the face was presented alone in the center of the computer screen for 2 s. The faces then disappeared, and subjects made a rating as in the previous experiments. Stimuli were presented in color in the actual experiment. The graph shows the average standardized attractiveness ratings for each group size. To calculate attractiveness ratings, we first obtained within-subjects z scores as in the previous experiments. For each subject, we then subtracted the average (across subjects and conditions) z score given to each face from each rating of that face. Error bars show 95% confidence intervals.

presentation condition was not significant $F(1, 152) = 0.106$, $p = .75$.

General Discussion

In the five experiments reported here, we found evidence consistent with the cheerleader effect: Both female faces (Experiment 1) and male faces (Experiment 2) in a group appeared more attractive than those same faces seen alone.[1] This effect seems robust to presentation timing (Experiment 3), to whether groups are created from natural photos or are synthetically created (Experiment 4), and to image manipulations such as blurring (Experiment 5). We propose that this effect arises from the fact that the visual system represents objects as an ensemble (Ariely, 2001), individual objects are biased toward the ensemble average (Brady & Alvarez, 2011), and average faces are perceived to be more attractive than faces in isolation (Langlois & Roggman, 1990). Together, these phenomena should cause faces in a group to appear more like the

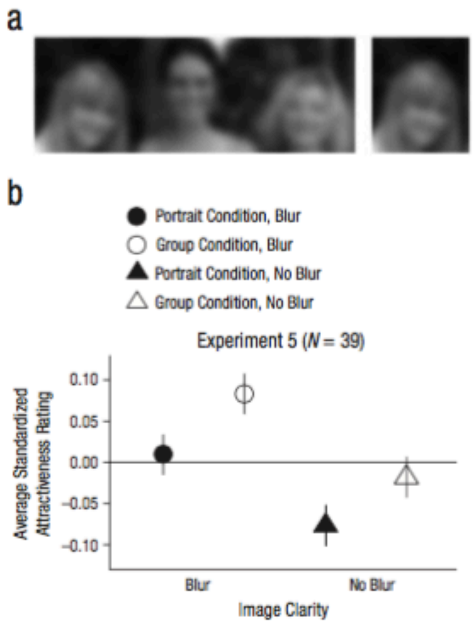


Fig. 4. Example stimuli (a) and results (b) from Experiment 5. Subjects rated the attractiveness of 50 faces four times each: in blurred group and portrait conditions (shown here) and in unblurred group and portrait conditions. Stimuli were presented and ratings were made as in Experiment 1. Stimuli were presented in color in the actual experiment. The graph shows the average standardized attractiveness ratings as a function of image clarity and presentation condition (group or portrait). Attractiveness ratings were calculated as in Experiment 4. Error bars show 95% confidence intervals.

group average than when presented alone, and that group average should tend to be more attractive than the individual faces, on average. However, some of our results should give readers pause in accepting our interpretation: We predicted that increasing group size (Experiment 4) or decreasing image quality (Experiment 5) should increase the bias of individuals to the group average and would thus increase the cheerleader effect, but we found no evidence of these effects. Despite this caveat about our interpretation, the cheerleader effect was robust: Across a wide range of settings, people in groups were rated as more attractive than the same people alone. Thus, having a few wingmen—or wingwomen—may indeed be a good dating strategy, particularly if their facial features complement, and average out, one's unattractive idiosyncrasies.

Author Contributions

Both authors contributed to the design of the experiments. The experiment was programmed by D. Walker. E. Vul and D. Walker analyzed and interpreted the data. D. Walker and E. Vul drafted the manuscript. Both authors approved the final version of the manuscript for submission.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Note

1. Critically, the same face when seen in a group of different faces is rated as more attractive than when seen alone. However, Post et al. (2012) found no such effect for a face presented in an arrays of the same face.

References

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12*, 157–162.

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science, 22*, 384–392.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43*, 393–404.

Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance, 35*, 718–734.

Langlois, J. H., & Roggman, L. A. (1990). Attractive faces are only average. *Psychological Science, 1*, 115–121.

Parks, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience, 4*, 739–744.

Post, R. B., Haberman, J., Iwaki, L., & Whitney, D. (2012). The frozen face effect: Why static photographs may not do you justice. *Frontiers in Psychology, 3*, 22. Retrieved from http://www.frontiersin.org/Journal/10.3389/syg.2012.00022/full

Rashid, R. (Writer), & Fryman, P. (Director). (2008). Not a father's day [Television series episode]. In C. Bays & C. Thomas (Creators), *How I met your mother*. New York, NY: CBS.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17*, 592–598.

Chapter 1, in full, is a reprint of the material as it appears in *Psychological Science*. Walker, Drew E.; Vul, Ed. 2013. The dissertation author was the primary investigator and author of this paper.

CHAPTER 2

The fundamental attribution error is reasonable in an uncertain world

Walker, D.E., Smith, K.A., and Vul, E.

Manuscript submitted for publication

ABSTRACT

We do not directly observe the internal qualities of others, such as their attitudes and dispositions, 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 commit the Fundamental Attribution Error: we are systematically biased to overestimate the influence of disposition on behavior. We propose that the social judgments made in classic studies of attribution have been interpreted as biased only because they have been compared to an inappropriate benchmark predicated on the assumption of deterministic dispositions and situations. We review six classic results and present empirical data that demonstrates that social inferences are consistent with unbiased probabilistic attribution of the influence of situations and dispositions in an uncertain world

INTRODUCTION

Imagine you're visiting a donation only museum. You see a fellow patron drop $5 into the "pay what you can" donation jar. Do you conclude that she is generous? This inference would likely be influenced by how much pressure there was to make a donation. For example, whether or not there was a museum docent monitoring the donation jar. Since internal qualities such as "generosity" cannot be directly observed, we must infer them from others' behavior. But behavior is influenced not only by these internal qualities, but also by external circumstances, such whether or not a docent is observing your donation. Thus, attributing a behavior to internal qualities or external situations is an underdetermined problem.

An extensive literature suggests that there are systematic discrepancies between the inferences about internal qualities that people *should* make, and the inferences they *do* make. A considerable number of behavioral experiments using rich social situations have concluded that we have a tendency to attribute a person's behavior disproportionately to their disposition, disregarding the influence of known situational pressure. That is, when we witness someone make a donation we are prone to think that she is a generous, and not properly consider the pressure imposed by a watchful docent. *The Fundamental Attribution Error*[1] *(FAE)* is one of the most famous concepts in social psychology, has spawned numerous theoretical explanations (for review see Gilbert & Malone, 1995), is featured prominently in introductory psychology texts, and is referenced often in popular culture (e.g. Gladwell, 2000). In the classic demonstration of the FAE, university students read an essay, ostensibly written by a classmate, which either favored or opposed Fidel Castro (Jones & Harris, 1967). Even when told that the writers were assigned their position by a course instructor, readers still thought that the author actually held the view expressed in the essay. A large number of studies have since produced similar results, yielding a net assessment in the literature that people are "lay dispositionalists" (Ross & Nisbett, 1991) – wired to neglect situational pressure and attribute actions to stable internal qualities. The inferences observers make in such experiments are typically considered to be logically unwarranted. For example, the most comprehensive review of the FAE states:

---

[1]The Fundamental Attribution Error (FAE) is often used interchangeably with the Correspondence Bias (CB) in the literature, but some authors have distinguished the FAE as the tendency to underappreciate the influence of situation, which is then a *cause* of the CB (e.g., Gawronski, 2004), the tendency to over-attribute behavior to disposition.

> *"..when people observe behavior, they often conclude that the person who performed the behavior was predisposed to do so—that the person's behavior corresponds to the person's unique dispositions—and they draw such conclusions even when a logical analysis suggests they should not." (Gilbert & Malone, 1995)*

The classic normative models reason that when you witness an outcome, and then learn that it was caused by one event, it is inappropriate to use that outcome as evidence for another potential causal event (e.g. Kelly, 1973; Jones & Davis, 1965). For example, if you flip on a light switch and the light doesn't turn on there are two likely explanations: the light-bulb is burnt out, or there is no electricity. If you then learn that the power is out across the city, there is no reason to believe that there is anything wrong with the bulb. According to this deterministic logic, if you read a pro-Castro essay and learn that the author had been assigned by an instructor to write in support of Castro, it is illogical to also take the essay as evidence that the author favors Castro.

Most isolated social situations, however, are not so powerful that they completely constrain behavior, like an electrical outlet carrying no power. In daily life people rarely encounter situations that are so extreme that everyone acts uniformly within them. 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 *totally* deterministic (people still escape from jail). Outside of such extremes, situational pressures are far from deterministic, but rather interact with internal qualities to produce behavior: some people would not make an optional donation even when a docent were watching, while others would make donation even without a witness. The situations akin to those used in FAE tasks are also far from deterministic; for example, when Sherman (1980) asked university students to write an essay supporting a controversial school policy, less than 70 percent 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. But how should we reason about peoples' dispositions when actions only partially constrain behavior? For this we turn to the formalism of causal attribution in probabilistic inference.

Recent work suggests that human causal learning and inference can be explained within a framework of probabilistic inference in Bayesian networks (Pearl 1988; Griffiths & Tenenbaum, 2005; 2009) On this account, we form a causal model of the world, and condition on our observations to infer what might have been true of the world to yield the outcomes we observed. This reasoning framework can account for a wide range of the causal inductions made by humans (Holyoak & Cheng, 2011), including social inferences, such as inferring goals from the movement of simple animated agents (e.g. Baker, Saxe, & Tenenbaum, 2009). We propose that the social inferences made in classic FAE studies can also be accounted for within this framework, and that given uncertainty about the strength of situational variables, human tendencies to infer internal qualities in these paradigms do not reflect a bias, but are actually quite sensible.

In the past there have been scattered proposals that the social inference process could be accounted for in terms of probabilistic inference (Azjen & Fishbein, 1975; Morris & Larrick, 1995). However, as these proposals predated modern computational statistical methods (Griffiths & Tenenbaum 2005, 2009; Kemp & Tenenbaum 2009), they could offer only verbal descriptions and conceptual possibilities[2] without being able to make concrete predictions about the variety of manipulations in the decades of established FAE literature.

In this paper we first explain how an unbiased agent should make inferences about dispositions in non-constraining situations using Bayesian inference. We then show that this framework yields six classic results in the FAE literature[3]: attributing to disposition in the presence of explanatory situational pressures (Jones & Harris; 1967), the influence of prior beliefs on dispositional attribution (Jones et al; 1971), modulation of attitude attribution by action intensity (including attribution of opposite attitudes from weak actions; Jones et al; 1971), interpreting ambiguous behavior (and disposition) in light of the situation (Snyder & Frankel, 1976), and attributing to situations in the presence of explanatory dispositional pressures (Quattrone; 1982). Our results show that the patterns of behavior interpreted as systematic flaws in our social inferences instead reflect the best possible inferences that observers can make given their uncertainty about other people and the situation.

Probabilistic Social Attribution

Given the uncertainty inherent in reasoning about the causes of others' behavior, we cannot expect people to make errorless social inferences. Consequently, the relevant question for assessing whether we make such inferences with systematic flaws is not

[2]However, see Jennings (2010) for an alternative quantitative Bayesian conceptualization of the FAE and the our Discussion for a comparison of his framework with the account we propose here.

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 under the same information. How, for instance, would an unbiased observer infer the generosity of a museum patron who makes a donation 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 three-node graphical model (Figure 2.1; 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).

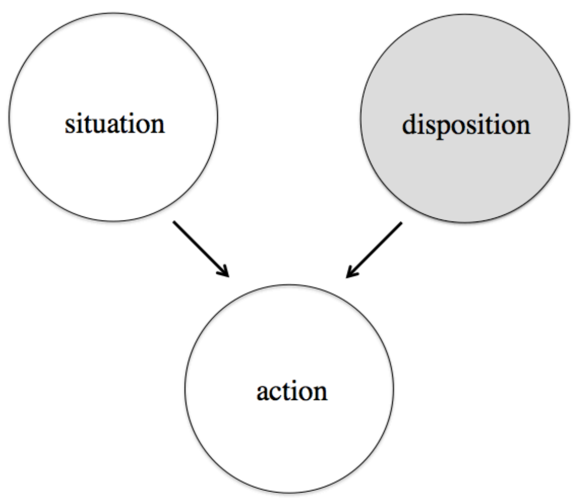


Figure 2.1. Graphical model of an action arising from the combination of two classes of causes: Situation and disposition both influence the probability that an action will occur. Various attribution experiments amount to conditioning on (observing) two of the three nodes (usually situation and action), and inquiring about the third (disposition).

In this scenario, we treat behavior as a simple, binary action: the museum patron either leaves a donation, or doesn't. We can express the binary action *a* (donate or not) as a draw from a Bernoulli distribution with some latent probability of donating (*q*). Situation (*s*) and disposition (*d*) will both influence this probability; for simplicity and convention, we will assume that their influence on *q* is additive in log-odds. Thus, situation influence (*s*) and disposition 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: $log[q/(1-q)]=s+d$. Therefore, the probability of donating can be obtained via the logistic transformation of the sum of the situational and dispositional influences:

$$q = \frac{1}{1 + e^{[-(s+d)]}}$$

Under this model, the probability of making a donation is *q*, and the probability of not donating is 1-*q*:

$$p(a \mid q, \theta) = \begin{cases} 1 & q \\ 0 & 1-q \end{cases}$$

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

This formulation 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(q/(1-q)) = 1; q = 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 non-constraining 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 (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 $s=1.5$ (log(0.92/(1-0.92)) = s+1; s=1.5: or equivalently,).

Although in most FAE 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 (*q*) 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 actor 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, we link action propensity (*q)* 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 (q) to action (a) as a Beta distribution:

$$P(a \mid q, \theta) = \beta(a \mid q\theta, (1-q)\theta)$$

where θ is the concentration parameter, indicating dispersion around the central tendency of. This formulation yields action strengths ranging from strongly negative (0), through neutral (.5) to very positive (1). 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)* – e.g., how many people are more generous or stingy than average? Given this prior distribution on dispositions, and the observation of an action *a* in a situation of some strength *s*, we can calculate the posterior probability of the generosity of the actor using Bayes rule:

$$P(d \mid a,s) = \frac{P(a \mid d,s)P(d)}{\int_{d'} P(a \mid d',s)P(d')}$$

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 2.2 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 a we still infer more generosity when we observed a donation compared to the prior. It’s only when the pressure to donate is so strong that nearly no one would refuse that the inferred generosity becomes indistinguishable from the prior.

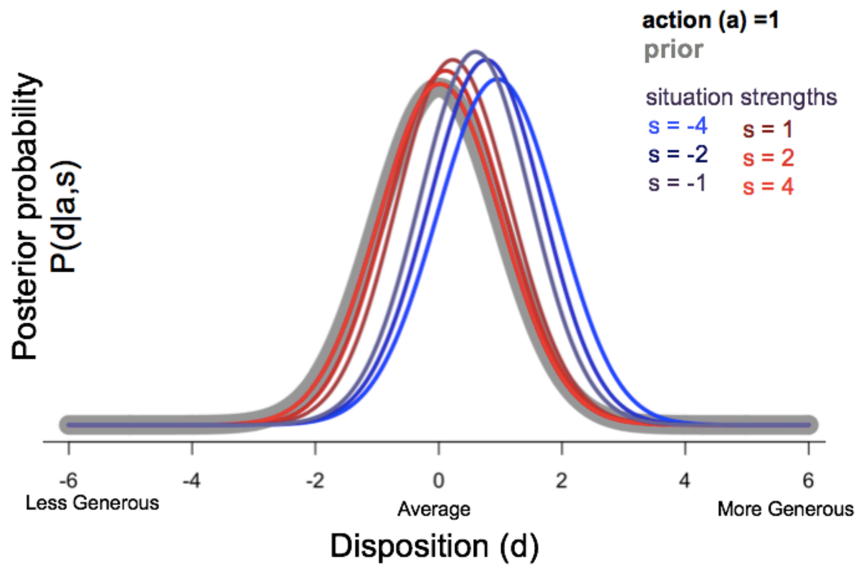


Figure 2.2. The posterior distribution over disposition after observing a positive action (e.g., donating; a=1), under different assumed situation strengths. The grey distribution represents the prior distribution of generosity (the generosity that would be inferred if nothing else was known about the situation or action). When the situation strongly discourages a donation (e.g. s = -4, bright green) 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 everyone would act this way regardless of their disposition. Thus, as situation becomes extremely strong, implying that it alone deterministically caused the action, probabilistic attribution will yield results consistent with deterministic, 'logical' attribution: no inferences about disposition will be made when situations force the observed action.

Let's imagine we are now in a world with somewhat more stingy museum patrons than before, where only 50% people would donate when no docent is watching (here represented by the prior). Now you observe someone leave a donation (*s=0*). Based on the equation above, you should infer that the visitor is somewhat more generous than average ($E[d|s=0,a=1]=0.39$, Figure 2.3 point A).

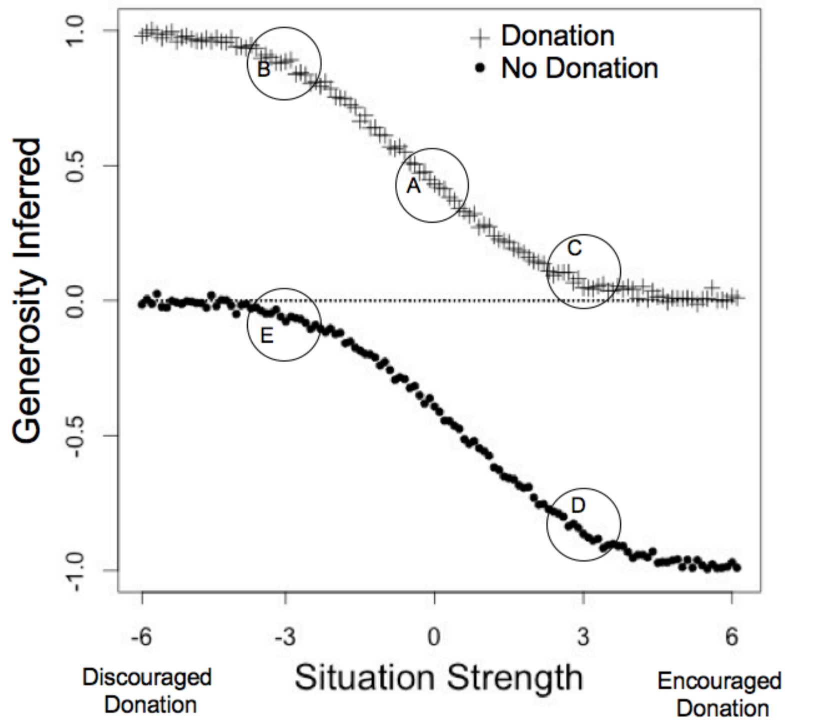


Figure 2.3. The strength and direction of the inferred disposition by a probabilistic social attribution model depends on the situational pressure in combination with whether the action occurred. For example, if a donation occurred (+) when there was a lot of pressure to donate, the ideal observer makes weaker inferences about the actors' generosity (point C). 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 (•) when there is strong pressure to donate (point D) suggests the person is far below average on this trait, compared to when the person doesn't donate but there was also strong pressure against donating (Point E). The ideal observer will always infer a disposition consistent with the observed action (though this will become vanishingly small as situations become nearly deterministic; more extreme than points C and E, respectively).

But what if there is strong pressure against donating? For instance, if the donation jar is missing (s=-3) yet a the person leaves a donation anyway, you should infer even more strongly that they are generous ($E[d|s=-3,a=1]=.75$, Figure 2.3 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, for example the docent says to the person "the museum will likely be closing if patrons refuse to donate" ($s=3$), the unbiased observer will still infer something about the patron's disposition ($E[d|s=3,a=1]=.08$, Figure 2.3 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.

Applying Probabilistic Attribution to Classic FAE Results

For the remainder of this paper we ask how well the probabilistic attribution framework captures human inferences in classic social attribution experiments from the *Fundamental Attribution Error* literature. We limit our discussion to the classic studies that have been directly interpreted as evidence for a FAE.[4] We first consider Jones and

[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 FAE 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 & Steinmetz, 1977) which requires accounting for the behavior of multiple actors simultaneously, and is thus beyond the scope of this instantiation of our model.

