Big Brazen Lies

Sebastian Deri

“An oddity of swindling as a fine art is its lack of perspective. From the smallest to the largest, most swindlers live in a two-dimensional world of outright falsehood, reveling in the reflection of their own conceit. Some of their “sure-fire” schemes may fail because they oversell them, yet in contrast, some of the most stupendous swindles succeed simply because it would be fantastic to suppose anyone would have the nerve to put over such an outrageous fakery.”

Walter Gibson

The Fine Art of Swindling (p. 11)

“Holmes came out swinging almost from the start. That was no surprise: we had expected her to be combative. What we hadn’t fully anticipated was her willingness to tell bald-face lies in a public forum.”

John Carreyrou

Bad Blood p. 277 (on Elizabeth Holmes and Theranos fraud)

“An editor at a rival publishing house told me, ‘I totally fell for it. After all, who would fabricate such a story?’”

Ian Parker

writing in The New Yorker, on writer Dan Mallory’ (A Suspense Novelists Trail of Deceptions)

“Surely no one would tell a lie of such enormity.”

Randall Kimpton

in Sophie Hannah’s *Closed Casket* (a novel)

“All this was inspired by the principle—which is quite true within itself—that in the big lie there is always a certain force of credibility; because the broad masses of a nation are always more easily corrupted in the deeper strata of their emotional nature than consciously or voluntarily; and thus in the primitive simplicity of their minds they more readily fall victims to the big lie than the small lie, since they themselves often tell small lies in little matters but would be ashamed to resort to large-scale falsehoods.

It would never come into their heads to fabricate colossal untruths, and they would not believe that others could have the impudence to distort the truth so infamously. Even though the facts which prove this to be so may be brought clearly to their minds, they will still doubt and waver and will continue to think that there may be some other explanation. For the grossly impudent lie always leaves traces behind it, even after it has been nailed down, a fact which is known to all expert liars in this world and to all who conspire together in the art of lying.”

Adolph Hitler

Mein Kampf (as translated by James Murphy)

A Shepherd-boy beside a stream
“The Wolf, the Wolf,” was wont to scream,
And when the Villagers appeared,
He'd laugh and call them silly-eared.
A Wolf at last came down the steep—
“The Wolf, the Wolf—my legs, my sheep!”
The creature had a jolly feast,
Quite undisturbed, on boy and beast.

For none believes the liar, forsooth,
Even when the liar speaks the truth.

William Ellery Leonard

1: Question

"We assume that the more a new piece of information fits with people's prior knowledge and experience, the more likely they will be to learn/believe it. The more discrepant a piece of information (the less it fits, the more surprising, or extreme it seems, etc.), the less likely people will be to learn/believe it. Please review why this is the common assumption, drawing from whatever literatures you think are the most relevant. What evidence is there for it? And, are there any exceptions to this rule? Are there ever cases in which someone is more likely to learn/believe an unexpected (versus expected) piece of new information? In other words, was Hitler correct in arguing that a big lie can sometimes be more believable and effective than a small one?"

- Dr. Melissa Ferguson

2: Thoughts on Question

I don't know how to approach this question. And I've tried multiple times. Either my thoughts turn so hopelessly abstract that I begin to doubt that I am writing anything that meaningfully corresponds to reality. Or I render insights that seem true but stem from the ordinary experiences of being a living person and are so obvious that they hardly need to be laid out in some "scientific" essay or treatise or whatever this is supposed to be.

This is my most recent shot. And I still don't have a solution. But here's what I got.

The prompt above can be chopped into two questions. The first question (highlighted in blue) concerns the extent to which we believe new information, depending on how surprising we find it. The scope of such a question is, hypothetically, enormous—swallowing into it research from areas of study from learning, to persuasion and beyond. The second question (highlighted in green) is narrower. Are big lies ever more believable than small lies? This question seems simpler and more straightforward (although, we'll quickly run into problems when we realize that the "size" of a lie is more a figure of speech than something that we can (or need) quantify). In any case, it's the lesser of two evils so let's just go with it.

word count (from beginning of next page, to right before references start): 12,129

3: Core Response

Dan Mallory is an author and former editor. His book, *The Woman in the Window*, authored under the penname A.J. Finn, sold millions of copies and gained him acclaim. He is also a prolific liar, as revealed by a recent profile in the New Yorker (Parker, 2019). His lies are extensive and shocking. Here's one. Some time in the winter of 2012, while working as an executive editor at William Morrow, he stopped coming into his office. Some time later, his professional acquaintances received an email (almost definitely written by Dan, although purportedly from his brother Jake) claiming he's been battling cancer and would soon undergo surgery for a tumor in his spine. This helped explain his previous absences and served to excuse his coming absences. It was also a lie. Those not already suspicious of him for other reasons, believed it.