Harris' (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 FAE literature is so extensive and often uses this essay task with modifications that are tangential to testing predictions of probabilistic social attribution (e.g. Miller, Jones & Hinkle, 1981), we only focus on versions of the classic essay task that are modified in ways that allow us the opportunity to test novel predictions beyond what could be demonstrated with this classic study.

The first modification of the essay task that we address is a slightly more complex scenario in which people have pre-existing beliefs about the author (Jones et al, 1971). Next, we examine a study wherein the strength of the argument presented in the essay is manipulated (Jones et al, 1971), and then explore a puzzling result for the FAE hypothesis: what do people do if the classic paradigm is inverted -- when they are asked to infer the strength of the situation after reading an essay written by an author with a known disposition (Quattrone, 1982). Finally, we consider a more subtle experiment than the essay paradigm in which behavior is held constant, but observers have different beliefs about the situation (Snyder & Frankel, 1976).

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 have scaled the posterior beliefs of the ideal observer to the scales used in the respective empirical studies.[5]

---

[5]For example, if the original experiment asked for attitude reports on a 1-6 Likert scale the ideal observer's judgments would be converted in this way: $1+5/(1+\exp(-E[d|a,s]))$. Furthermore, even though we presume that human intuitions are probabilistic, we report

Inferred attitude when action is encouraged by the situation (Jones & Harris, 1967)

In the classic FAE experiment, Jones and Harris (1967) examined how people account for situational pressures when reasoning about others' dispositions. Based on a deterministic view of such inferences, they reasoned that a behavior is evidence of a person's disposition, but when there is a situational explanation for the behavior it should longer reveal anything about the actor's disposition. They asked university students to read an essay that either opposed or endorsed Castro. Participants were told that the essay was written by a classmate who was either instructed to argue for a particular position, or was free to choose whether to write a pro or a con essay. After reading the essay participants answered ten 7-point Likert scale questions (1: strongly anti to 7: strongly pro) about what they thought the author's true attitude toward Castro had been; these ten responses were summed, yielding an overall scale from the strongest anti-Castro beliefs (10) to the strongest pro-Castro beliefs (70). If the essay position were freely chosen, then it obviously reveals the authors' attitude; however, if instructions to write in support of one position or another would make any person — regardless of their disposition — produce a compelling essay for the instructed position, then Jones and Harris suggest that the essay content should not be informative of the authors' attitude.

---

point-estimates from the ideal observer by averaging over the posterior distribution in order to mirror the point-estimates that human observers were asked to make.

As predicted, when readers were told the essay position was freely chosen they believed that the author had an attitude about Castro consistent with the views expressed in the essay. However, when the reader was told that the position had been assigned, readers continued to estimate the authors true attitude to be consistent with the essay's position, albeit more weakly (original data re-plotted in Figure 2.4, panel A). According to Jones and Harris this suggests that people behave illogically: people infer something about the writer's attitude when they should explain behavior based on the situation.

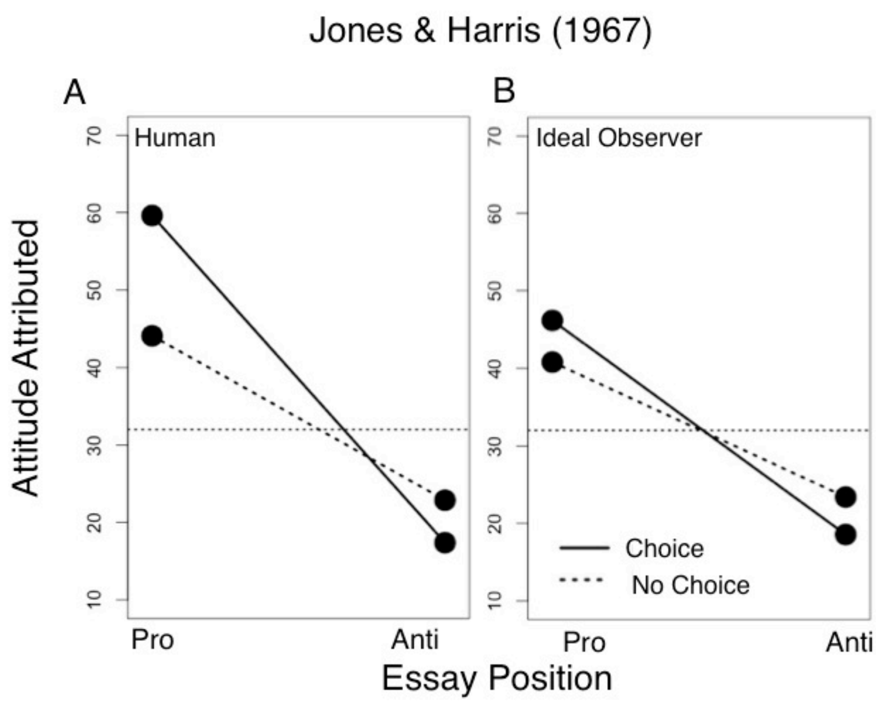


Figure 2.4. Inferred attitude as a function of essay position, and whether this position was chosen or assigned. *A:* People inferred that the position expressed in an essay was indicative of the author's true attitude both when they chose their position, and when it was assigned (Jones & Harris, 1967). The attitude attributed to the author was stronger when the author chose his position (solid line), and weaker when assigned (dotted line). *B:* An ideal observer also infers that the essay is indicative of the author's true attitude, but more informative when the position was chosen, and less informative when assigned (dotted line).

But what inferences should we expect from an unbiased observer who *did not* believe that instructions to write a particular essay are completely deterministic? We can characterize the behavior of such an unbiased observer via a probabilistic social attribution model: 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 2.3, 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.

So, an unbiased observer believing in non-deterministic influence of situations will still infer some disposition, but just how much depends on the observers' assumptions about how compelling the situation is. To formalize this, we must specify the "situation strength" of being offered a free choice about which position to take in an essay, that of being assigned to write an essay taking a particular position, as well as the prior distribution about Castro attitudes. The majority of subjects (readers) reported being anti-Castro (Mean=32.5, Variance=35.4 on a scale of 10-70); we used these summary statistics to set the prior distribution of attitudes about Castro ($d \sim N(-.53, 0.44)$[6]; changing this distribution yields a roughly uniform shift in inferred attitude across all four

[6] $d \sim N(-0.53, 0.44)$ yields the corresponding mean and variance when transformed to the 10-70 scale: pro-Castro attitude $x = 10 + 60/(1 + \exp[-d])$.

conditions). 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 have situation strengths that would compel a perfectly neutral person to write the assigned essay 80% of the time (s=1.38 and -1.38, respectively). Finally, we obtained "likert-like" ratings from the posterior belief about disposition by scaling the logistically transformed expected disposition to the 10-70 scale used in the paper.[7]

Under these assumptions, an ideal observer infers the same pattern of dispositions as people do: when the situation does not exert any pressure (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. When the situation does exert pressure to take a particular position (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 2.4, panel B).

It's critical to note that the *qualitative* pattern of results (weaker, but non-zero, attribution of attitudes in the no-choice condition) holds regardless of the specific assumptions we make about prior distributions over d, or situation strength in the no choice condition.

The influence of preconceptions on inferred attitude (Jones et al., 1971)

The probabilistic social attribution model yielded the critical pattern of inferred attitudes observed in the original FAE experiment (Jones & Harris, 1967) under the

[7]Reported pro-Castro attitude = 10+60/(1+exp(-E[d|a,s])).

assumption of an expectation that on average people disfavor Castro. This was sufficient because participants in that experiment were only told that a classmate wrote the essay. However, real-life social situations typically contain much more context that we use to flavor our inferences about other people. For example, in the USA, if you meet someone at a National Rifle Association rally, that person is more likely to be politically conservative, whereas if you meet someone at a Vegan potluck, that person is probably politically liberal. The background context should therefore affect how people attribute internal qualities to others.

Jones et al (1971) investigated how prior expectation and action intensity affect attribution of beliefs about the legalization of marijuana. Subjects first read a questionnaire that the essay author ostensibly filled out. The questionnaire responses were designed to alter subjects' expectations about how conservative the author was, and thus how likely they are to support, or oppose, the legalization of marijuana (e.g., an author who favors strict abortion laws is likely to be conservative and to oppose legalization while an author who opposes strict abortion laws is likely to be liberal and support legalization). They then read an essay that either favored or opposed legalization of marijuana, and were told that the essay position was either freely chosen, or assigned). Then subjects estimated the author's attitude about marijuana 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 someone who was portrayed as 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 had been assigned the position or chose it freely. 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: readers inferred an attitude more consistent with the position in the essay when an author ostensibly chose his position, than when the position had been assigned (Figure 2.5, panel A).

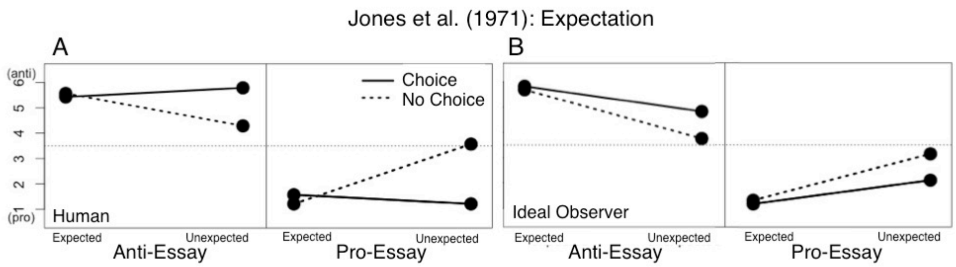


Figure 2.5. When the essay position was consistent with expected attitude, people (top) attributed the corresponding, consistent attitude to the author, regardless of whether the essay position was assigned, or chosen by the author. However, when the essay position deviated from the expected attitude, people 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 (bottom) 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 are no extenuating situational pressures.

What inferences should we expect from probabilistic social attribution under the assumptions that (1) the instructions to take a certain essay position are not deterministic, and (2) 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 2.3), 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 most reason to change our beliefs about authors' disposition, and this change in beliefs will be weaker in the no-choice condition – when 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 consistency, we retain the same situation strength as in the previous scenario: $s=1.38$ for instructions to write an anti-legalization essay and $s= -1.38$ 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 $d\sim N(1.38,.1)$, and an ostensibly liberal author is assumed to have a disposition $d\sim N(-1.38, .1)$; 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.

The influence of action intensity on inferred attitude (Jones et al., 1971)

Organizing an NRA rally is a more extreme action with regard to Second Amendment beliefs than having an NRA membership, or merely owning a hunting rifle. In general, since real-world behaviors are not easily classified as dichotomous (donate / do not donate), but rather fall along a continuum of extremeness, 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 of the essay (pro- or anti- marijuana legalization), and the extremeness: strongly anti-legalization, weakly anti-legalization, weakly pro-legalization, and strongly pro-legalization. Each subject read one essay, expressing a position that they believed had been either chosen or assigned. Based on this information, they estimated the author's attitude about marijuana legalization on a 6 point Likert scale.

When the essay position was purportedly freely chosen, subjects thought that the author's attitude scaled with the expressed position: inferred legalization attitude changed monotonically from a strong anti essay, to weak anti essay, to weak pro essay, to strong pro essay. However, when the essay position (pro or anti) had ostensibly 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 Marijuana legalization, and vice versa (Figure 2.6, panel A). 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 isn't the letter itself, but the strength of the letter that provides us with the most information. A recommendation that praises penmanship and punctuality, though technically positive, actually seems to suggest a more negative opinion on account of how much more positive it could be.

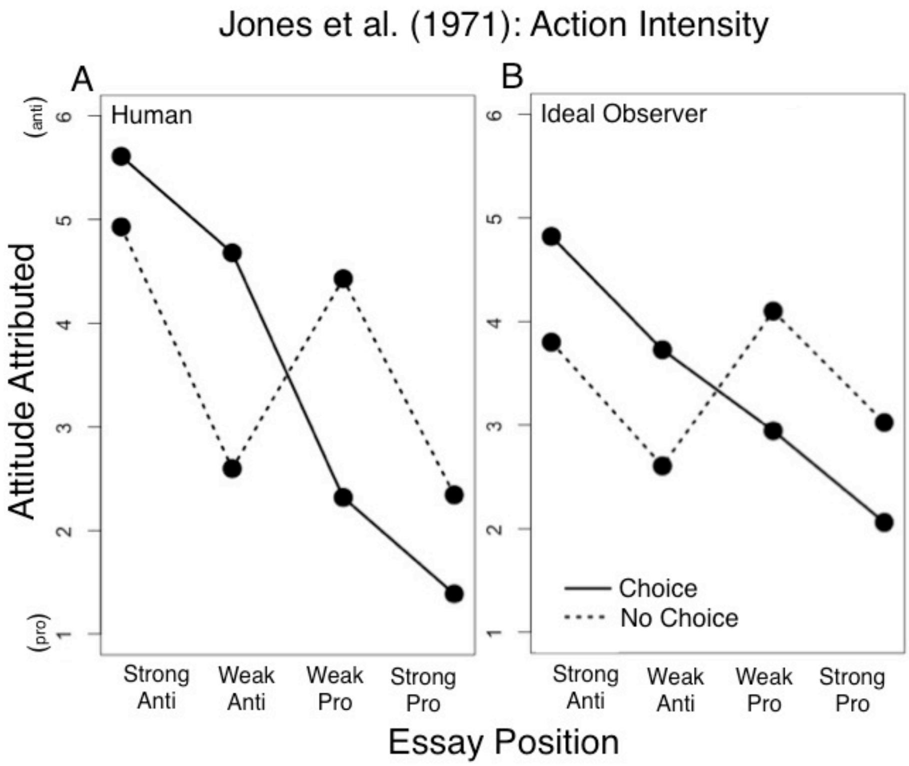


Figure 2.6. Both human subjects (left) and the probabilistic attribution model (right) infer that a strong essay reflects the author's attitude (and more so for a freely chosen essay). However, a weak essay only indicates a consistent attitude if the essay position was freely chosen; a weak essay taking a position assigned by the experimenter indicates that the author is actually likely to oppose the essay position.

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~N(0,1)).

Since in this experiment the actions were explicitly non-binary, we switched to the Beta-likelihood function (eq. 3), which relates action propensity to a continuous scale from 0 to 1 (θ was, somewhat arbitrarily, set to 10). Accordingly, we rescaled the 4 essay strength ratings from the 1-10 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.

Figure 2.6, panel B shows that similar to the inferences people make, when the situation exerts no pressure, probabilistic attribution infers authors' attitudes that scale with expressed essay strength. However, when the essay position was externally motivated, a weak essay is taken as evidence that the authors true attitude is actually opposed to the expressed position: a situation that encourages a pro-Legalization essay would yield an 80^{th} percentile essay strength from a neutral person; and seeing a 60^{th} percentile essay instead implies that the author's disposition pushed the author in the direction opposite from the situational pressure.

Inverting the FAE: Inferring situation when attitude is known (Quattrone, 1982)

The FAE hypothesis posits that people overestimate the influence of disposition and underestimate the influence of situations. Thus, under the FAE 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 "over-attribute" the actor's action to situational pressures. This result is symmetric with the over-attribution to disposition observed in classical FAE experiments, but is completely inconsistent with the conventional interpretation of those results.

Quattrone (1982) asked subjects to read an essay favoring or opposing the legalization of marijuana, but instead of telling readers that the essay position was chosen or assigned, he told them that the author was known to have either a neutral opinion about legalization, or an opinion consistent with the attitude expressed in their essay. Subjects were told that the purpose of the study was to determine if extraneous experimental factors (e.g. experimenter bias) might be influencing the opinions people expressed. After reading the essay, subjects were asked to estimate the likely situational pressure on a 30-point Likert scale (-15: pressure to oppose, 15: pressure to favor). Thus, just as the classic FAE paradigm manipulates situation strength and measures inferred disposition, Quattrone manipulated disposition strength and measured inferred situational influence. 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 2.7, panel A).

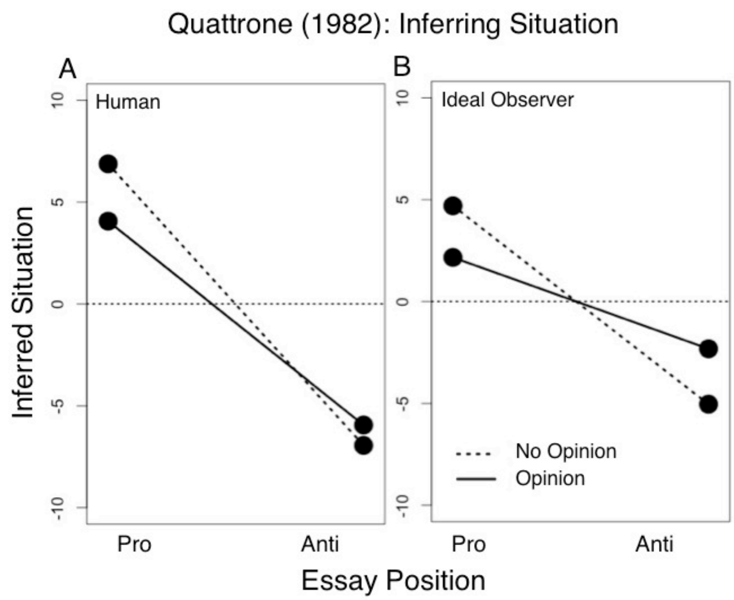


Figure 2.7. Inferred situation as a function of the essay and the known attitude. *A:* Subjects inferred that the position expressed in an essay was evidence for how much the situation was motivating behavior, both when they thought the author had a pre-existing attitude, and when the didn't. The situation inferred was stronger when they thought the author had no pre-existing opinion. *B:* The ideal observer also infers that the situation was pressuring the essay position, but more so when the author had no existing opinion.

According to the logic of the FAE account, this pattern of results could be considered an "over-attribution" to situational factors, since the pre-experiment attitude is known to have caused behavior. This finding is inconsistent with classic explanation of the FAE and calls into question the theoretical accounts of the FAE that claim that we have a inclination to over-attribute behavior to dispositions, and not attribute enough to situations (e.g. Taylor & Fiske, 1978; Gilbert, Pelham & Krull, 1988).

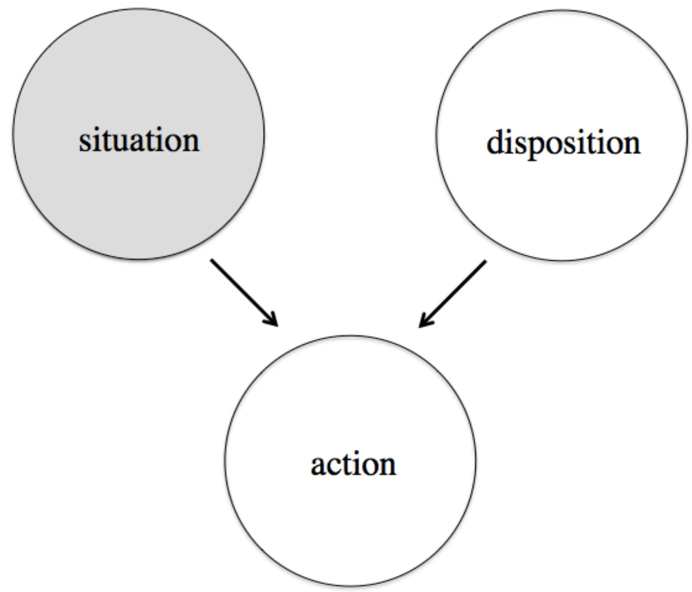


Figure 2.8. Graphical model show that situation and disposition influence the probability that an action will occur. Instead of conditioning on (observing) situation and the action, and inquiring about situation, here we condition on disposition and action, and inquire about the strength of the situation.

A probabilistic attribution account, however, 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 2.1), inferring the unknown situation strength given a known disposition is symmetric inferring the disposition given a known situation (Figure 2.8). 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:

$$P(s \mid a,d) = \frac{P(a \mid s,d)P(s)}{\sum_{s'} P(a \mid s',d)P(s')}$$

.

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 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- Marijuana essay) but now knows the author's legalization 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 FAE account (Figure 2.8). 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 FAE effect, as well as the inverse FAE effect, where human behavior is directly opposite to the predictions of the FAE hypothesis.

Inferring disposition when actions are ambiguous (Snyder and Frankel, 1976).

In the classic FAE paradigm readers are asked to infer an essay author's disposition while the experimenter varies the essay. Snyder and Frankel (1976) worried that the FAE might arise because the behavior (i.e. an essay on a controversial topic) might be disproportionately salient compared to the situation (i.e., whether or not the essay was assigned), and thus held the observed behavior constant while manipulating observers' beliefs about the situation. Observers watched a silent video of a woman answering some questions. One group of subjects was told the conversation topic was either sex (a higher anxiety situation) or politics (a lower anxiety situation) before watching the video, and the other subjects learned the topic only after seeing the video. Observers then completed two sequential questionnaires: first, how anxiously the woman behaved *in the video*; second, how anxious was this woman in general (dispositional anxiety). Snyder and Frankel found that altering observers' beliefs about the situation changed how the same behavior was interpreted, and the subsequent dispositional attributions were comparable to the original FAE paradigm: observers who perceived anxious behavior inferred a more anxious disposition, and vice versa. We demonstrate that probabilistic attribution explains why belief about situations influences perceptions of behavior, and as a consequence, subsequent questions to generalize from the perceived behavior will yield corresponding dispositional inferences.

Displayed Anxiety

Immediately after watching the interview observers judged how comfortable/uncomfortable and how tranquil/upset they thought the woman *behaved during the interview*. They answered both question on a 1 (not at all anxious) to 9 (extremely anxious) scale, yielding a summed "perceived anxiety" score between 2 and 18 (by summing the two responses). Observers who knew ahead of time that the topic was sex interpreted the behavior as more anxious than observers who thought she was discussing politics. However, observers who found out the topic only *after* watching the video estimated anxiety to be the same, regardless of the topic (Figure 2.9, panel A).

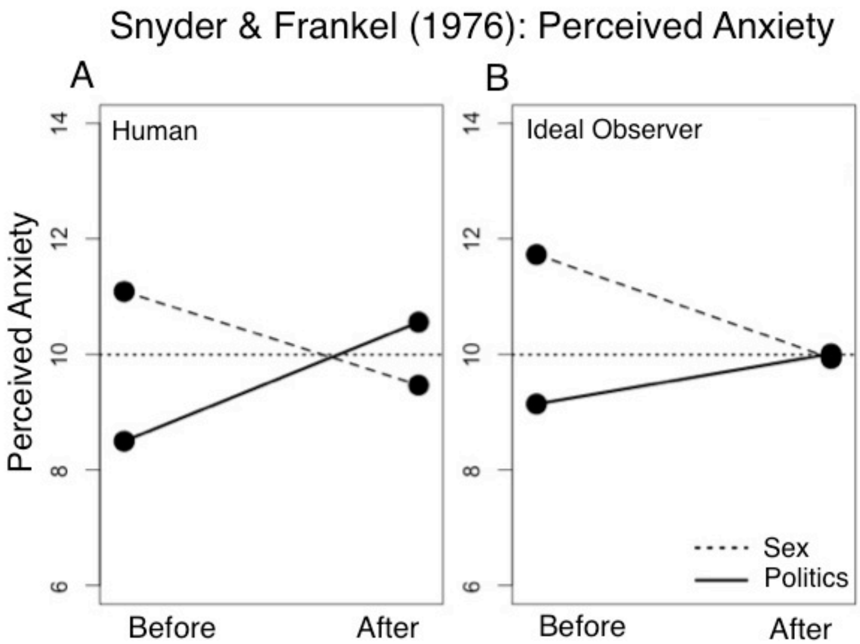


Figure 2.9. Perceived anxiety as a function of interview topic, and whether this was known before or after watching the video. *A:* Observers who had a belief about the topic when they watched the video interpreted the woman's behavior to be more anxious than average when they believed she was talking about sex (dotted line), but less anxious than average when they thought she was discussing politics (solid line). Observers who were unaware of the topic while watching the video thought she displayed neutral anxiety. *B:* The ideal observer infers that an ambiguous behavior was more likely to have arisen from anxiety when the topic is thought to encourage anxiety (sex; dotted line), compared to an ambiguous behavior produced in a situation that discourages anxiety (politics; solid line). When nothing is known about the situation, ambiguous behavior is no more or less likely to have arisen from anxiety. In this uninformative situation nothing is known about the cause of the behavior, and so anxiety level is inferred to be neutral.

It seems intuitive that that we use context to interpret behavior, particularly ambiguous behavior: e.g., crying can be an expression of happiness (e.g., at a wedding) or sadness (e.g., at a funeral). The probabilistic attribution framework provides a natural mechanism to capture the intuition that our inference about someone's current state or mood combines observations of their ambiguous behavior with information about the situation. We extend the generative process underlying the probabilistic attribution framework (disposition and situation combine to produce the propensity for an action), by separating the current internal state (the propensity for an action) from the observed behavior itself with the assumption that the observed behavior is a noisy (but unbiased) reflection of the current internal state (Figure 2.10). In the Snyder and Frankel case, the internal state is the anxiety that the interviewed woman feels (which is influenced by the conversation topic), while the observed behavior is her ambiguously anxious body-language.

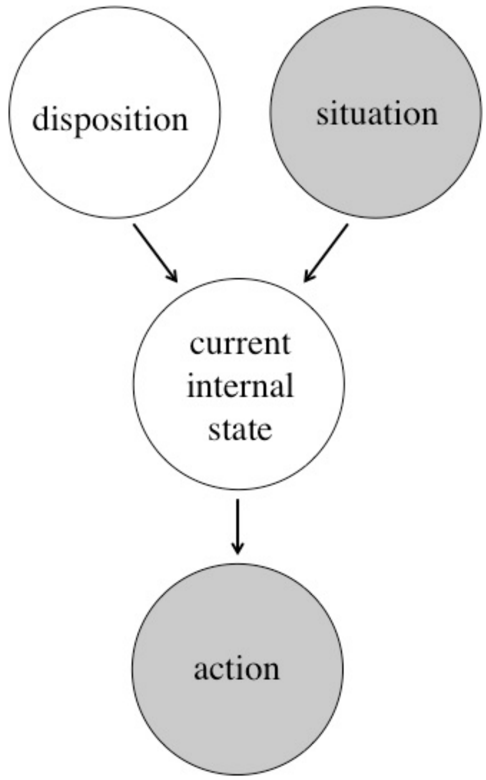


Figure 2.10. Inferring the current state from ambiguous behavior: Situation and disposition combine to influence the current internal state (as in Figure 2.1), and the observed behavior is a noisy reflection of the internal state.

Formally, we consider the internal state (*x*) to be drawn from a Beta distribution as in eq. 3, and the observed action (*a*) to be drawn from a beta distribution centered on *x:*

$$x \sim Beta\left(q\theta,(1-q)\theta\right)$$

$$a \sim Beta(xk,(1-x)k)$$

as in our previous cases, we take θ=10, and we set k=2. We can obtain inferences about the internal state of an actor by specifying some beliefs about the situation, and the ambiguity of the behavior. We take the woman's actions in Snyder and Frankel's stimulus to be perfectly ambiguous (behavioral anxiety of a=0.5 on a sale of 0 to 1), the topic of sex to be an anxiety provoking situation (s=1, i.e., 73% of dispositionally neutral people would be anxious in that circumstance), and politics to slightly reduce anxiety (s= -0.5, i.e., a person who feels anxious half the time would be anxious only 37% of the time while discussing politics[8]). In the "after" condition, nothing is known about the topic while watching the video, so we assume the situation is neutral (s=0). As in all previous simulations, we rescale our model estimates to the empirical scale (2-18) used in the experiment.

Figure 2.9 shows that the current (internal) anxiety estimated by probabilistic attribution is consistent with human judgments: compared to an unknown topic, an ambiguous action is perceived as more anxious in the context of an anxiety-provoking topic and less anxious when the topic discouraged anxiety.

Inferred Dispositional Anxiety

After answering questions about how anxious the woman was in the video, subjects next answered a variety of questions about how anxious they thought the woman

---

[8]Snyder and Frankel's choice of politics as a non-anxiety inducing topic may be contentious now, but seems more plausible of the early 1970s when there was less political polarization (Layman et al, 2006). Here, this assumption is only useful to capture the lower perceived anxiety in the "before-politics" than the "after" condition: if we assume that politics is a "neutral" (s=0) topic, then perceived anxiety would be equal in the "before-politics" and "after" conditions, but the discrepancy between "before-sex" and "before-politics" would remain.

was in general (e.g. how apprehensive she would be in a future situation, her percentile rank on anxiety, etc.). From this questionnaire, Snyder and Frankel calculated a "dispositional index" on a scale of 0 to 100 (Figure 2.11, panel A). Observers who had used the topic to inform their perception of the woman's anxiety in the video (because they had been told the topic ahead of time) inferred that she was generally more anxious when the topic was purported to be sex rather than politics. In contrast, observers who did not know the topics while watching the video, and thus did not perceive any differences in the woman's anxiety, showed a striking reversal of this pattern: they estimated that the woman was less anxious in general if they were told she had been discussing sex, rather than politics (Figure 2.11, panel A).

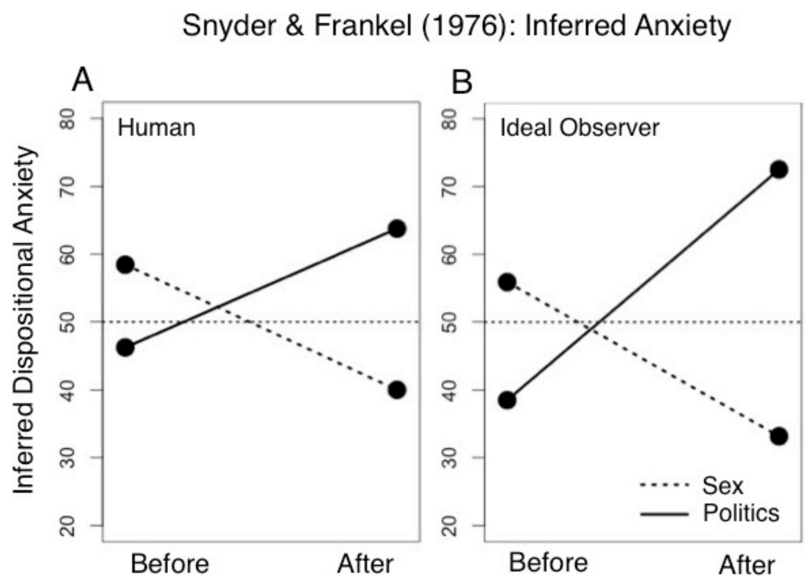


Figure 2.11. Inferred dispositional anxiety as a function of the topic, and when the observer learned the topic. *A:* Observers who knew the topic before watching the interview inferred that the woman who they thought had been discussing sex had a more anxious disposition in general (dotted line) compared to the woman who was asked about politics. (solid line). Observers who only learned about the topic after seeing the video made the opposite inference, thinking the woman who had been discussing politics had a more anxious disposition than the woman who'd been asked about sex. *B*: An ideal observer infers that the woman who displayed anxious behavior, (sex condition, dotted line) has a more anxious disposition than the woman who didn't display anxious behavior (politics condition, solid line). However, when there was no difference in perceived anxiety (after condition) the ideal observer infers that the woman has a more anxious disposition when it learns that she was discussing politics, and a less anxious disposition when it learns she was discussing sex.

Although the judgment of dispositional anxiety while knowing the situational pressures is superficially similar to the classic essay paradigm, the crucial difference here is that the perceived anxiety has itself been inferred from the situation, rather than directly observed (as discussed in the previous section). Thus, observers try to infer dispositional anxiety from the anxiety they had perceived in the video and the current situation; in other words, observers condition on the perceived anxiety in the video to judge general dispositional anxiety.[9] Consequently, the perceived anxiety used to estimate dispositional anxiety is high in the before-sex condition, low in the before-politics condition, and somewhere in the middle in the after conditions. In the before conditions, attribution to dispositional anxiety unfolds as in the classic FAE paradigm: an action perceived as high anxiety, although somewhat explained by the situation, still suggests higher dispositional anxiety (and vice versa for low anxiety). In contrast, in the "after" conditions, observers could not use the topic to judge the woman's anxiety in the video (thus they perceived the same neutral level of anxiety), but they can use the topic to estimate her dispositional anxiety. Observers seem to reasonably infer that someone who seemed neutral when in an anxiety provoking situation is likely to not be an anxious person in general, and someone who was neutral in an anxiety-reducing situation is more likely to be generally anxious.

Probabilistic attribution naturally captures this intuition, and thus reproduces the striking reversal in human judgments of dispositional anxiety in the before and after conditions. We use the same model parameters as in the previous section, and infer the dispositional anxiety by conditioning on the temporary, state anxiety the model had previously estimated. The qualitative pattern of inferred dispositional anxiety matches that of human observers, and the intuition outlined above (Figure 2.11, panel B). Again, we note that this result is obtained without altering any parameters from the application

---

[9]While it might at first seem odd to assume that observers condition on perceived anxiety to infer dispositional anxiety (rather than jointly inferring both properties, which would not yield the behavioral patterns), such sequential conditional inference seems to be the norm for humans; for instance, people seem to condition on their previous answers when making new judgments about the same visual stimulus (Jazayeri & Movshon, 2007; Stocker & Simoncelli, 2008).

of probabilistic attribution to estimating perceived anxiety – this result falls naturally out of conditional inference under uncertainty.

Experimental Support for Unbiased Social Attribution

We have demonstrated that six experiments classically interpreted as evidence that observers show a bias to over attribute behavior to situation (as well as a seemingly contradictory finding that we over-attribution behavior to situations) is predicted by a Bayesian inference framework. Using the information reported in the existing literature we have shown that the *general pattern* indicative of a FAE can be accounted for by this framework under *reasonable* parameter settings. However, reliance on the classic data does not allow us to test all of the critical predictions of this model, so we conducted three studies aimed to test the predictions delineated in the *What makes a situation informative about disposition?* section above. Specifically, if observers' social reasoning is consistent with an ideal observer, we expect: (1) if the same *action* is observed, attribution to internal qualities should decrease systematically as situational pressure is perceived to increase. Likewise, attribution to situation should decrease systematically as the influence of internal qualities are perceived to increase. (2) We would also expect to see inferences that have classically been interpreted as an FAE – observers should infer that the unobserved feature (either internal quality or situational influence) is consistent with the observed behavior (e.g. a pro-Castro essay should indicate a pro-Castro attitude). Making no inference here– as was previously proposed as the normative model – would not be consistent with optimal inference. However, (3) when the observed variable (either situation or internal qualitie) *is* perceived to totally compel the observed behavior, then no attributions should be made regarding the unobserved variable – this would be evidence of a *genuine* attribution error. Finally, this inference framework views social reasoning as essentially a process of inferring, under uncertainty, the most probable unobserved variables given the observed variables– within this framework there is no reason to think that situations and internal qualities are reasoned about differently than other latent causal variables, therefore, (4) reasoning about situation or internal quality should be equally subject to predictions delineated above.

A New Experimental Paradigm: Bucket Toss

According to the framework we have proposed, the relevant variables involved in social inference, which are sometimes known and sometimes unknown, are situational pressure, the relevant internal quality, and resulting behavior. To more directly test the predictions of our social inference model we need a scenario that involves an unambiguous situation and an unambiguous internal quality that maps directly to an observable behavior. We therefore invented a scenario that involves a made-up carnival game called bucket-toss. The objective of bucket-toss is to throw a quarter into a container (the observed action) from a fixed distance. In bucket-toss the size of the container can change (situational influence) and – as is usually the case in the world – individuals have varying skill (the internal quality) at this game. We first obtain objective measure of the perceived strength of the situation (e.g. how many people would make a quarter into a given container from five feet away), and then in two experiments we varied the container size and known skill, respectively, to compare human inferences to an ideal observer making inferences about either an unknown internal quality or an unknown situation given identical information.

Obtaining empirical estimates of situation strengths

Methods

The goal of this first study was to simply determine how strong different situations are perceived to be. This experiment was run online using participants recruited using Amazon Mechanical Turk. There were 143 participants who each completed five trials of the experiment. Participants first read that they would see five different containers and, after each, would be asked to judge how hard it would be to toss a quarter into the container from 5 feet away. On each trial the participant saw a profile image of a person standing 5 feet away from a container (Figure 2.12, see Appendix C for all profile images).

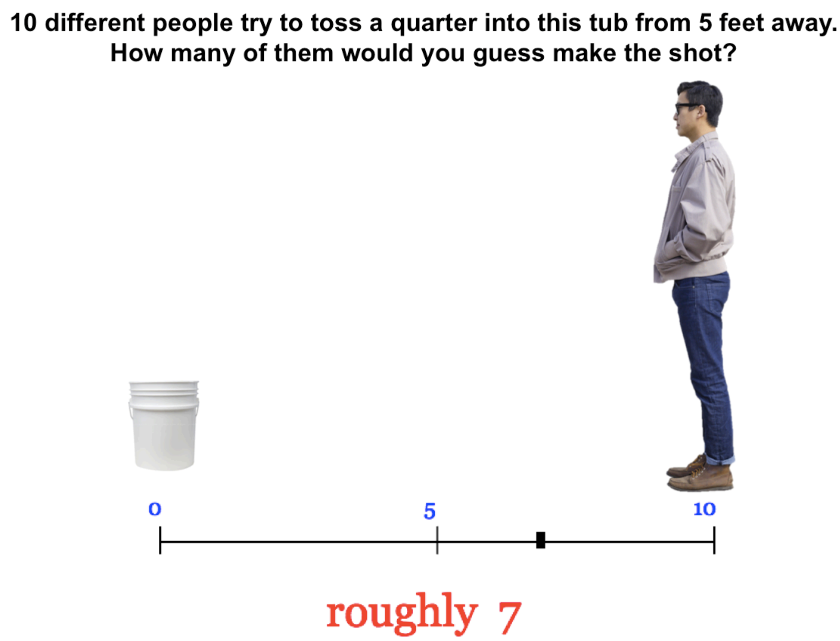


Figure 2.12. On each trial participants were presented with an image of a person and a container randomly selected from the five possible containers. Participants could select how many people out of ten they thought would make the quarter into the particular container shown by moving the toggle left or right from five. As the toggle moved the word below the image changed between integers indicating "roughly x" number of people

For each participant, each of the five containers (Figure 2.13) appeared once in a random order. In our model, the strength of the situation is defined as the proportion of people who would be compelled to perform a certain action, so questions were phrased in a way meant to elicit responses appropriate to this definition. On each trial participants were asked if "ten different people tied to toss a quarter into this container from five feet how many of these ten would you guess make the shot?" Participants used a mouse to move a toggle on a sliding scale. At the beginning of each trial the toggle started at the center (which corresponded to five people) and participants could indicate they thought more than five people would make the shot by moving the toggle to the right to, or less than five people by moving the toggle to the left. The number of people indicated by the slider position appeared below the scale, and this changed dynamically as the participant moved the slider.

Figure 2.13. In this study, to obtain objective perceptions of a semi-continuous situation we asked observers to rate how many people out of ten would be able to make, from five feet away, a quarter into each of these five containers.

Results

Although the number that participants saw appeared as integers (e.g. "roughly 7 people"), for finer granularity we recorded responses on a continuous scale based on the toggle location as a proportion between 0 (0 people) and 1 (10 people). On average (responses were skewed, so the median response was used) observers believed that only .06% of people could make the quarter into the shot glass, 23% of people could make it into the Solo cup, 59% of people could make the 5-gallon bucket, 79% of people could make the ice-tub, and 99% of people would make the shot into the kiddie pool. Importantly, as the container size increased, judgments about the proportion of people who could make the shot increased monotonically and were largely non-overlapping

(Figure 2.14). These features are important as they validate that these container sizes effectively communicate distinguishable situational pressures.