I start with this story to ground us in the concrete. This is a lie that was really told, really worked (until it didn't), and it is as good of an example of a "big lie" as I can think of. I also think all the essential psychology of a big lie can be pried out of it and examples like it. Including why such big lies may often be believed, and perhaps more so than "small" lies.

Imagine this example. A friend invites you over for dinner, but you feel like staying in. Rather than telling the truth, you decide to lie. Here's two lies you could tell. Lie one: you are just not feeling well and should really spend the rest of the evening resting up. Lie two: your brother committed suicide and you are in the midst of all the emotional and logistical complications of that, so dinner won't really work out tonight. (This is another lie that Dan told—that his brother committed suicide.) Both lies would probably suffice for getting you out of dinner, but the second seems "bigger", in some sense, than the first. It's also probably a safe bet that your friend is more likely to wholeheartedly believe the bigger lie, about the suicide. People often say they're not feeling well to get out of engagements, and your friend, who is tuned into this fact, might at least be suspicious about your conveniently timed illness. He might not call you out on it, but he's likely to have his doubts. On the other hand, he might be shocked to hear that your brother committed suicide, but it seems probable that he'd pretty unequivocally believe you.

After all, who would lie about such a thing? Indeed, it was in response to Mallory's lie about the tumor in his spine that one his colleagues is quoted as saying "I totally fell for it. After all, who would fabricate such a story?" While spinal cancer or suicide are somewhat uncommon, it also seems extremely unlikely, perhaps even more unlikely, that someone would lie about such things. This is the essence of why "big lies" can work. Some lies seems so egregious—by which I mean something like morally abhorrent—that we think people are extremely unlikely to tell them. So when people do (infrequently) tell them, they are believed. That's basically it.

This also raises two new interesting questions. (1) Why don't people tell these big, morally abhorrent lies more often? (2) And what exactly makes them so morally offensive?

Let's take the second question first. We must begin by noting that lies don't occur in a vacuum. They occur in the context of relationships with others and the communities within which those relationships exist. Such big lies violate the trust, that forms the foundation of these relationships and communities, and they do so in an extreme way.

In any relationship or community, we can imagine there exists somewhere in people's minds something like an awareness of different channels of communication, that vary in their seriousness and thus the extent to which they must be kept free of "noise". Barroom banter and casual storytelling are the types of talk that occur at low importance channels. We allow these channels to be noisy, because the stakes are low. And there is an understanding that the dishonesty, lying and exaggeration that occurs on these channels is to be confined there, or to channels of similar levels of importance, with little leakage upwards. Thus, lying that occurs on these channels is not taken to be indicative of lying at more serious, higher fidelity bandwidths. It's why we don't necessarily expect that someone who exaggerates about their high school sports accolades will cheat on their taxes. Communication on more important channels is understood to be more serious, and there is something like an implicit agreement that the signal on these channels is to be kept to a higher level of fidelity. This understanding and compliance with it is why we need not doubt when someone says "help, I'm drowning", screams "fire" in the proverbial theater, or warns us that they saw a rattlesnake on our hiking trail. It's understood that we don't lie about such things. Because of this understanding, we can trust that communication about important matters is true. (If people speak falsely on these channels, it is only because they are mistaken, not deceptive intent.)

Lying on these channels exhibits extreme disregard for the people with which one has relationships and one's community. It indicates a willingness to engage in behavior, that if committed in mass, would render everyone liable to being seriously taken advantage of. Doctors could falsely diagnose us with serious diseases to receive bigger payouts, and police officers could detain us for made up crimes. Because of this, such lies are severe moral offenses, on par with things like physical violence. Indeed a person's relationships and community standing is quite possibly more damaged by lying about things like having brain cancer and or a relative committing suicide than throwing a punch in the heat of the moment.

People who lie on such high fidelity channels indicate they can't be trusted at all. They make themselves liable to be labeled as extremely self-interested, reckless, and unpredictable, easily willing and able to disregard social mores and moral injunctions, to which the "rest of us" abide. They may be seen as morally reprehensible. Such people are to be avoided, excluded, and kept at the fringes of society.

This helps now answer the first question, about why people don't often tell such lies. They are a big risk to take. And in most cases not worth it. In almost all situations, whatever benefit you might get from telling a big lie is likely out-weighted by the cost that such a lie would incur if it were to be found out. Essentially, such lies are only worth it when their benefit outweighs their potential cost of losing an entire relationship, one's standing in the community, and indeed an indefinite number of future relationships, given that word travels. And this is not often, perhaps never.