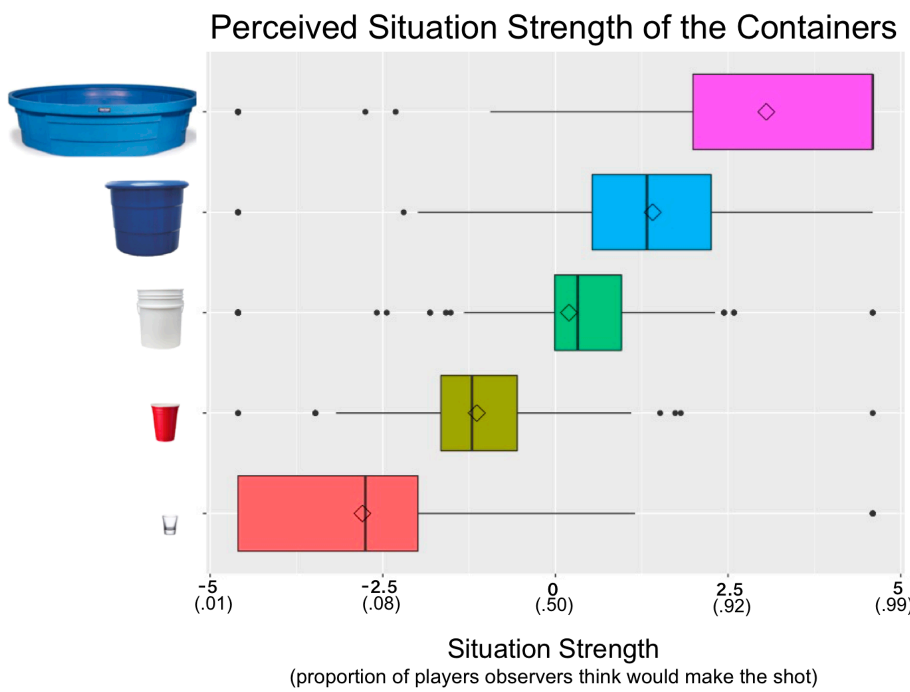


Figure 2.14. Each of the five containers we perceived to have distinguishable situation strengths. Because the parameters in the social attribution model are additive in log-odds we logit transformed the proportions, which are the units shown on the x-axis. The proportions that corresponds to each are shown in parentheses. For example, a situation that would compel a proportion of .92 people to make the shot correspond to $s = 2.5$.

Experiment 1: What internal-quality is inferred when situational pressure is varied?

Methods

In this experiment, we measured observers' inferences about a player's bucket-toss *skill* when they perform a certain action (make or miss the shot), which was either motivated or discouraged by the situation (container size), so that we can compare this behavior to an ideal observer model constrained by empirically estimated perceptions of situation strength. This experiment was conducted online with participants recruited from Amazon Mechanical Turk. There were 299 participants, who each made ten judgments. Participants were told the rules of bucket-toss: it is a game in which players try to toss a quarter into various containers from a fixed distance. These initial instructions were presented along with an image of the five containers that would be seen during the experiment. Participants were also told that they would see a series of players and that they would need to guess how skilled each person was at the game. On each trial, participants saw a profile image of a person standing a fixed distance away from a container, and were told the person either made or missed the shot (Appendix D). Observers were asked to report how skilled they thought the player was, and they made this judgment by moving the toggle on a numberless sliding scale that ranged from "very unskilled" on the right and "very skilled" on the left, with "average" in the center. As the participant moved the toggle the words below the scale changed dynamically to reflect how the toggle position corresponded to the percentage of people better or worse than the target player. On each of the ten trials the container presented was selected randomly with the constraint that each would appear twice per participant. Whether the player was shown to make or miss the shot was determined randomly on each trial. The name and image of the player shown was selected randomly from a bank of 12 (6 males and 6 females; Appendix C) with the constraint that each person/name would only appear once per participant.

Model and Results

We can now compare how humans attribute actions to traits of an individual and the attributions of an ideal observer. The ideal observer uses estimates of situation strength (container difficulty) obtained previously, to estimate the skill level of a bucket-toss player who either made, or missed, a shot into a container of a particular size (Figure 2.15, panel A). Likewise, we obtain human observers' ratings of skill level on those same types of trials (Figure 2.15, panel B).

Human judgments track the critical predictions of the probabilistic social attribution model: (1) For a particular *action* (made or miss) attribution to internal qualities decreases as perceived situational pressure increases. (2) The unobserved variable (skill) is estimated to be in a direction consistent with the observed behavior (above average skill is inferred when the person makes the shot, below average skill is inferred when the person misses the shot). (3) However, no attributions are made when the situation is perceived to be totally compelling – when the player makes the quarter into the kiddie pool or fails to make the shot glass.

These data also show the expected "fundamental attribution error" style behavior: when situations weakly motivate the behavior (missing the Solo cup v. making the ice bucket), inferred ability differs between making and missing a toss ($t(603) = 10.78$, $p < .001$). This inference is weaker than in the least constraining situation (five-gallon bucket; $t(597) = 11.51$, $p < .001$), but it vanishes entirely in the very influential situations (shot glass vs kiddie pool; $t(580) = .075$, $p = .94$. Together, these results show that when we parametrically manipulate situation strength, and measure the strengths people estimate for these situations, human behavior is qualitatively and quantitatively consistent with probabilistic attribution.

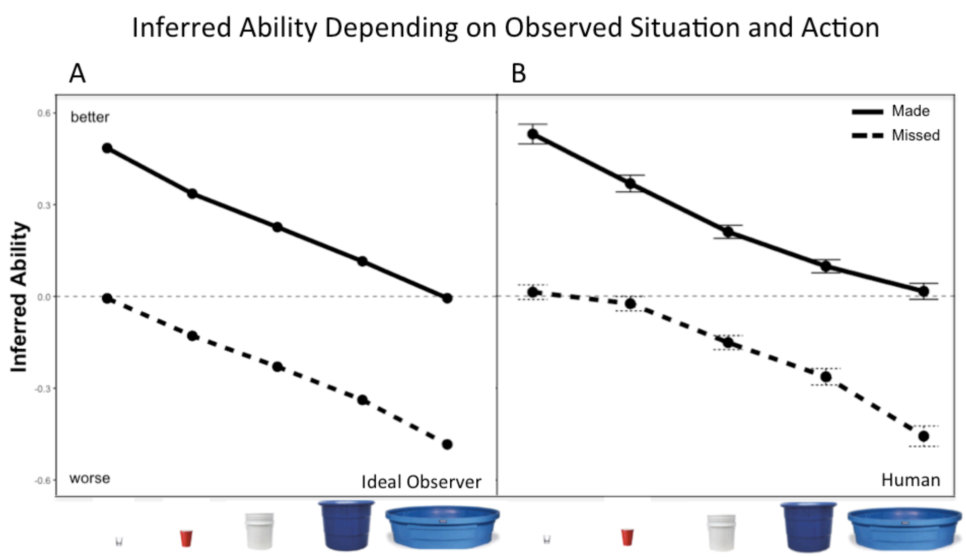


Figure 2.15. Inferred ability at bucket-toss as a function of container size, and whether the player made or missed the shot. *A*: The inferences made by the ideal observer illustrate the critical predictions of the social inference framework. When the player makes the shot (black line) the ideal observer infers that this is indicative of above-average ability at bucket-toss, but as the situation exerts more pressure to miss the shot (smaller containers) more exceptional ability is inferred. Likewise, when a player misses the shot (dotted line) the ideal observer infers that the player is below average at bucket-toss, but as the situation becomes more conducive to making the shot (larger containers) more exceptional inability is inferred. No behavior-consistent inference are made when the situation becomes nearly deterministic – missing the shot glass or making the kiddie pool. *B*: Human inferences are consistent with the critical predictions of the social inference model. When action is held constant, attribution to ability decreases systematically as situational pressure is perceived to increase. Behavior-consistent inferences are made until the situation is perceived to be totally compelling.

Experiment 2: What situational pressure is inferred when internal-quality is varied?

In the previous experiment the unknown variable was bucket toss skill, which was inferred given knowledge about the success of the shot and the container size. This is analogous to most attribution experiments—participants are asked to infer an unknown internal quality given an observed situation and action. However, the Bayesian social attribution framework says that the pattern of inferences in these studies arises due to inferring unknown variables under uncertainty, patterns of inference should be similar regardless of whether the unknown feature is the situation or the disposition. So, in this experiment we were interested in the inferences people would make about the situation (container size) when the internal-quality (skill at "Bucket Toss") is known, and how this compares to the inferences of an ideal observer operating with the same information.

Methods

This experiment was conducted online and 305 Amazon Mechanical Turk users participated. Each participant competed eight trials. It was important that participants thought the situation (the container the player was aiming for) was chosen at random and was not related to their skill at "Bucket Toss." (e.g. it was not a "level" they had achieved). To this end, we told participants that "Bucket Toss" is a popular carnival game of "skill and chance," and that the first step in the game is the "the luck of the spin" -- that players first begin the game by spinning a large wheel which determines which container they'll aim for. An image of a person spinning a wheel was shown along with images of the five containers on which the spinner could land (these were the same containers as used previously; Figure 2.13). Participants were told that players would have to try to toss a quarter into whichever container the pointer indicated from a fixed distance away. On each trial participants were presented with a profile image of a person. When the trial began a graphic of a question mark was in the place where the container would be. The participant was told the players' skill-level in terms of percentile (e.g., "Jenifer is better than 99% of people at Bucket Toss") as well as whether the person made or missed the shot they were aiming for. Participants then indicated which container they thought the player had been shooting for. On each trial the skill percentile (1^{st}, 25^{th}, 50^{th}, 75^{th}, or 99^{th}) and whether or not the person made or missed the shot was selected at random. The person-image and name were selected at random with the constrained that they could only appear once per subject.

Response Scale

Participants were asked which container they thought the player had been aiming for, but their responses were given in the form of a bet. We changed the scale in this way in order to provide participants with a meaningful way to express little or no information[10]. They were told to imagine that that they had $10 to bet, and that they could

[10] The reason we used this more complicated response scale was because the center of the simpler scale used in Experiment 1 would not have been meaningful. If an observer has little or no information about a person's skill (for example, in Experiment 1 when the player made a shot that about 50% of people could make) it makes sense to guess this player is about average – if you have some assumption that skill is normally distributed, than average (the middle of the scale) is the optimal guess when you have no other information. However, in this experiment if a participant has little or no useful information about container size there is no meaningful way to express this on the conventional scale, since the center of the scale (the 5-gallon bucket) is no more likely to

distribute the bets over multiple containers. They could distribute these bets by dragging a toggle along a one-dimensional line. The response scale started with the toggle in the center of the scale with monetary bets distributed equally over the 5 buckets ($2 on each bucket; Figure 2.16, panel A). Below the response scale buckets were ordered from smallest to largest. As the participants dragged the toggle to the right this indicated they were betting more money on the larger containers; as the subject dragged the toggle to the left this indicated they were betting more money on the smaller containers. A histogram that corresponded to the proportion of money bet on each container acted as a visual cue that changed dynamically as participants changed their bets. Before the experiment participants were given examples of the meaning of the response scale. For example, they were told that if you were asked to bet on which container a man had used to give his dog a bath than they might distributed bets over many containers (Figure 2.16, panel B), dragging the toggle toward the right to place money disproportionately on the larger containers, meaning that they bet roughly $6 that he gave his dog a bath in the kiddie-pool, roughly $3 that he used the drink-tub, roughly $1 that he used the 5-gallon bucket, and less than a $1 that he used the red-cup. On the other hand, if they wanted to bet all $10 that the man gave his dog a bath in the kiddie-pool they could drag the toggle all the way to right, indicating that all $10 was placed on this container (Figure 2.16, panel C). If they were asked which container they thought a man used to hold liquid they

---

occur than any other container (there is a one in five chance for each container. Being able to distribute bets allows for a uniform distribution of bets, which is optimal with no other information.

might leave the toggle in the middle of the scale, indicating that they were distributing bets equally among all containers (Figure 2.16, Panel A).

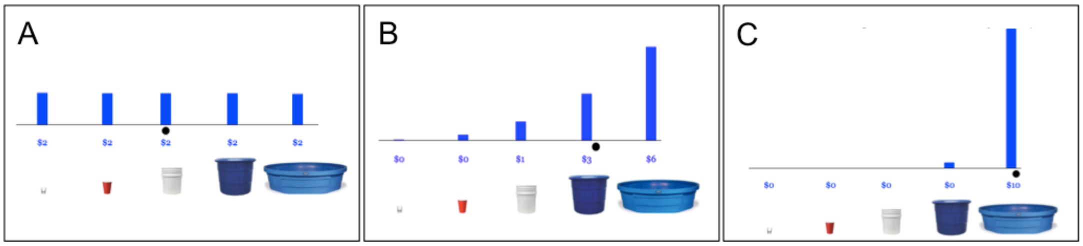


Figure 2.16. The response scale started with the toggle in the center of the scale with monetary bets distributed equally over the 5 buckets (A) Participants could move the toggle toward the right to place money disproportionately on the larger containers (B), or to the left to place bets on smaller containers (not shown). Moving the toggle all the way to the right (or left) indicated a bet of all $10 on the most extremely size container, for example the kiddie pool (Panel C).

Model and Results

Again, we can compare human attributions to the attributions of an ideal observer. The ideal observer uses internal-quality (skill percentile) to estimate the situation strength (container size) when the player either made, or missed a shot (Figure 2.17, panel A). Likewise, we obtain human observers' judgments of the most probable container size on those same types of trials (Figure 2.17, panel A).

Again, we see the graded pattern of inferences predicted by probabilistic attribution: as we increase the ostensible skill of the player (strength of the internal quality), observers' estimate a missed shot to be ever more diagnostic of container size, and made shots become ever less diagnostic. However, in these data, human judgments are not consistent with all of the predictions of the probabilistic social attribution model: container size is only estimated to be in a direction consistent with the observed behavior when the players' skill is at or above the 50th percentile and makes the shot, or at or below the 50th percentile and misses. The pattern for missed shots might make sense as a consequence of the response scale we used – there is no difference in judgments between the 25th and 1st percentile conditions, perhaps because people are encouraged to distribute bets homogenously. However, the counterintuitive pattern for made shots is more puzzling: people estimate that a 99th percentile player who made a shot is likely to be throwing into one of the *smaller* containers; this peculiar pattern might be a consequence of people not fully believing that container size is allocated randomly to players, and thus believe that an especially skillful player is likely to be matched up with an especially challenging target. This might imply a fascinating direction for future research: people do not believe situations and dispositions to be independent – dispositions likely influence what situations we find ourselves in.

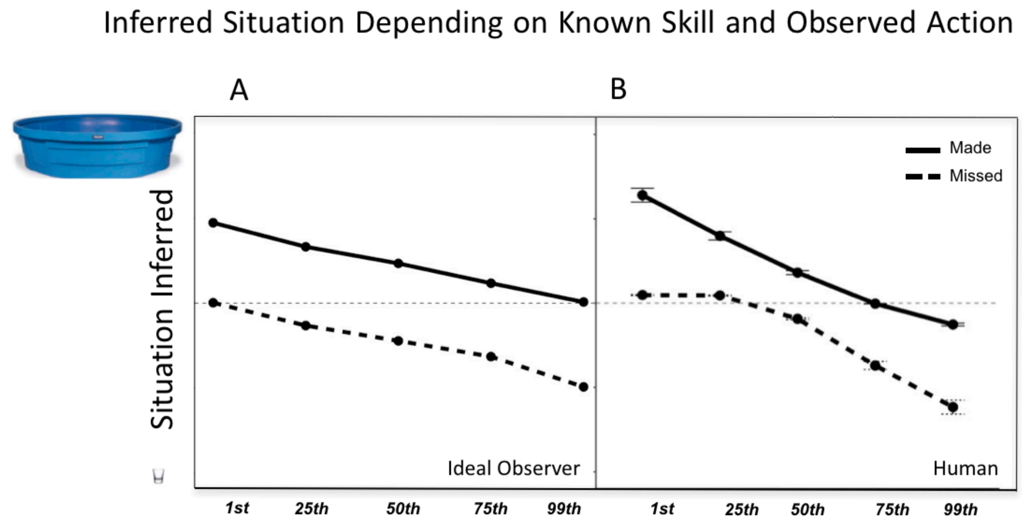


Figure 2.17. Inferred situation as a function of skill, and if the player made or missed the shot. *A*: The inferences made by the ideal observer illustrate the critical predictions of the social inference framework. When the player makes the shot (black line) the ideal observer infers that this is indicative of a situation that facilitates the behavior (a larger container); a missed shot (dotted line) is indicative of an impeding situational force (a smaller container). No situational inference is made when the skill becomes deterministically strong – the exceptionally skilled player would make any shot presented, the exceptionally unskilled player would miss. *B*: Human inferences are consistent with the prediction that when action is held constant, attribution to situation decreases as skill is perceived to increase. Unexpectedly, however, no inferences about container size is made when a player in the 75^{th} percentile makes the shot, or a player in the 25^{th} percentile or below misses the shot, and one of the smaller containers is inferred when a player in the 99^{th} percentile makes the shot.

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 FAE 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 exactly antithetical to The Fundamental Attribution Error 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.

Our intention in this work is to broadly demonstrate that behavior traditionally interpreted as a bias in the classic FAE 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 FAE 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 FAE: over-attributing 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 cooking in mis-calibrated expectations about situations, but rather arise from the structure of probabilistic causal inference.

Beyond demonstrating the classic literature is qualitatively consistent with this framework, we also directly test predictions of the social inference model by obtaining objective measure of the perceived situation strengths so we are able to compare human attributions to internal-qualities to the attributions of an ideal observer without having to make any assumptions about how strong observers think situations are. We show that when we parametrically manipulate perceived situation strength, human observers' inferences are qualitatively and quantitatively consistent with probabilistic attribution. We also show that when we manipulate disposition parametrically inferences about situation are largely consistent with an ideal observer (though there are inconsistencies that arise and warrant further investigation).

Indeed, the critical contribution of our paper 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 Appendix A and B for elaboration) would yield the qualitative effects in the classic FAE 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.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329-349.

Gawronsi, 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

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological bulletin*, *117*(1), 21.

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.

Gladwell, M(2000). *The Tipping Point.* Boston: Little, Brown

Griffiths, T. L., & Tenenbaum, J.B. (2005). Structure and strength in causal induction. *Cognitive psychology*, 51(4), 334-384.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological review*, *116*(4), 661.

Holyoak, K.J. & Cheng, P.W.(2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology*, 62, 135-163.

Jennings, K. E. (2010). Determining the Internal Consistency of Attitude Attributions. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society. Austin, Texas: Cognitive Science Society* (pp. 978-984).

Kim, J. Y., Natter, M., & Spann, M. (2009). Pay what you want: A new participative pricing mechanism. *Journal of Marketing*, *73*(1), 44-58.

Jones, E.E., & Davis, K. E. (1965). From acts to dispositions: the attribution process in person perception. *Advances in Experimental Social Psychology*, *2*, 219-266.

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, *3*(1), 1-24.

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.

Kelley, H. H. (1973). The processes of causal attribution. *American psychologist*, *28*(2), 107.

Kemp, C., & Tenenbauum, J.B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, 116(1), 20.

Layman, G. C., Carsey, T. M., & Horowitz, J. M. (2006). Party polarization in American politics: Characteristics, causes, and consequences. *Annu. Rev. Polit. Sci.*, *9*, 83-110.

Machado, F., & Sinha, R. K. (2013). *The viability of pay what you want pricing*. Working Paper.

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.

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.

Nisbett, R. E., Caputo, C., Legant, P., & Marecek, J. (1973). Behavior as seen by the actor and as seen by the observer. *Journal of Personality and Social Psychology*, 27(2), 154

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.

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.

Ross, L., Greene, D., & House, P. (1977). The “false consensus effect”: An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*(3), 279-301.

Sherman, S. J. (1980). On the self-erasing nature of errors of prediction. *Journal of Personality and Social Psychology*, *39*(2), 211.

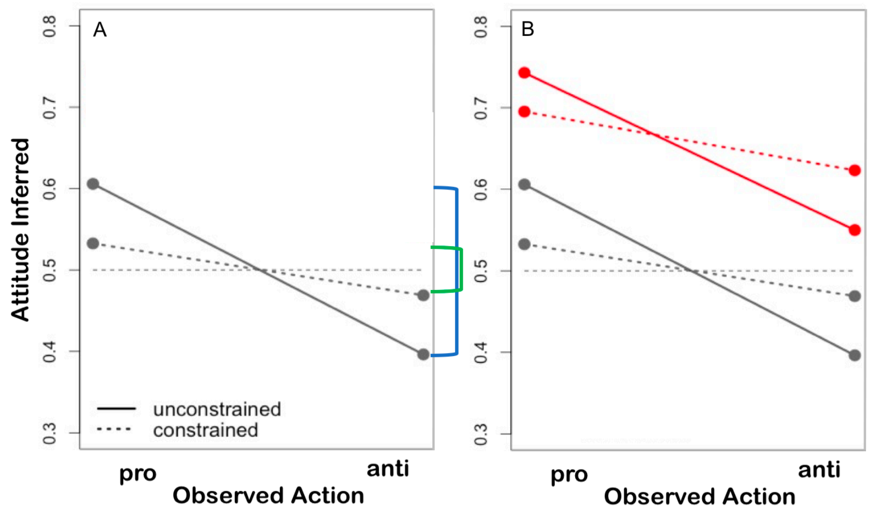
Snyder, M. L., & Frankel, A. (1976). Observer bias: A stringent test of behavior engulfing the field. *Journal of Personality and Social Psychology*, *34*(5), 857.

Taylor, S. E., & Fiske, S. T. (1978). Salience, attention, and attribution: Top of the head phenomena. *Advances in experimental social psychology*, *11*, 249-288.

Trope, Y., & Alfieri, T. (1997). Effortfulness and flexibility of dispositional judgment processes. *Journal of Personality and Social Psychology*, *73*(4), 662.

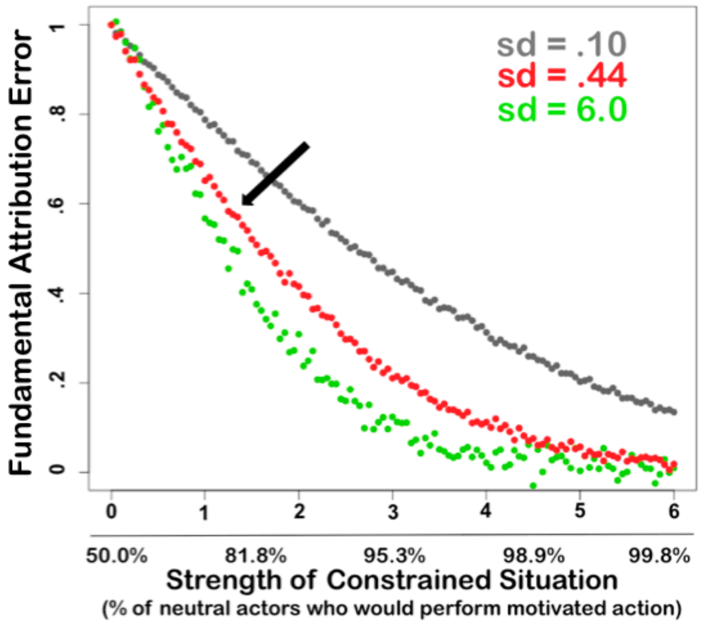
Trope, Y., Cohen, O., & Maoz, Y. (1988). The perceptual and inferential effects of situational inducements on dispositional attribution. *Journal of Personality and Social Psychology*, *55*(2), 165.

Appendix A



The Fundamental Attribution Error is not a peculiar phenomenon that arise 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: 1) observer's belief about situation strength, 2) observers' belief about the central tendency of peoples' dispositions and 3) variability of dispositions in the world. When a binary action (e.g. a pro or anti essay) is observed, the degree of FAE can be thought of as the difference between the attitudes inferred from the two distinct actions (difference between the points on the dotted grey 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 grey 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 FAE, it simply shifts all of the inferences, leaving the ratio representing the degree of FAE intact (red lines in panel B)

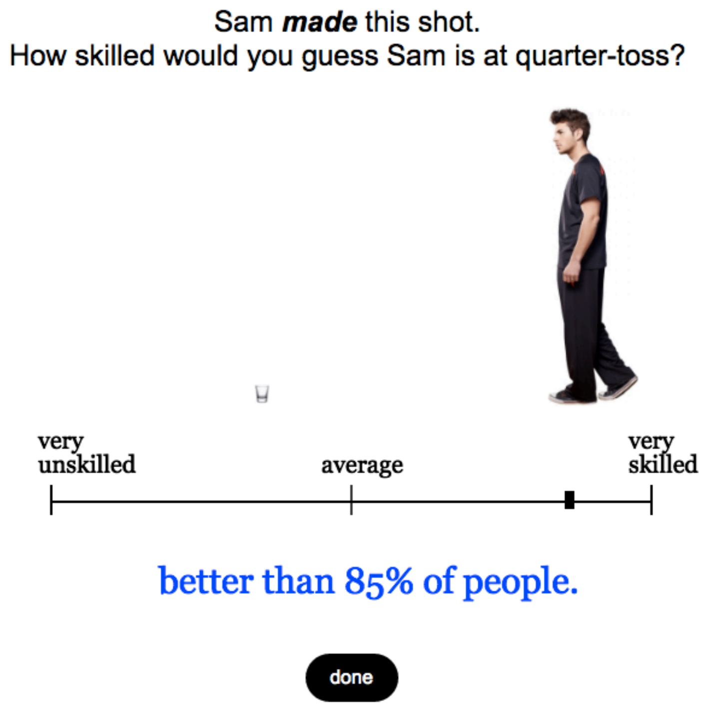
Appendix B



Although belief about the central tendency of the disposition doesn't change the degree of the FAE, the perceived strength of the constraining situation and the perceived dispersion of attitudes in the world *does* have an effect on the degree of FAE. Critically, however, the probabilistic social attribution model produces the pattern of inferences indicative of a non-negligible FAE 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 an FAE. 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.

Appendix C

Appendix D


Sam *made* this shot.
How skilled would you guess Sam is at quarter-toss?
very
unskilled
average
very
skilled
better than 85% of people.
done

Appendix E

In general, Jennifer is **BETTER than 99%** of people at "quarter-toss."
At the fair this year she spun the wheel and then **MISSED** the shot she was told to aim for.
Which container do you think Jennifer was trying to make the quarter into?

Chapter 2, in full, is currently being revised for publication of the material. Walker, Drew E.; Smith, Kevin; Vul, Edward. The dissertation author was the primary investigator and author of this paper.

CHAPTER 3

Successes of the intuitive psychologist

Walker, D.E. and Vul, E.,

Manuscript in preparation for publication

Abstract

In a now classic experiment, Ross, Amabile & Steinmetz (1977) showed that observers think that a participant who is randomly assigned to invent questions has more general knowledge than a participant assigned to answer these questions. These questioner-contestant results have been interpreted as conceptually related to the Fundamental Attribution Error: observers don't properly adjust for the advantage afforded by arbitrary social roles when making judgments about others (e.g. they ignore "role-conferred advantages"). The implication in the literature is that this error is evidence of a non-normative reasoning process. However, since errors are not always indicative of irrationality, here we specify and test the non-normative processes that could cause this behavior: (1) observers don't consider all of the available information that is relevant to this judgment, or (2) they do consider relevant information, but systematically under-appreciate situational influences. In two series of experiments, we demonstrate that observers (a) are sensitive to the advantages conferred and disadvantages imposed by social roles, (b) are better calibrated to the social-roles used in this paradigm than has been argued, and (c) make sensible use of other information available to them.

Imagine you are a university student listening to a professor lecturing on her topic of expertise. Do you take her behavior at face-value and conclude she is generally brilliant, or do you consider that she is speaking on a topic for which she has unique idiosyncratic knowledge? It has been proposed that “in drawing inferences about actors, perceivers consistently fail to make adequate allowance for the biasing effects of social roles upon performance” (Ross, Amabile, & Steinmetz, 1977). That is, that we do not account for these *role-conferred advantages* when forming impressions of others – we make judgments based on the behavior we witness without properly adjusting for the advantage afforded or disadvantages imposed by social roles, and fall victim to a “a special case of a more fundamental attribution error.” In the classic demonstration of this phenomenon, Ross, Amabile and Steinmetz (1977) asked students to play a quiz game in which one participant was randomly assigned the task of inventing questions (the “questioner”) and the other was assigned to try to answer these questions (the “contestant”). Unsurprisingly, the questioner was able to invent questions that the contestant failed to answer correctly, but despite explicit awareness of the different and arbitrary roles, observers still inferred that the questioner had more general knowledge than the contestant.

The implication in the literature is that this error – like errors in general—is evidence of non-normative behavior. However, errors can easily arise from normative behavior— even the most rational observer will make errors when operating with incomplete or inaccurate information. Since there is more than one decision-making process that could result in this pattern of inference, if we hope to understand the source of the error, and whether or not it results from a non-normative social reasoning process, it is important to articulate these possibilities.

One non-normative reasoning process that would result in the *role-conferred advantage* behavior is that observers don't consider all of the available information that is relevant to the judgment they are asked to make. For example, they don't consider the different role constraints, or they don't utilize information present in the question content, and they make judgments based simply on which player knew the answer to more questions. A classic explanation for errors in social inference is the idea that "behavior engulfs the field" (e.g., Heider, 1958, p. 54; Snyder & Frankel, 1976) and observers are so influenced by behavior that they fail to consider more nuanced features of the scenario. If the role-conferred advantage does indeed arise from this simple heuristic account, we would expect a questioner who asks more questions to always be judged as superior on general knowledge but a more rational observer to be sensitive to their relative performance in light of their social roles. We would also expect an observer who is taking advantage of relevant information in the contestant-questioner exchange to also be sensitive to features beyond social roles, such as the difficulty or cleverness of the trivia questions the players ask or answer.

Another possibility is that the role-conferred advantage arises not because observer doesn't consider social roles or other relevant information, but because they systematically under-appreciate the situational influences – observers take the social role into account, but they underestimate the advantage it provides the questioner. This might arise if observers generally underestimate how much idiosyncratic knowledge the average person has; thus evidence that the questioner knows something the contestant does not would mean that the questioner simply knows more things. Even though the average person would be able to invent ten challenging trivia questions if put in the same position, observers don't realize this and so average behavior seems above average.

In two series of experiments we investigate to what extent human behavior is consistent with either of these non-normative explanations. By modulating the number of questions asked and answered, we demonstrate that it is not the case that observers are using a simple heuristic that totally ignores differential social roles. Players who know the correct answer to more trivia questions are not always judged to have superior general knowledge. Furthermore, observers also seem to utilize other relevant information, as smart and clever questions have a predictable influence on the inference observers made. We did find evidence when the use of idiosyncratic knowledge is made salient observers no longer judge the questioner to be above average, or superior to the contestant, and the classic role-conferred advantage disappears. However, *underestimating* others' use of idiosyncratic knowledge may not be what is driving the classic inference pattern in the quiz game paradigm. Surprisingly, we found evidence that the average contestant actually *can't* invent ten challenging questions in the time allotted, and so it's questionable if the inferences observers made in the original scenario are actually errors at all.

Experiment 1

In Chapter 2 we showed that observers are sensitive to the intensity of others' actions when making inferences—for example, observers infer that a *stronger* pro-marijuana essay indicated a stronger pro-marijuana attitude compared to a weak pro-marijuana essay and the inferences they make about different actions is predicted by an ideal observer (Figure 2.6). If observers are making rational inferences in the quiz-game paradigm, we would expect the degree of the inference to scale with the observed behavior, but also that the relevant behavior should depend on which player is being evaluated. That is, an observer who considers the different roles should not simply make judgments by comparing the simplest common action: who knew the correct answer to the most questions. The questioner demonstrates their knowledge by *inventing* trivia questions, and so this person could be evaluated on dimensions such as the difficulty and breadth of their questions, or even more simply: how many trivia questions she is able to invent. The contestant on the other hand demonstrates her level of general knowledge by *answering* questions-- this person could be evaluated on dimensions such as how many questions they are able to answer correctly, and how difficult those questions were.

Experiment 1a

In Experiment 1a we held the number of questions the questioner asked constant and manipulated the number of questions the contestant was able to answer. One possibility is that observers are totally insensitive to the different roles, and are simply judging the players based on which player knows the correct answer to the most questions. If so, we would expect the contestant to never be rated equivalent to the questioner until this person can answer all of their questions. On the other hand, it could be the case that observers are aware that different roles make it easier for the questioner to know the answer to more questions, but they simply underestimate *how* advantageous the role of questioner is. That is, they are aware of the difference but underestimate the magnitude of this difference. If this is the case – they know that *inventing* ten challenging trivia questions is easier than correctly *answering* ten questions someone else invented – we would expect that at some point the contestant's behavior to be impressive enough to be judged as equivalent to the questioner, even before this person can answer all ten questions. From the original study the questioner is judged to be above average and superior to the contestant, who is rated about average. If observers demonstrate sensitivity to the different roles, this manipulation also allows us to identify at which point the questioner's behavior is seen as impressive enough to be judged to have general knowledge that is as exceptional as the questioner's.

Stimulus Creation

In order to produce a set of trivia questions that accurately reflected what undergraduate students could produce in a 15-minute period (the time allotted in the original experiment) we brought 19 UCSD students into the lab (one group of eight and one group of eleven) who received partial course credit for their participation. Students were seated around a conference table and the instructions described in Ross, Amabile, and Steinmetz (1977) were read aloud to the group by the researcher. Subjects were told that their task was to create ten challenging but not impossible trivia questions for which they knew the answer, and that these answers should be only one to a few words long. They were given some examples of typical trivia questions (e.g. Who is the Greek Goddess of victory?) and told to avoid both easy questions (e.g. number of days in April) and unfair questions (e.g. the name of their brother), and were encouraged to draw from their own areas of interest or expertise. They were told that they would next take turns asking their trivia questions to one another.

In order to both produce a set of realistic incorrect answers and get an objective measure of question difficulty subjects next took turns reading their questions aloud while the other participants answered the questions silently with pen and paper. The questioner was instructed to go at a pace that would allow others to generate and write a response; those answering questions were told to ask the questioner to slow down if necessary. The experimenter wrote answers to questions along with the subjects to make sure the pace felt appropriate and asked the reader to slow down when necessary. Subjects produced 124 usable questions total (Appendix A; some subjects were unable to generate 10 questions in the time allotted; some of the questions generated did not meet the specified criterion. This is discussed in more detail in Experiment 1c).

Participants

Subjects were recruited from the Amazon Mechanical Turk Marketplace and received money for participating. Three-hundred and sixty people attempted the task, but fifty-six participants were eliminated before data analysis for clicking through the experiment without moving the response scale or for not passing an attention check, leaving 304 participants that were included in data analysis.

Procedures

Participants read that they would be witnessing the interaction of two undergraduate students who had participated in this activity in an academic psychology lab setting. They were told that these students arrived at the lab and were asked to participate in a general knowledge quiz game in which one person was randomly assigned (by coin flip) to have the job of being the questioner and the other was assigned the job of being the contestant. The participants read the instructions the questioner was given (described previously in the stimuli section), and were told that the questioner posed each question to the contestant one at a time and the contestant answered as best as he/she could. The participant then watched what they were told was a replay of this interaction.

In each of ten trials the participant first saw the trivia question the posed by the questioner (Figure 3.1, A) and next saw the contestant's response and whether or not the response was correct (Figure 3.1, B) or incorrect (Figure 3.1, C). If the response was incorrect the correct answer was also shown. Between-subjects, we varied the number of questions the contestant had answered correctly (either zero, four, six, eight or ten). For each participant, the ten questions were drawn at random from the 124-question stimuli set, and which questions were answered incorrectly by the contestant was also determined randomly. The incorrect answers displayed were the modal incorrect answer given by subjects in the stimuli creation phase (or if there was no modal incorrect answer, we used an incorrect answer that seemed reasonable). After watching the ten trials unfold chronologically participants were asked to judge how much general knowledge the contestant and questioner, respectively, had relative to other people. Each judgment was elicited in a randomly determined order. This rating scale was specified as the proportion of people with more or less general knowledge than the target person, and participants made these ratings using a sliding scale that ranged from "less general knowledge" to more "general knowledge," centered on "average." Thus, effectively, we asked observers to rate the general knowledge percentile of the questioner and contestant. The proportion that corresponded to the toggle position was shown, and changed dynamically as the toggle was moved.

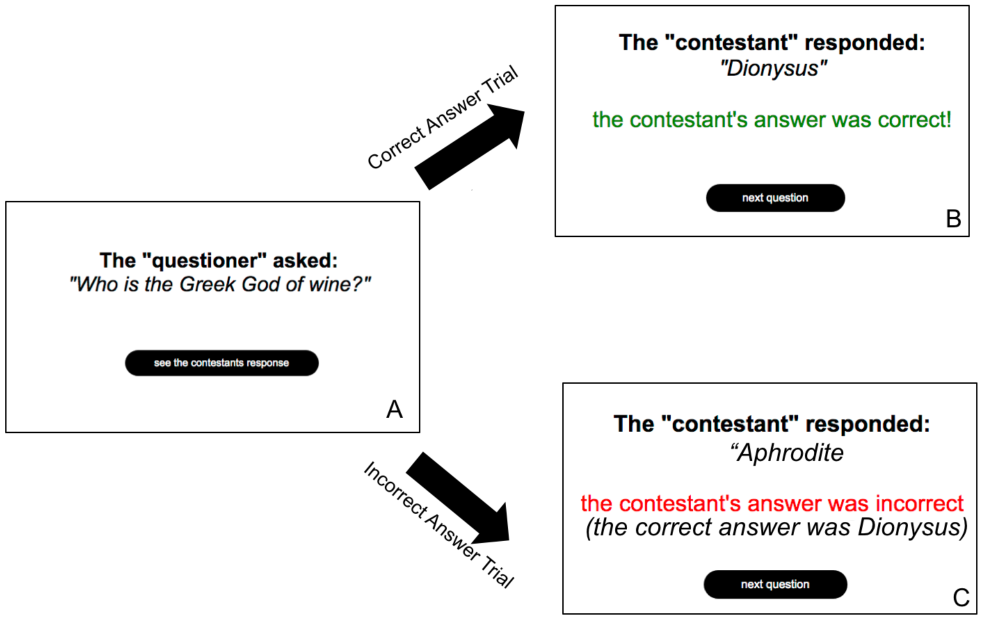


Figure 3.1. On each trial the trivia question posed by the questioner appeared alone on the screen (A) and when the participant had finished reading the question they could advance to the contestant's response. On contestant-correct trials (B) the *contestant's* response appeared on the screen followed by the words "the contestant's answer was correct!" in green lettering. On contestant-incorrect trials (C) the *contestant*'s response was given followed by the words "the contestant's answer was incorrect" in red lettering and with the correct answer, provided in parenthesis.

Results and Discussion

Importantly, we were able to replicate the original Ross, Amabile and Steinmetz (1977) finding on this online platform. When the questioner asked ten trivia questions and the contestant answered four of these (the number the average contestant was able to answer in the original study) observers judged the questioner to be in the 66^{th} percentile and have general knowledge superior to the contestant who they judged to be in the 50^{th} percentile ($t(118) = 4.9$, $p < .001$; Figure 3.2, A).

As expected, when the contestant was able to answer a larger number of questions correctly, observers inferred they had more general knowledge. When the contestant answered two or fewer questions, observers thought they were below average on general knowledge; when they could answer six or more, they thought they were above average; and when they could answer four correctly, observers inferred the contestant had an average amount of general knowledge. Interestingly, this inference is well-calibrated to ground truth: participants in the stimuli generation phase were, on average, able to answer about four questions ($M = 3.7$, $sd = 1.7$). This is also similar to the average number of questions that Ross, Amabile and Steinmetz (1977) reported their contestants were able to answer. So, as expected, observers are sensitive to the contestant's actions: the more questions they answer, the more general knowledge they are inferred to have. Moreover, observers seem well-calibrated to constraints imposed on the contestant: the average person can answer about four questions, and when contestants answer four questions out of ten, they are judged to be average.

The questioner (who is always seen to ask ten questions) is inferred to be well above average, and thus deemed to possess more general knowledge than the contestant, up until the point that the contestant can answer eight questions correctly. At this point the players are judged be no different on general knowledge ($t(118) = .73$, $p = .46$) and both to be well above average in the 73rd percentile. This suggests that people are not using a simple heuristic that knowing more questions amounts to having more knowledge, regardless of the role. If observers were insensitive to social roles, we would expect the questioner to always be rated as superior until the questioner was able to answer all ten questions. Experiment 1a suggests that the discrepancy between the ratings of the contestant and questioners' general knowledge is not due to a total blindness to different social roles, or to a miscalibration to the constraints imposed on the contestant, but -- consistent with the original study -- arises from a miscalibration to the role of the questioner – the questioners' ability to invent ten challenging Trivia questions seems to exceed our expectations.