Note that my claim about big lies is that they are likely to be believed. Being believed, however, is not the same as never being found out. Assuming someone telling a big lie isn't already a known liar, doesn't contradict anything that the person they are speaking already knows to be incontrovertibly true, and hasn't raised suspicions for any other reason, their big lies are likely to be believed on their face—for the reasons above. But, regardless of whether they are believed or not, big lies misrepresent reality, often more severely than typical, smaller lies. Eventually, contradictions are likely to arise, making these big lies susceptible to being exposed.

A wonderful and illustrative example comes from Clifford Irving, who in 1971 claimed to have been authorized by Howard Hughes, a public icon who by the 1960s had become a total recluse, to help write Hughes's biography. Indeed, Irving produced an entire manuscript, replete with details about Hughes' life. On this basis, Irving was paid $765,000 by the publishing company Mcgraw-Hill for the story rights (Jackman, 2003, p. 37-38). While his big lie was initially believed, it eventually came into more and more in conflict with reality as Irving and his story gained attention. Most damningly, Howard Hughes briefly emerged from hermitage and called reporters to tell them he'd never met, seen, or heard of any man named Clifford Irving, until he'd heard about the news of him claiming to have been commissioned to write his biography.

This case highlights an interesting feature that seems to characterize many big lies. They are concrete. (Irving, for example, forged detailed letters from "Your truly" Howard Hughes and supplied florid anecdotes about his experiences interviewing Hughes, such as Hughes offering his research assistant organic prunes from a cellophane bag; Jackman, 2003, p. 40-41). This concreteness lends them credulity. And also makes them susceptible to being eventually found out.

That's the bulk of what I have to say about big lies and why they work. If it were acceptable to end my response there I would. Since it's not, I'll spend the rest of the essay fleshing out these ideas, some other related ones, and pointing to relevant empirical work.

PART 1: Definitions

4: What is a lie exactly?

Lying

Since we're going to be talking about lying a lot, let's start by being a clearer on what lying is. The literature in philosophy on lying is most useful here. A common and precise definition of lying in the philosophical literature is the act of making a statement that the speaker knows or believes to be false with the intent that the receiver believe it to be true (Bok, 1999, p. 13; Mahon, 2016). Thus, a communicated statement must meet three conditions to be considered a lie: (1) it must be false, (2) the speaker must know this, (3) the speaker must have an intent to deceive the person to whom the communication is directed.

This rules out from consideration a few types of communications which might otherwise uncritically be considered lies. Most importantly, it makes clear that lying is more than simply making a false statement. No matter how outrageous their claims, conspiracy theorists, cut leaders, religious and political zealots—if they are sincere in their beliefs—are not liars. To be lying, a speaker must believe that the statement they are making is false. Imagine two cult leaders make the same claim—that a commune must be set up to prepare for the end of the world, which will occur on May 14, 2019. One cult leader is a schizophrenic and genuinely believes this to be true, while the other does not and hopes to use the commune as a way to achieve personal glory and wealth. Despite making the same claim, only the second cult leader has lied. Thus, when trying to reason about lying, we must account for the speaker's beliefs.

Further, we see that for a statement to be a lie, we must take into consideration a speaker's intent. Consider the same statement—"Your dress looks great that way, honey"—made by Lucas to his wife, Sonia, in two different scenarios. In both scenarios he believes the statement to be false; that is, he does not actually think the dress looks great. However, in one scenario, Sonia has spilled marinara sauce on her dress when rushing through the kitchen. And when speaking, Lucas intends to make light of the situation and assumes that Sonia shares the common belief that dresses do not look good with globs of marina sauce on them, and that she will understand his sarcastic tone to indicate his lack of seriousness. In the other scenario, it is their wedding night. And when speaking, Lucas does indeed intend to implant the false belief that he thinks her dress looks lovely, in his wife's head, in order to avoid hurting her feelings. Only in the latter case is he lying. This further curtails the cases that constitute lies.

Deceiving

This essay will concern lying, but note there is a more general notion of intentionally misleading in interpersonal interactions that often comes up when talking and thinking about lying—deception. The definition of this concept is also worth clarifying as the literal word is sometimes (confusingly) used interchangeably with lying, despite the two not meaning the same thing.

The dictionary definition of deceiving—"to cause to accept as true or valid what is false or invalid" ("deceive," Merriam Webster Dictionary, 2019)—matches the standard philosophical definition closely (Mahon, 2016), although the latter adds the requirement that, again, the perpetrator must have an intent to mislead.