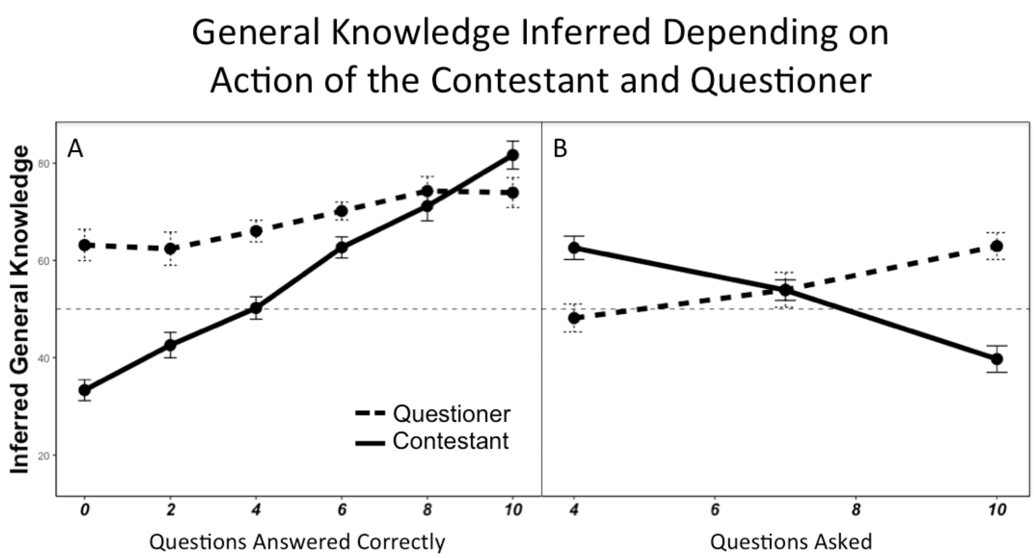


Figure 3.2. The rated general knowledge of the contestant and questioner as a function of (A, Expt 1a) the number of questions answered correctly (given that ten were asked), and (B, Expt 1b) the number of questions that were asked (given that three were answered correctly). (A) As the contestant answers more questions she is rated as having more general knowledge, at about four out of ten answered correctly, she is estimated to be about average. The role-conferred advantage (contestant is rated as less knowledgeable than the questioner) holds when the contestant answers fewer than eight out of ten questions. (B) As the questioner asks more questions she is rated as having more general knowledge. When she asks only four question the role-conferred advantage reverses and she is estimated to have less general knowledge than the contestant and to be numerically below average. When she asks seven questions she is rated equal to the contestant, and when she asks ten questions the role-conferred advantage reemerges.

Experiment 1b

Observers infer that the questioner is well above average on general knowledge when they are able to invent ten challenging trivia questions in ten to fifteen minutes, suggesting that although observers are aware that questioners have an advantage, they underestimate the magnitude of this advantage. Observers find the ability to invent ten challenging trivia questions for which they know the answer indicative of exceptional general knowledge ability, even though in Ross, Amabile and Steinmetz (1977) all of the students randomly assigned to this role were able to do this. In this experiment we investigate what observers *expect* the performance of the questioner to be by varying the number of questions the questioner is able to invent. The logic is that the number of questions at which the questioner is judged to be average should reflect observers' expectations of average performance in this role.

Participants

Participants were recruited from the Amazon Mechanical Turk marketplace and received money for participating. One hundred and fifty participants attempted the task, but 31 participants were eliminated before data analysis for either clicking through the experiment without ever moving the response scale, or for not passing an attention check, leaving 119 participants that were included in data analysis.

Materials and Procedures

The procedures used in Experiment 1b were similar to Experiment 1a except here we varied the number of questions the questioner produced, holding the number of questions the constant was able to answer constant at three. The participant always watched all ten trials unfold. In one condition the questioner asked all ten question, but in the other two conditions the questioner was shown to be unable to produce all ten questions, and instead asked only four or only seven. Trials in which the questioner asked a question were identical to trials in Experiment 1a. Trials in which the questioner was unable to think of a questions always appeared last (the last three when the questioner asked seven questions and the last six when the questioner asked four questions). On trials where the questioner was shown to be unable to think of a question the screen read "The questioner asked: [questioner was unable to think of another question]." These trials were followed by a screen that read "The contestant responded: [no response given because no question was asked]." The stimuli used were the same 124 questions used in Experiment 1a, chosen at random for every participant. As in Experiment 1a, after seeing the ten trials the participant rated the general knowledge of the contestant and the questioner, again in a randomly determined order using the same response scale described previously.

Results

Observers were indeed sensitive to the different number of questions the questioner was able to invent – rating questioner as having less general knowledge when they invented fewer questions. We again replicated the original role-conferred advantage pattern – when the questioner asked ten questions and the contestant answered three, observers judged the questioner to be in 62nd percentile and to have superior general knowledge ($t(73) = .73$, $p = .46$). In this case, observers now rated the contestant to be in the 40th percentile in general knowledge (perhaps because here they answered one question fewer than empirically demonstrated in our data and the original paper). When the questioner is able to invent only four questions observers rate their general knowledge to be below average in the 48th percentile (but statistically indistinguishable from average; $t(38) = 16.8$, $p < .001$), and when they invent seven questions they are seen as being slightly above average in the 53rd percentile and to have general knowledge that is on par with the contestant $t(40) = 15.01$, $p < .001$),

Experiment 1c

From Experiment 1b it seems that observers expect questioners to be able to invent between four and seven questions. But, according to Ross, Amabile and Steinmetz (1977) the Stanford students who participated in the original study were able to invent ten appropriate trivia questions in the time allotted. So, from this is seems that the role-conferred advantage arises because we expect the average contestants to invent fewer questions than they can, and when we witness more we infer they are above average. However, surprisingly, in the stimulus generation phase (described above) the 19 UCSD students tasked with inventing ten trivia questions in accordance with the criterion specified in Ross, Amabile and Steinmetz (1977) were actually *not* all able to complete the task. Of the 190 questions expected, only 174 questions were produced, and only 124 of these were deemed appropriate to use in Experiment 1a and 1b. Questions were eliminated if the answer provided by the writer was incorrect or if the question was ambiguous (e.g. “What is considered a popular Thai food?) or if it did not meet other criteria specified in the instructions, such as having no precise or concise answer (e.g. “why were tomatoes once called poison apples?”). On average, participants were able to invent about seven ($M = 7.15$, $sd = 2.19$) usable trivia questions in the time allotted.

From the behavior of the questioners in the original study, observers' expectations in Experiment 1b deviate quite a bit from ground-truth (they expect players to be able to invent between four and seven questions, when player can invent ten) however, if we compare observers expectations to performance of our participants in the stimulus generation-phase the deviation between expectations and reality is not as extreme (they expect players to invent between four and seven questions, and players can invent roughly seven). Because this behavior is not consistent with the behavior reported in the original study we wanted to replicate the stimulus generation phase. Forty-three UCSD undergraduates participated in exchange for partial course credit, and procedures were the same as described previously with the exception that students were not actually required to read their questions aloud after they had produced them. Students were *told* that they would be required to do this, though, in order to ensure that these participants had as much incentive to produce ten high-quality questions as previously.

Results

As before, this group of 43 students were not all able to produce ten acceptable trivia questions in the fifteen minutes allotted. Figure 3.3 shows the number of questions students were able to produce before any filtering of question content was performed. These numbers were obtained by simply counting the number of blanks in which participants wrote *any words*. Of the 43 students 21 of them filled in fewer than 10 blanks, and 22 students filled in all of the blanks. To systematically determine which questions were actually appropriate trivia questions and answers, a research assistant who was blind to the purpose of the study filtered the questions. She believed she was choosing questions that would be appropriate to use as trivia question in a future study. The number of questions that participants generated that met criteria is shown in Figure 3.3. Twenty-five questions were eliminated because the answer provided was incorrect (e.g. the United States became a country in 1769; San Diego is the second largest city in the U.S), twenty-four questions were eliminated because they did not have a clearly defined answer (e.g. “What type of bias can be found in our society?”; “How many hours of sleep is best?”), and nine were eliminated for having no precise or concise answer (e.g. “Define test-retest reliability.”) After this filtering, participants, on average, were able to invent 7.39 appropriate trivia questions ($M = 7.39, sd = 1.90$). This is very similar to the number of trivia questions invented by the 19 participants in the stimulus generation phase ($M = 7.15, sd = 2.19$).

There are several possible reasons that participant behavior in this task differs between UCSD students and students in the original study. It could be that Stanford students are better than UCSD students at inventing trivia questions. However, there is no description of question-vetting in the original paper, so another possibility is that Stanford students were also not able to invent ten questions that met the standards, but this was not discovered. If the researchers didn’t check the answers the questioners provided, they might not realize that the indicated answer was incorrect (e.g., there are 206 bones in the human body, rather than the provided answer: 260). Since the main issue in our study was that participants asked questions with multiple possible answers, or the answer provided was incorrect, and this was not discovered until systematically vetting the questions using the internet, it seems plausible that this could be the source of the discrepancy (since answers to arbitrary trivia questions are considerably harder to find without modern search engines).

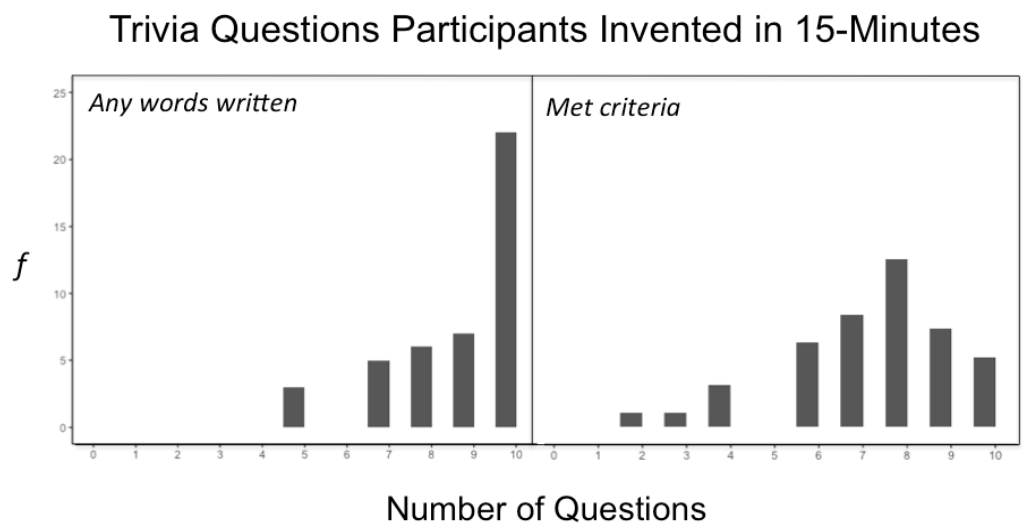


Figure 3.3. The number of question participants invented. The left panel shows the number questions for which participants wrote anything. Of the 43 participants, 22 students attempted all ten questions, while the other 21 students attempted fewer than ten questions. The panel on the right shows the number of questions that were complete and accurate trivia questions. Participants, on average, wrote 7.39 appropriate questions on average.

Discussion

In Experiment 1a we showed that observers are, at least coarsely, sensitive to different role constraints. We showed that observers' ratings of the contestant were sensitive to changes in the contestant's behavior, but also that these inferences were consistent with ground-truth. On average, people can answer about four trivia questions in this scenario, and contestants shown to answer four questions were inferred to have average general knowledge. This is consistent with the theory that when we know the impact a situation should have on behavior, we adjust our judgments appropriately.

Because observers have direct experience with the contestants' role constraints (as they are aware of how few of the questions they are able to answer themselves), they discount based on this role appropriately. Further, there is some awareness that the questioner is in an advantageous position. When the contestant is able to answer eight trivia questions they are rated to have general knowledge as exceptional as the questioner's, and when they can answer all ten questions they are rated to have general knowledge that is more exceptional. Thus, when participants observe that both the questioner and the contestant can answer all of the questions correctly they judge the contestant as more knowledgeable, which suggests that they are aware that the questioner has an easier job. However, the fact that average performance is seen as exceptional suggests that observers may underestimate just how much easier their job is. Indeed, in Experiment 1b we showed that observers infer the questioner to be average when he can invent between four and seven questions. Although this could suggest that observer underestimate average performance, we do not believe this is the case. Inventing ten challenging trivia questions in fifteen minutes actually is exceptional behavior – on average our questioners were only able to invent between 7.15 (stimulus generation phase) and 7.39 (Experiment 1c). In this light, inferring that a questioner who invents ten trivia questions is above average is consistent with actual performance on this task. That said, observers do not seem to be perfectly calibrated to the strength of this situation as they judge a questioner who invents between four and seven questions to be average, which underestimates the actual average performance. Perhaps the most revealing condition to examine is the condition that most closely captures ground-truth: when the observer invents seven questions (only slightly less than true average performance) and the contestant correctly answers three of these (also only slightly less than true average performance). In this case the role-conferred advantage disappears: the two players are judged to be equally above average on general knowledge – which isn't necessarily an over-attribution of general knowledge considering that it was college students – a group likely to have greater general knowledge than the average person – who could produce about seven questions on average.

Experiments 2

In Experiment 1 observers' ratings showed sensitivity to the number of questions the questioner invented and the contestant answered. But these are very blunt indicators of general knowledge, and if judgments are made by rationally estimating the general knowledge that the questioner and contestant are drawing on then these judgments should be sensitive to details of the questions, not merely how many were asked or answered correctly. We next tested if observers use more nuanced information, such as question difficulty (Experiment 2a), whether the questions content is diagnostic of general knowledge (Experiment 2b), and the breath of topics reflected in the questions (Experiment 2c) to inform their judgments about the players' general knowledge.

Experiment 2a

To manipulate question difficulty, we selected the ten most and least difficult questions of the 124 trivia questions produced in the stimuli creation phase of Experiment 1 (Appendix B). The easy question set consisted of the ten questions that had been answered correctly by participants most often (>90% of the time) and the difficult question set consisted of the ten questions that had been answered correctly least often (~1.8% of the time). To put this in context, questions from the stimulus generation phase were answered correctly 36.5% of the time on average.

The procedures were very similar to Experiment 1a, except that here we varied between subjects whether the questioner asked ten easy questions or ten hard questions. Additionally, on contestant-incorrect trials, the incorrect response read “I don’t know” rather than an incorrect guess. Participants were recruited from the Amazon Mechanical Turk marketplace and received money for participating. One hundred and sixty-seven people attempted the task, but 45 participants were eliminated before data analysis for either clicking through the experiment without ever moving the response scale, or for not passing an attention check, leaving 122 participants that were included in data analysis.

Observers were sensitive to question difficulty when judging the general knowledge of questioners and contestants. When questions were harder, observers rated general knowledge as greater for both questioner and contestant ($F(1,120) = 21.45$, $p < .001$). Moreover, ratings showed the classic “role-conferred advantage” pattern of inference: questioner rated as more knowledgeable than contestant ($F(1,120) = 24.28$, $p < .001$). Both questioner and contestant were rated to be above average in the difficult question condition, but only the questioner was rated above average in the easy-question condition.

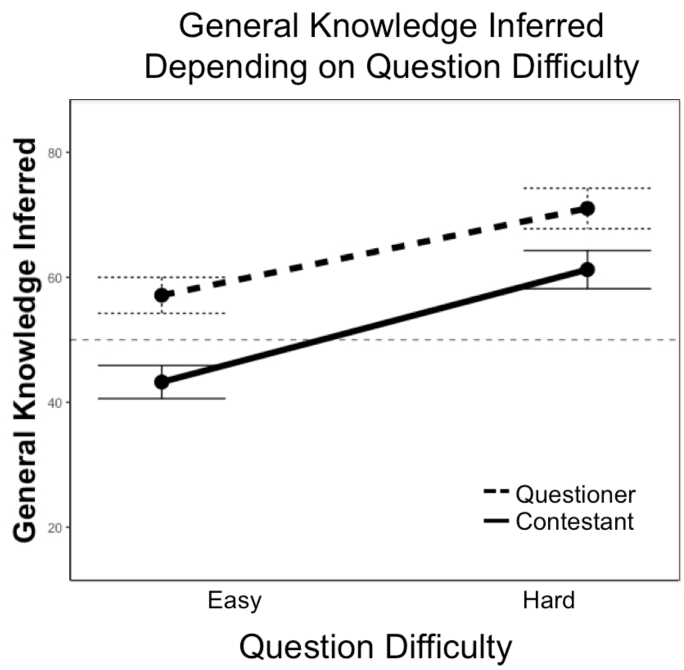


Figure 3.4. The rated general knowledge of the contestant and questioner as a function of whether easy or hard questions were asked. The questioner is always rated higher than the contestant, but both players are rated higher when the questions are harder.

Experiment 2b

Using question difficulty to infer the players' general knowledge is only a rational thing for observers to do in cases when difficulty is in fact indicative of general knowledge. A person who has a lot of general knowledge will be able to ask public knowledge questions that few people know the answer to. However, being able to ask questions that are hard to answer is not necessarily indicative of having a wealth of general knowledge. Even a questioner with exceptionally little general knowledge can stump a Jeopardy champion by asking them to report the questioner's middle name. Here we tested if observers distinguish between question difficulty and question diagnosticity when making judgments.

We varied question difficulty (hard v. easy) and how diagnostic the questions were of general knowledge (diagnostic v. non-diagnostic) between subjects. The procedures were very similar to those used in Experiment 1, except that we used four distinct sets of ten questions, designed by the experimenter (Appendix C). The hard-diagnostic question set comprised challenging questions about collective knowledge (“What was the code name for the Battle of Normandy?”), the hard-non-diagnostic question set included difficult to answer questions about facts that one would expect to be specific to someone’s life experience (“I am from Chadron Nebraska, what high school did I attend?”); the easy-diagnostic questions had answers that most people know, but revealed deeper knowledge on the part of the questioner (“Liberty Island was named Bedloe's island, before the construction of which NY monument?”) and the easy-non-diagnostic questions were easy, and revealed no particular knowledge on the part of the questioner (“What is the day after Saturday?”). Participants were recruited from the Amazon Mechanical Turk marketplace and received money for participating. One hundred and seventy-six people attempted the task, but eighteen participants were eliminated before data analysis for either clicking through the experiment without ever moving the response scale, or for not passing an attention check, leaving one hundred fifty-eight participants that were included in data analysis.

In the case where questions were diagnostic, we observed the classic “role-conferred advantage” pattern in which the questioner was rated higher than the contestant, and both the players were rated as having more general knowledge when the question set was difficult compared to easy, however question difficulty differentially effected players when questions were non-diagnostic ($F(1,154) = 18.09$, $p < .001$). In the non-diagnostic case there was a similar pattern of judgments about contestant: they were rated as having more general knowledge when they answered harder questions compared to easier questions, $t(86) = 7.8$, p <.001When the contestant could not answer easy questions they were judged to have less general knowledge than the questioner who invented these easy questions ($t(106) = 4.7$, p <.001), and both players were inferred to be below average. However, a contestant who could answer some of the private-knowledge (hard non-diagnostic) questions was rated above average.

The question of primary interest was whether observers made more sophisticated inferences than simply using question difficulty as proxy for general knowledge, and this does seem to be the case. How much general knowledge observers rated a questioner to have didn't depend on how difficult the questions were when the questions were non-diagnostic – a questioner who asked questions that were not diagnostic of general knowledge was inferred to be below average, regardless of whether their questions were easy or hard to answer ($t(86) =.20$, $p = .84$). Likewise, a questioner who asked diagnostic easy questions was inferred to have higher general knowledge than a questioner who asked questions of the same difficulty that were not diagnostic of general knowledge ($t(89) = 2.9$, $p = .004$). This means that observers are not simply evaluating knowledge based on the difficulty of the questions, but are instead using the questions to estimate the contestant's knowledge.

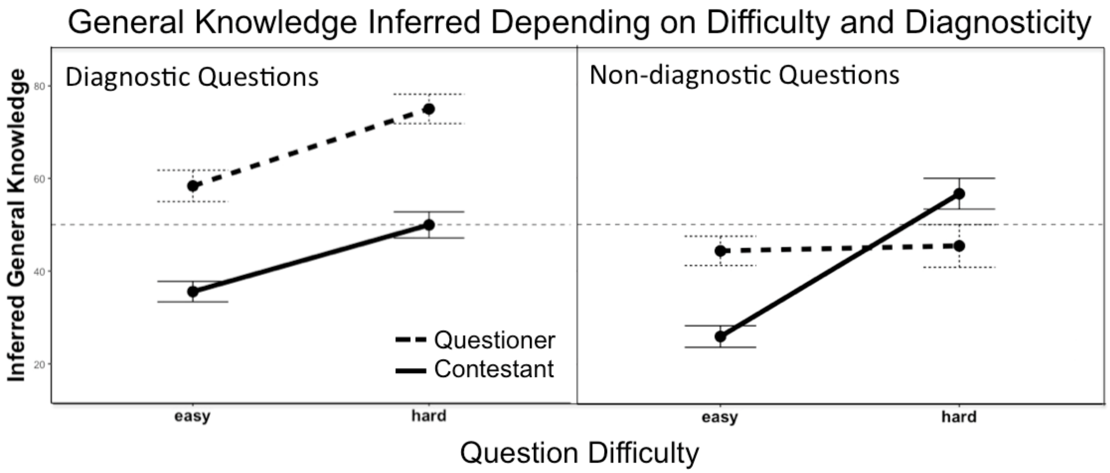


Figure 3.5. The rated general knowledge of the contestant and questioner as depending on question difficulty and whether questions were diagnostic (left panel) or non-diagnostic (right panel). When questions are diagnostic, the questioner is always rated higher than the contestant, but both players are rated higher when the questions are harder. When questions are non-diagnostic the contestant is rated below average and below the questioner, but she is rated above average and above the questioner when she answers hard questions that are non-diagnostic. However, when the questions are non-diagnostic the question difficulty does not affect ratings of the questioner, and she is rated below average regardless of question difficulty.

Experiment 2c

In the previous experiment observers realized that difficult questions about unique private knowledge are not reliable indicators of a questioners' general knowledge. In this experiment we test if observers are sensitive to more subtle indicators that questions are being drawn from idiosyncratic knowledge. Being able to invent ten challenging trivia questions on a wide variety of topics should be more indicative of general knowledge than inventing ten challenging questions on a specific topic (which might suggest specific expertise or a pet-hobby). Here we vary between-subjects whether the questioner is shown to ask questions from a variety of topics or on only a specific topic. The stimuli consisted of ten sets of ten trivia questions designed by the experimenters. Each question set was on a particular topic (e.g. Harry Potter, Modern Politics, see Appendix D). In the *specific* condition the questioner asked all ten questions from one question sets, which was chosen at random. In the *breadth* condition the questioner asked one question selected at random from each of the ten separate question sets. Contestants always answered four of these questions correctly. Participants were recruited from the Amazon Mechanical Turk marketplace and received money for participating. Eighty people attempted the task, but ten participants were eliminated before data analysis for either clicking through the experiment without ever moving the response scale, or for not passing an attention check, leaving 70 participants that were included in data analysis

Results

We find that observes are sensitive to this more subtle cue that the questioner is relying on idiosyncratic knowledge when inventing questions, and observing a "conferred advantage" depends on question set, $F(1,68) = 4.3$, $p = .043$. When the questions are drawn from ten different categories the questioner, as usual, is rated above average (64^{th} percentile) on general knowledge that is superior to the contestants (49^{th} percentile; $t(28) = 2.8$, $p < .001$). However, in the *specific* question condition this discrepancy vanishes –- the questioner is no longer inferred to be above average, or to have more general knowledge than the contestant, $t(40) = .28$, $p = .78$.

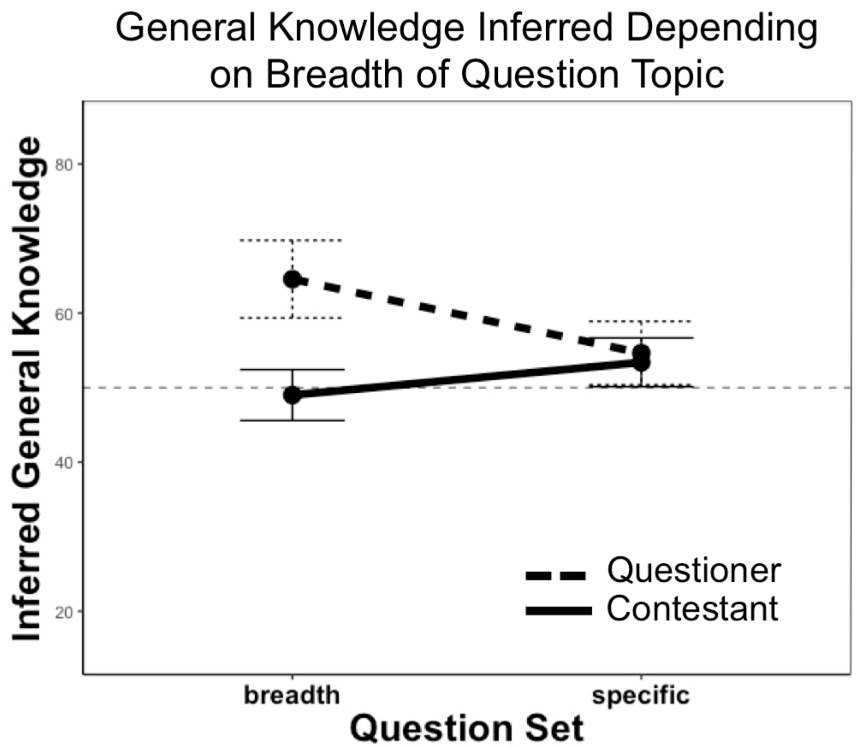


Figure 3.6. We replicate the classic role-conferred advantage paradigm when the questioner asks questions from ten different categories (breadth condition). However, the questioner and contestant are rated equal on general knowledge when the questioner asks questions that are entirely from the same category.

Discussion

Experiments 2a, 2b and 2c showed that observers are sensitive to progressively more nuanced features of question content, suggesting they are using this information in a rational way to estimate that questioner and contestants' knowledge. First, observers seemed to use question difficulty to inform their judgments, rating both the questioner and contestant as having more general knowledge when harder questions were asked/answered and having less general knowledge when easier questions were asked/answered. From Experiment 2b and 2c it seems that observers use more than superficial qualities of difficulty (e.g. can they themselves answer the question), but other features diagnostic of general knowledge, such as whether the questions revealed private or public knowledge of whether they were drawn from a narrow or wide range of topics. When questions were hard but non-diagnostic, or drawn from only a narrow range of topics observers did not use these questions as positive indicators of general knowledge ability.

General Discussion

Experiments 1a, 1b and 1c reveal that observers are not only aware of the different role-constraints, but they are also pretty well calibrated to the difficulty of each role. In Experiment 1a we showed that observers rate a contestant who answers eight or more questions equal to or superior to the questioner, suggesting that observers are not just comparing how many questions each player answered correctly regardless of the role constraints. Further, observers' judgments suggest that they have a good sense of average contestant behavior -- participants, on average, can answer about four questions, and this is what observers judge to be average. Surprisingly, observers also have a good sense of average questioner behavior. Initially it seemed like observer underestimate how easier the questioners' role was since observers judged average behavior to be above average. However, it turns out asking ten questions is above average behavior – in Experiment 1c we found that most people actually can only invent about seven questions.

In addition to being reasonably calibrated to and using social roles when making judgments, observers also seem to utilize other information present in the exchange. Experiments 2a, 2b and 2c suggest that observer use nuanced information present in question content as cues to infer general knowledge. They are sensitive to question difficulty, question diagnosticity, and the breadth of the question set when reasoning about questioner and contestant's behavior.

Overall, it seems that the judgment pattern in the *role-conferred advantage* is not evidence of non-normative behavior. Observers seem to rationally consider role-constraint as well as other available information that is relevant to the judgment they are asked. The one error observers seem to make is slightly (but not as much as previously thought) underestimating the number of questions the questioner can invent. Previously we have shown (Chapter 2) that people use perceived situation strength optimally. When presented with the task of tossing a coin into a bucket, for example, observers have a good sense of the strength of this situation and inferences about disposition match those we expect from an ideal observer. However, this does not preclude the possibility that people are not perfectly calibrated to all situations, particularly ones they have never experienced. Do we really expect observers to be know exactly how many questions the average person can invent in 15-minutes? It seems likely that some situations seem a bit *easier* than they are (e.g. saying no to an experimenter in a white lab coat who asks you to shock someone), and in these cases average performance will seem unimpressive. Likewise, there are some situations that are seem *harder* than they look (e.g. inventing trivia questions), and average performance will seem impressive. Imperfect estimation of strange situations is not necessarily evidence of non-normative reasoning.

References

Heider, F. *The psychology of interpersonal relations.* New York: Wiley, 1958.

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.

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.

Appendix A

Question	Correct Answer	Incorrect Answer
Do male or female sea horses carry the fertilized eggs?	*male*	*female*
How many bones are there in the adult human body?	*206*	*260*
How many keys are there on a standard modern piano?	*88*	*110*
in the ancient Greek language, what does the word "agape" mean?	*unconditional love*	*Wine*
In where did the word "Mafia" originate?	*Sicily, Italy*	*South America*
In which state was the ice cream brand "Ben & Jerry's" created?	*Vermont*	*California*
In which year did World War I end?	*1945*	*1845*
Shakespeare's sonnets were all addressed to which two people?	*young man and dark lady*	*mother & lover*
The sinking of which ship is notoriously associated with reports that it's crew were eaten by sharks?	*USS Indianapolis*	*Titanic*
What are the animals that represent the Democrat and Republic parties in U.S., respectively?	*Donkey & Elephant*	*Bears*
What are the names of the two visual sensory receptors?	*Rods and Cones*	*neurons and axons*
What are the three primary colors?	*Red, Yellow, Blue*	*red, yellow, green*
What disorder is caused by the loss of Dopamine?	*Parkinson's*	*Depression*
What does DNA stand for?	*Deoxyribose nucleic acid*	*Deasocherite Nucleus A*
What does the last pair of chromosomes code for?	*sex*	*eye color*
What is the capital of California?	*Sacramento*	*Los Angelus*
What is the capital of Spain?	*Madrid*	*Mexico*
What is the largest artery in the human body?	*aortic*	*vena cava*
What is the metropolitan area with the most population in North America?	*Mexico City*	*New York City*
What is the metropolitan area with the most population in the United States?	*New York*	*Los Angelus*
What is the name for the prion disease discovered in sheep?	*Scrapie*	*Alzheimer's disease*
What is the name of the extreme right political party in France, led by Marie Le Pen?	*The National Front*	*Free French*
What is the name of the largest island in the Republic of Fiji?	*Viti Levu*	*Fiji*
What is the name of the last U.S. state when they are listed in alphabetical order?	*Wyoming*	*Washington*
What is the space between neurons called?	*Synapse*	*neural junction*
What is the state bird of California?	*Quail*	*California Condor*
What is the term for a group of lions?	*a pride*	*cluster*
What is the term for the discriminatory laws against Chinese immigrants in the 1800s?	*Chinese Exclusion Act*	*Restriction immigration laws*
What kind of triangle has a different length on each side?	*scalene*	*Isometric*
What structure on a single neuron transmits electrical signals?	*Axon*	*conductor*
What tear was the attraction Twilight Zone Tower of Terror at Disneyland closed?	*2017*	*2016*
What the second most widely spoken language in the world?	*Spanish*	*English*
Which American president involved in the Watergate scandal?	*Richard Nixon*	*Kennedy*
Which country has the most land area?	*Russia*	*China*
Which direction does the sun rise from?	*East*	*West*

Which film did Hitchcock regard as his singular masterpiece?	*Vertigo*	*Carrie*
Which part of the brain is responsible for emotion?	*Amygdala*	*Thalamus*
Which river carved out the Grand Canyon?	*Colorado*	*San Juaquin*
Which state became the 50th state of the United States?	*Hawaii*	*Alaska*
Which state is located above Kansas?	*Nebraska*	*Michigan*
Which Swedish actress was barred entry into the US when it was discovered she was having a marital affair with her director?	*Ingrid Bergman*	*Katrina Zach*
Who found the uncertainty principle in quantum physics?	*Heisenberg*	*Isaac Newton*
Who invented the lightbulb?	*Thomas Edison*	*Franklin*
Who is the 44th president of the United States?	*Barack Obama*	*Bush*
Who is the inventor of Alcoholics anonymous & the 12-step program?	*Bill Wilson*	*Kay Patrick*
Who made the sculpture Moses in Italy during the Renaissance?	*Michelangelo*	*Da Vinci*
Who was the first African American woman to headline the music festival Coachella?	*Beyoncé*	*Zedd*
Who was the winner of the 2010 soccer world cup?	*Spain*	*Germany*
Who's the Greek goddess of marriage?	*Hera*	*Aphrodite*
American sign language evolved from what spoken language?	*French*	*English*
Cingular is now owned by what cell-phone provider?	*AT & T*	*Verizon*
How many books are in the series "A Series of Unfortunate Events"?	*thirteen*	*eight*
How many season of the show "Friends" are there?	*ten*	*twelve*
How many wives di Henry VIII	*six*	*eight*
Immediately before going to the Lakers, Stephen Nash played for what NBA team?	*Phoenix Suns*	*Clippers*
In "Big Bang Theory" Raj's red-headed girlfriend's name was?	*Emily*	*Amy*
In operant conditioning, rewarding behavior by removing a negative stimulus is called?	*negative reinforcement*	*operant conditioning*
In the book "Of Mice and Men" which character dies at the hand of his best friend?	*Lennie*	*George*
In the cartoon "The Simpsons" Homer is known for working in a…	*Nuclear Power Plant*	*Office*
In the show "Breaking Bad" what was the prop used to convey meth actually made of?	*Rock Candy*	*Cement Dust*
In what city is Richard Nixon buried?	*Yorba Linda, CA*	*North Dakota*
On what continent does it never rain?	*Antarctica*	*Australia*
On what planet was Luke Skywalker raised?	*Tattooing*	*Vega*
On which hemisphere are Penguins found?	*Southern*	*North Pole*
The Terracotta Warrior are from which province in China?	*Xi'an*	*Shih*
What are brain cells that send signals called?	*neurons*	*neurotransmitters*
What candy do the characters of "13 going on 30" enjoy eating?	*Razzles*	*Hot Tamales*
What contiguous U.S. state does not participate in daylight savings?	*Arizona*	*Florida*
What device was used to execute the French during the French Revolution?	*Guillotine*	*Cannon*
What form of government did Hawaii have before joining the union?	*Monarchy*	*Oligarchy*
What Harry Potter word is now in the Oxford English dictionary?	*Muggle*	*Butter Beer*
What is the 10th largest city in the United States?	*San Jose, CA*	*Sacramento*
What is the 4th most populated country in the world?	*Indonesia*	*U.S.*

What is the capital of Brazil?	*Brasília*	*Rio De Janerio*
What is the capital of China?	*Beijing*	*Hong Kong*
What is the capital of Cuba?	*Havana*	*La Habana*
What is the capital of South Korea?	*Seoul*	*Ho Chi Minh*
What is the capital of Uruguay?	*Montevideo*	*Uruguay*
What is the capital of Washington state?	*Olympia*	*Seattle*
What is the fourth farthest planet from the sun?	*Mars*	*Uranus*
What is the largest organ in the human body?	*skin*	*stomach*
What is the largest type of bear?	*Polar Bear*	*Grizzly*
What is the largest U.S. state?	*Alaska*	*California*
What is the main drummer for the Beatles?	*Ringo Star*	*Leonard*
What is the name of Sherlock Holmes sister in the British TV show?	*Eurus*	*Lucy*
What is the name of the Mayan god of rain?	*Chaac*	*Owexyloti*
What is the pen name of J.K. Rowling uses to write detective novels?	*Robert Galbraith*	*Joe*
what is the plastic end of a shoelace called?	*an aglet*	*tensile*
What is the state bird of New Mexico?	*Roadrunner*	*Humming Bird*
What is the structure that connect the two brain hemispheres?	*Corpus Colosseum*	*Cortex*
What medication is prescribed for stage-fright?	*beta-blockers*	*marijuana*
What month does winter start in?	*December*	*10*
What part of the body does an Ophthalmologist treat?	*the eye*	*tongue*
What was the first Capitol of California?	*San Jose*	*Sacramento*
What were the years of the American Civil War?	*1861-1865*	*1843-1850*
What's the name of the style of the painting "Starry Night"?	*post-impressionist*	*surrealism*
Where is the deepest point on each located?	*Mariana Trench*	*Santa Fe Trench*
Where is the tallest building in the world?	*Dubai*	*New York City*
Which album sold the most in the year 2015?	*25 by Adele*	*If You're Reading this it's too Late*
Which color is the absence of light?	*black*	*white*
Which country use to rule Hong Kong?	*Britain*	*China*
Which franchise hold the most NBA championships?	*Celtics*	*Lakers*
Which lobe in the brain is responsible for hearing?	*Temporal*	*Occipital*
Which lobe in the brain is the most important for visual function?	*Occipital*	*Parietal*
Which NBA team holds the record for most NBA finals appearances?	*Los Angelus Lakers*	*Spurs*
Which psychologist experimented on dogs and was involved in classical conditioning?	*Pavlov*	*Watson*
Which state has the largest population in the U.S.?	*California*	*Texas*
Which state is the biggest in the U.S.?	*Alaska*	*Florida*
Which two character always come and do health inspections in "Bob's Burgers"?	*Hugo and Ron*	*Linda's ex husband*
Who are the two main protagonists in the TV show Gilmore Girls?	*Lorelai and Rory*	*Jane and Wanda*
Who created the first plane?	*Wright Brothers*	*brothers*

Who holds the record for the most career points scored in the NBA?	*Kareem Abdul-Jabbar*	*Michael Jordan*
Who invented PCR?	*Kary Mullis*	*Newton*
Who is known as the All Father in Norse Mythology	*Odin*	*Thor*
Who is the Greek God of war?	*Ares*	*Apollo*
Who is the Greek god of war?	*Ares*	*Apollo*
Who is the Greek God of wine?	*Dionysus*	*Bachus*
Who is the Greek God of wisdom	*Athena*	*Hera*
Who is the Greek Goddess of love?	*Aphrodite*	*Venus*
Who is the writer of the "Importance of Being Earnest?	*Oscar Wilde*	*Charles Dickens*
Who was the lead singer for the and Queen?	*Freddie Mercury*	*Prince*
Who was the only president to serve more than 2 terms?	*Franklin Roosevelt*	*George Washington*
Who was the president during the Watergate scandal?	*Richard Nixon*	*Reagan*
Who was the third president of the United States?	*Thomas Jefferson*	*Andrew Jackson*

Appendix B

Easy		Difficult	
Question	Correct Answer	Question	Correct Answer
What is the capital of Japan?	*Tokyo*	*The sinking of which ship is notoriously associated with reports that it's crew were eaten by sharks?*	*USS Indianapolis*
What is the capital of China?	*Beijing*	*What is the name of the extreme right political party in France, led by Marie Le Pen*	*The National Fron*
Who painted "Starry Night"	*Vincent van Gogh*	*What is the pen name of J.K. Rowling uses to write detective novels?*	*Robert Galbraith*
What is the chemical formula for water?	*H20*	*Who invented Polymerase chain reaction technique?*	*Kary Mullis*
Who wrote the drama Macbeth and Romeo and Juliet?	*Shakespeare*	*Who is the writer of the "Importance of Being Earnest?*	*Oscar Wilde*
From the game Mario Brothers, who was Mario's brother?	*Luigi*	*Which state is located above Kansas?*	*Nebraska*
Which country recently left the EU?	*Great Britain*	*What is the name of the Mayan god of rain?*	*Chaac*
who voices Dory in "Finding Dory"?	*Ellen DeGeneres*	*American sign language evolved from what spoken language?*	*French*
In which year did World War II end?	*1945*	*What is the name of the largest island in the Republic of Fiji?*	*Viti Levu*

Appendix C

Diagnostic (Smart)

Difficult		Easy	
Question	Correct Answer	Question	Correct Answer
What was the code name for the Battle of Normandy?	*Operation Overlord*	*Which day of the week comes from the Latin Saturn dies, or 'day of Saturn'?*	*Saturday*
What is the oldest active ballpark in Major League Baseball?	*Fenway Park*	*In 1964 Phil knight and Bill Bowerman founded Blue Ribbon Sports, a company that would turn into the biggest athletic brain in the world, today know as what?*	*Nike*
What is the term for visible traces of an earlier painting beneath newer artwork on a canvas?	*Pentimento*	*What yellow fruit is curved because it grows toward the sun?*	*Banana*
What is the independent record label on which nirvana released their first album?	*Sub Pop*	*What is the acronym do we use for Light Amplification by Stimulated Emission of Radiation?*	*Laser*
What was the name of the character played by Audrey Heburn in Breakfast at Tiffany's	*Holly Golightly*	*Julius Pringle is mustached logo of which brand of potato chip?*	*Pringles*
What is the name of the legislature of Israel?	*Knesset*	*Which cartoon characters'' full first name is Scoobert Doo?*	*Scooby Doo*
Who is the Russian chemist who published the first version of the periodic table?	*Dmitri Mendeleev*	*Which international coffee company is named after a character in Moby Dick?*	*Starbucks*
The word 'yacht' comes from the Dutch word that means….?	*Hunter*	*What herbivorous marsupial 'bears' from Australia do not drink water?*	*Koala*
The logo for the video game company Atari has a name. What is it?	*Fuji*	*What sports beverage was first developed for the Florida Gators football players?*	*Gatorade*
What is the name of the main character in the children's books by Astrid Lindgren?	*Pipi Longstocking*	*Liberty Island was named Bedloe's island, before the construction of which NY monument?*	*Statue of Liberty*

Non-diagnostic (Dumb)

Easy: Question	Easy: Correct Answer	Difficult: Question	Difficult: Correct Answer
When driving in the United States what color light means stop?	*Red*	*I am from Chadron Nebraska, what high school did I attend?*	*Chadron Senior High school*
What is the day after Saturday?	*Sunday*	*What was my high school mascot?*	*Cardinal*
What is the biggest coffee chain in the world?	*Starbucks*	*Who was the principal of my high school?*	*Jerry Mack*
Who is the current president of the United States?	*Donald Trump*	*What is the closest interstate to Chadron, Nebraska?*	*interstate 90*
What is the name of the player that throws the ball for the batter in baseball?	*Pitcher*	*What county is Chadron, Nebraska in?*	*Dawes County*
What animal is said to have nine lives?	*Cat*	*How many public swimming pools are in Chadron, Nebraska?*	*Two*
How many cards are in a standard deck of playing cards?	*52*	*What is the zip code in Chadron, Nebraska?*	*69337*
What language do they speak in Mexico?	*Spanish*	*What is the mascot of Chadron State College?*	*Elmo Eagle*
What is the name for a person who goes to outer space?	*Astronaut*	*What street is the Wal-Mart on in Chadron, Nebraska?*	*Linden St.*
What is the name of a person who does not eat meat?	*Vegetarian*	*What is the name of the only museum in Chadron, Nebraska?*	*Museum of Fur Trade*

Appendix D

Question	Correct Answer
In the 4th Harry Potter book, what is the name of the sweet that Dudley eats when the Weasley's visit Privet drive?	*Ton-tongue toffee*
In the game quidditch in Harry Potter, how many points is the Golden Snitch worth?	*150*
In Harry Potter, who drives the Knight Bus?	*Stan Shunpike*
In Harry Potter, where is the location of the Order of the Phoenix headquarters?	*12 Grimmauld place*
In the book Prisoner of Azkaban, what item of clothing does Harry trick Mr. Malfoy into giving Dobby to free him?	*A sock*
In the book the Goblet of Fire, what type of dragon does Harry have to get past	*Hungarian Horntail*
In the Harry potter series, what is Dumbledore's sister's name?	*Ariana*
In Harry potter, who was the best man at Lily and James' wedding?	*Serious Black*
What is the name of Voldemort's snake in the Harry Potter series?	*Nagini*
In Harry Potter, what was Tom Riddles mother's maiden name?	*Guant*
In the Harry Potter series, Hermione's parents do what for a living?	*Dentists*
In World War II, what were the two opposing military alliances called?	*Axis and Allies*
What year did World War II begin?	*1939*
What was the name of aerial branch of the German military during WWII?	*Luftwaffe*
What was the name of Hitler's autobiography?	*Mein Kampf*
What was the name of the plane that dropped the first atomic bomb on Hiroshima?	*Enola Gay*
What was the name of the research project that produced the first nuclear weapons during WWII?	*The Manhattan Project*
What was the code name for the Battle of Normandy?	*Operation Overlord*
What was the name of a suicide attack pilots from the Empire of Japan during WWII?	*Kamikaze*
In a Nazi concentration camp, a pink triangle identified a prisoner as what?	*Homosexual*
What English code-breaker created a machine that helped decrypts Nazi messages?	*Alan Turing*
What does the term Blitzkrieg mean in English?	*Lightening War*

A substance that has a pH lower than 7 is considered what?	*Acidic*
What is element #1 on the periodic table?	*Hydrogen*
What is the chemical element with abbreviation Fe?	*Iron*
The unreactive gases such as Helium, Neon and Krypton are known as what?	*Noble gases*
Who is the Russian chemist who published the first version of the periodic table?	*Demintri Mendeleev*
What is the only metal that is liquid at room temperature?	*mercury*
Does water shrink or expand when it freezes?	*expands*
Dry ice is the solid form of what?	*carbon dioxide*
What is the heaviest man-made element called?	*Ununoctium*
What is the heaviest naturally occurring element that we know of?	*Uranium*
What family of elements is in the left-most column of the Periodic Table?	*Alkali metals*
What artist is best known for a painting of his mother?	*Whistler*
What art movement was Yoko Ono associated with during the 1960s?	*Fluxus*
What is the name of the painting of a man in a bowler hat with a green apple covering his face?	*Son of Man*
What painter was married to the Mexican artist Diego Rivera?	*Frida Kahlo*
Who painted 'Sunday Afternoon on the Island of La Grande Jatte'?	*Georges Seurat*
What painted 'The Scream'?	*Edvard Munch*
What artist is best known for abstract expressionist drip paintings?	*Jackson Pollock*
Pablo Picasso co-founded what art movement?	*Cubism*
What art movement literally means the style of wild beasts?	*Fauvism*
What musical instrument often appears in Picasso paintings?	*a guitar*
What is the term for visible traces of an earlier painting beneath newer artwork on a canvas?	*pentimento*

In CrossFit, what does WOD stand for?	*Work out of the day*
In CrossFit, what is the name of the benchmark workout consisting of 95lb Thrusters and pull-ups?	*Fran*
What shoe brand is associated with CrossFit?	*Reebok*
People who do CrossFit call their gym a what?	*A box*
In what city was the first CrossFit gym?	*Santa Cruz, CA*
Critics of CrossFit have argued that it can lead to developing what disorder, that causes breakdown of muscles?	*Rhabdomyolysis*
In what season are the CrossFit games held?	*Summer*
What was the first year the CrossFit games were held?	*2007*
Who were the founders of CrossFit?'	*Greg Glassman and Lauren Jenai*
In CrossFit, what is the pull-up variation called when you use momentum to get your chin above the bar?	*Kipping*
What a CrossFit athlete uses the prescribed weight and reps in a work out, what is this known as?	*Rx*

Who was Nirvana's original drummer?	*Aaron Burckhard*
What is Kurt Cobain's middle name?	*Donald*
What was the independent record label on which Nirvana released their first album?	*Sub pop*
Where was city and state was the band Nirvana from?	*Aberdeen, Washington*
Who was Kurt Cobain's married to?	*Courtney Love*
What genre of music is the band Nirvana?	*Grunge*
What is the name of Nirvana's best-selling album?	*Nevermind*
What year did Kurt Cobain die?	*1994*
What was the name of Nirvana's bassist?	*Krist Novoselci*
What was Nirvana's debut studio album?	*Beach*
What is the name of the Nirvana album that has a transparent anatomical manikin with angel wings on the cover?	*In Utero*

Who played Lawrence of Arabia in the film of the same name?	*Peter O'Toole*
Who directed the film 'Citizen Kane'?	*Orson Welles*
Who directed the film 'The Seventh Seal'?	*Ingmar Bergamn*
What Stanley Kubrick movie is based on an Arthur C. Clarke novel?	*2001: A Space Odyssey*
What was the name of the character played by Audrey Hepburn in 'Breakfast at Tiffany's'?	*Holly Golightly*
What movie is the line 'I'm going to make him an offer he can't refuse' from?	*The Godfather*
What 1942 movie stars Humphrey Bogart and Ingrid Bergman?	*Casablanca*
What movie is the line 'Here's looking at you kid' from?	*Casablanca*
In what year was the first Academy Award ceremony held?	*1929*
Who won best actor in 20016?	*Leonardo DiCaprio*
What film won best picture in 2016?	*Spotlight*

In 2015 the United States re-established diplomatic relations with which country?	*Cuba*
Who is the current prime minister of Britain?	*Theresa May*
What is the name of the far-right political party in France led by Marie Le Pen?	*The National Front*
The Knesset is the legislature of which country?	*Israel*
Who is the current supreme leader of North Korea?	*Kim Jong Un*
What is the name of the Colombian Marxist guerilla group?	*FARC*
Which two countries impeached their presidents in 2016?	*South Korea and Brazil*
Who is the current president of Syria?	*Bashar al-Assad*
The rules for exiting the EU are contained in article 50 of which treaty?	*Treaty of Lisbon*
In October 2016 Iraqi and Kurdish troops launched an offensive to reclaimed what city?	*Mosul*
What political party does Austrian president Alexander van der Bellen belong to?	*Green*

In sailing, what is it called when you zig-zag to move upwind?	*beating*
In sailing, turning the boar toward the eye of the wind is called what?	*tacking*
What is it called when a sailboat is traveling roughly perpendicular to the wind?	*reaching*
What is the technical name for a traditional modern yacht?	*Bermuda sloop*
What is the name of the piston system of a sailboat used to control the shape of the sail?	*Boom vang*
What is the name of the round-the-world solo yacht race, sailed non-stop?	*Vendee Glove*
the term 'in irons' means that you are sailing into...	*the wind*
What is the term for the permanent structure to which a ship is secured?	*a mooring*
When looking forward toward the bow of a ship, what is the term for right?	*starboard*
What is the name of the straight piece of wood or metal which fits into the head of the rudder and is used to steer a boar?	*a tiller*
The word 'yacht' comes from the Dutch word that means...?	*hunter*

What team did Babe Ruth play for before joining the Boston Reds Socks?	*Baltimore Orioles*
Who has played the most consecutive games of baseball, breaking Lou Gehrig's record?	*Cal Ripken, Jr.*
What is the oldest active ballpark in Major League Baseball?	*Fenway Park*
What baseball great is known for funny expressions such as 'It ain't over till it's over.'?	*Yogi Berra*
What are the name of the two leagues that Major league baseball is divided into?	*The National League and The American League*
In baseball, what does 'the top' of the inning mean?	*The first half*
In baseball, what is the area where the relief pitcher warms up called?	*the Bull Pen*
In baseball, how many runs does a grand slam drive in?	*four*
Who was the first African American to play Major League Baseball?	*Jackie Robinson*
Who won the world series in 2016?	*Chicago Cubs*
What is the name of the baseball team in Kansas City?	*Royals*

Chapter 3, in full, is currently being prepared for submission for publication of the material. Walker, Drew E.; Vul, Edward. The dissertation author was the primary investigator and author of this

GENERAL DISCUSSION

The social information available to us at any given moment is ambiguous at best. Yet, remarkably we resolve this ambiguity efficiently and arrive at judgments and perceptions that are largely accurate. The fact that our people-perceptions do not always perfectly align with reality is not necessarily evidence that our social reasoning processes are flawed. In the last three chapters I have shown that although social judgments are not always without error, many of the systematic errors we make can arise from being globally rational agents operating with incomplete information.

In Chapter One I show that we have a bias to perceive faces in a group to be more attractive than they are, not because our information-processing is flawed, but rather as a consequence of our visual system's strategy to circumvent resource limitations and capitalizing on redundancy in visual scenes. The bulk of the previous work on ensemble coding has addressed simple features such as line-orientation and dot size (e.g., Ariely, 2001; Chong & Treisman, 2003), and findings in Chapter One are at the forefront of a growing body of work beginning to illuminate how statistical properties influence information processing in attention and perception for high-level visual features. Findings that high-level features like attractiveness are subject to set averaging have had implications for understanding the mechanisms of visual processing (Huang, 2015). Such findings have also influenced work on how fast and efficient people-perception can affect real-life social organization and interaction (e.g., Phillips, et al., 2015), and has been used to investigate the face processing deficits in individuals with Autism.

In Chapter Two and Three I proposed that two of the most well known biases in social psychology can be explained by a rational inference framework. However, there are two criteria we might use to evaluate whether behavior in a given domain is rational: whether the reasoning process is internally consistent (coherence) and whether the judgments themselves are empirically accurate (correspondence; Dunwoody, 2009). The *Fundamental Attribution Error* as classically presented is a claim about coherence. The fact that attributions are made in a direction consistent with behavior despite known situational pressure is suggested as evidence of a flawed process*:* we make dispositional attributions without properly discounting for the known influence of the situation (e.g., Gilbert & Malone, 1995).

These paradigms don't measure or compare behavior to ground-truth as observers in these studies make attributions from an invented behavior about a fake actor, without an empirical comparison to average behavior. Therefore, the studies underlying this phenomenon can't make a claim about *correspondence.* Thus, a sufficient challenge to the FAE is demonstrating coherence of human judgments in these paradigms, which we do in Chapter Two. We show that making attributions in a direction consistent with behavior when there is also known situational pressure is not in fact evidence of a flawed process*,* but is rather the rational judgment of a probabilistic observer. In our modeling of the six classic study results, our argument that these judgments show coherence requires only two extremely conservative assumptions about observer sensitivity to situations: that they know the direction of the situational influence (being asked to write a pro-essay imposes pressure to write a pro-essay as opposed to an anti-essay) and that the influence of these situations is not deterministically strong (See Chapter Two Appendix

A and B for further elaboration). Still stronger evidence of coherence in social reasoning are Experiment 1 and 2 in Chapter Two, in which we obtain empirical perceptions of situation strengths and show that participants judgments can be explained as optimal inference using these situation strengths (and vice versa: judgments about the influence of situations is explained as optimal inference using known dispositions). These results provide strong evidence the social judgments show coherence, a satisfactory challenge to the argument put forward in the original literature.

Further, although it might not be readily apparent, the original role-conferred advantage study (Ross, Amabile, & Steinmetz, 1977) itself provides some evidence that social judgments are coherent. In the original study, as well as in new data presented in Chapter Three, observers accurately judge the general knowledge of the contestant, suggesting that when observers have a good sense of the pressure imposed (having directly experienced the challenge of hearing questions they cannot answer) they make accurate judgments about the contestants' general knowledge , providing further evidence of coherence. Taken in isolation, an observer's lack of correspondence with regard to the questioner – their tendency to judge questioners as above average – could be explained as an issue of miscalibration to the situation or a lack of coherence. Because Chapter Two and Three provide converging evidence that social judgments are internally consistent, it seemed more plausible this tendency arises from coherent reasoning about a situation that was underestimated. However, this was not what caused observers to rate the questioner above average. Surprisingly, we instead found evidence that observers are fairly well calibrated to expected performance in this role: in the condition that most closely approximated how both contestants and questioners actually do perform, a "role-conferred advantage" bias disappeared, suggesting that observers are reasoning rationally here both with regard to coherence and correspondence. Similarly, in Chapter Two, when we used empirically inspired essay compliance rate from Sherman (1980), human inferences matched those of an ideal observer, suggesting that human estimates of the strengths of these types of situations are at least in the right ballpark.

Taken together, results from Chapter Two and Three strongly suggest that social reasoning is rational in terms of coherence, but also provides some evidence that – at least in these paradigms – there is correspondence in human social judgments. Of course, demonstrating evidence of calibration in these paradigms certainly cannot answer the question of how calibrated people are to situations in general. As we know from some classic findings in psychology, we are not always accurate in such estimates: We are astonished by Milgram's (1965) demonstration that an authority figure can compel 65% of people to ostensibly shock another person to death, and wildly underestimate how many people would do such a thing in this situation. But people may be bad at estimating the pressure that these particular situations impose, and so research into biases in social judgment might pick out and thus over-represent the situations that people have a particularly bad understanding of. This makes it difficult to get an unbiased estimate of how good observers are at evaluating the strength of everyday situations.

Although we have shown calibration to one situation classically thought to be misestimated, calibration to the strength of situations more generally is an open question. If observers coherently use a rational inference framework but systematically underestimate the strength of situations than this could also result in consistently and erroneously judging dispositions to have an inflated influence on behavior. From the classic FAE literature this is indeed a possibility, however, this proposal is difficult to reconcile with results that suggest that when we make attributions from known dispositions, we show a tendency to "over attribute" to situations (Quattrone, 1982; Chapter 2, Expt. 2). The more parsimonious explanation is that both results are produced by the same mechanism: estimating an unknown variable using unbiased probabilistic inference.

The view proposed in Chapter Two that social attribution can be better understood by using probabilistic reasoning has implications beyond explaining biased behavior in classic paradigms. Bayesian probabilistic inference has been widely used as a framework to understand how humans are able to intelligently resolve uncertainty in domains such cognitive psychology and sensation and perception, but has been only sparsely applied to help us understand how humans reason about other humans (e.g. Baker, Saxe, & Tenenbaum, 2009), This framework could be applied much more broadly to inform our understanding of human social reasoning, and could help elucidate very consequential social phenomena, for example, stereotypes. Stereotypes result from perceived correlations between observable demographic features and unobservable character traits (e.g. gender and intelligence; race and criminality) that yield an undesirable influence on human judgments of individuals. Most research in this domain focuses on reducing prejudice, but not much attention has been given to where in the cognitive inference process these errors arise. Bayesian cognitive modeling could be used a formal description of this process to help determine, for example, if stereotypes arise because people overestimate the correlations in the world, if they rely on the incorrect conditional probabilities, or if they overweight socially charged attributes (e.g. gender) compared to other features (e.g. occupation). Furthermore, research on prejudice often presumes that stereotypes are emotional rather than statistical biases and consequently misses one straightforward approach to reducing prejudice. Insofar as prejudices arise because we rely on perceived statistics in the world, greater information about an individual will reduce how much influence demographic traits effect our judgments of others. Ignoring the statistical nature of stereotypes can even lead to counterproductive interventions. For example, the ban-the-box policy, which prevented employers from asking about criminal record on initial job applications, actually decreases the probability that low-skilled black and Hispanic men would be hired (Doleac & Hansen, 2016) since employers now discriminate more widely against groups that are more likely to have a criminal record.

In addition to demonstrating the usefulness of applying a Bayesian probabilistic inference framework to social questions, Chapter Two and Three are also examples of how investigations of social phenomena can benefit from parametrically manipulating and measuring continuous variable. This approach can put behavior into a broader context. By observing judgments made across a spectrum of situational influences, and action strengths, for example, we were able to see that inference patterns are sensible, which is a feature that was obscured when narrower samples of behavior were taken in the classic FAE and role-conferred advantage literature (e.g. Gilbert & Malone, 1995). Overall, the quantitative and computational techniques applied here can produce reliable and precise predictions of behavioral measures and allow for theoretically motivated advances in social psychology.

References

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological science*, *12*(2), 157-162.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329-349.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision research*, *43*(4), 393-404.

Doleac, J., & Hansen, B. (2016). *Does "ban the box" help or hurt low-skilled workers? Statistical discrimination and employment outcomes when criminal histories are hidden*. Working Paper.

Dunwoody, P. T. (2009). Introduction to the special issue: Coherence and correspondence in judgment and decision making. *Judgment and Decision Making*, *4*(2), 113.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological bulletin*, *117*(1), 21.

Huang, L. (2015). Statistical properties demand as much attention as object features. *PloS one*, *10*(8), e0131191.

Milgram, S. (1965). *Obedience*. Penn State University.

Phillips, L. T., Weisbuch, M., & Ambady, N. (2014). People perception: Social vision of groups and consequences for organizing and interacting. *Research in Organizational Behavior*, *34*, 101-127.