This conception of deceiving differs from lying into principal ways. First, deception does not require communication in the form of a statement. For example, a person would be engaging in deception, but not lying, if they were to intentionally cause others to believe they are wealthy (when they are not) by wearing a fake Rolex to party. Second, deceiving implies an accomplished act. That is, if one deceives another it implies that a false belief has successfully been implanted in the mind of another. In lying, there is only the necessity of intent.

5: What makes a lie “big”?

Big Brazen Lies

When people use the term “big lie” they are understood. And, as the quotes that open this essay reveal, people are comfortable talking about things like the “enormity” of a lie. But, what exactly makes some lies “bigger” than others, and what lies are appropriately labeled “big lies”?

I am not aware of any attempt by philosophers to nail down the definition of a “big lie”. In economics, some have attempted to formalize the various dimensions that might characterize the “size” of a lie, which I will review at the end of the section. Ultimately, however, I want to stick with notions that hem close to real life examples of lies that people would or have identified as “big lies”. I see these as discrete types of lies, of the kind I discussed in Section 3 (“Core Response”), where various examples of big lies were given (e.g. lying about having spinal cancer to excuse one’s absences from work). If a definition is required, it would be something like lies that are morally offensive, because they are misrepresentations about important matters; lies which we don’t expect because of some implicit social understanding that there exist something like high fidelity “channels” of communication where only the most serious of claims are made and so should be kept free of noise; lies, which if discovered are likely to lead to very harsh and negative inferences about the character of those telling them, risking their relationships and standing in the community. For clarity, we can call these “big brazen lies”.

It is worth distinguishing these “big brazen lies” from other notions of the “bigness” of a lie.

Big Unlikely Lies

Big brazen lies often make claims about the existence or occurrence of something uncommon. For example, spinal cancer and suicide are relatively uncommon. This might lead us to define the “bigness” of a lie as corresponding to something like the extent to which a liar’s statement seems unlikely or improbable (from the perspective of a naïve receiver to whom it is told, who is judging something like the probably of content communicated in the statement, and not for example, the probability that someone falsely or truthfully make such a claim). Indeed, this is the type of definition that the prompt seems to implicitly espouse (e.g. “the more discrepant a piece of information (the less it fits, the more surprising, or extreme it seems, etc.), the less likely people will be to learn/believe it”).

We could, of course, define big lies in this way, as “big unlikely lies”. But I think it is an impoverished notion of what we intend to study. This is because a lie is more than a simple verbal statement or abstract proposition—like “I won the lottery”—to which we can assign a

probability. A lie is a social communicative act taking place between two parties, a speaker and a receiver, who have different states of knowledge about the true state of reality (e.g. only the speaker knows whether or not they actually won the lottery) and certain expectations about norms and rules governing how those different states of knowledge are to be bridged. To define the size of a lie as the extent to which a statement seems unlikely to be true is to concern ourselves mostly with things like abstract probability judgments, only considering the content of the statement itself (e.g. "how likely is it that someone would win the lottery"?) from the perspective of person to whom the lie is told. It does not account for the knowledge of speaker, nor the societal expectations surrounding their behavior (e.g. how likely are people to lie about winning the lottery?).

Note some of the consequences of failing to consider the knowledge of the liar in our definition of the bigness of a lie. Consider, for example, two lies that Jason—who is 5'1—might tell about his height on his OkCupid dating profile. He might overstate his height slightly, by only one inch, claiming to be 5'2. Or he might overstate much more dramatically, claiming to be 5'9, a full eight inches taller than he actually is. If we define the bigness of a lie as the extent to which a the content of a statement seems improbable to a naïve receiver, we need not arrive at this conclusion. Consider the perspective of Jasmine, who is browsing Jason's profile, and does not know Jason's real height. All she sees is either the profile of someone claiming to be 5'2 or the profile of someone claiming to be 5'9. If she is judging the probability of the abstract event that "a man on a dating website is 5'2" and "a man on a dating website is 5'9", the former event should seem more improbable—and thus the "bigger" lie under this definition—as there are simply fewer men who are 5'2 than 5'9. Arriving at this conclusion about which lie is "bigger" suggests something about our definition is off.

Big Lies As Large Deviations from Reality

A definition of a big lie which fixes the previous problem is to define the bigness of a lie in terms of the extent to which the liar has distorted reality. The more a lie deviates from reality, the bigger a lie it is. Thus, Jason claiming to be 5'9 when he is 5'1 is a bigger lie than him claiming to be 5'2 when he is 5'1 because the former misrepresents reality, his height, to a greater extent. Unlike the earlier definition which focused on the situation from the perspective of the receiver, this definition takes the perspective of the speaker.

This is not a bad definition of the "size" of a lie. But it runs into its own problems. It is easy enough to think about how much reality has been distorted, when we are thinking of lies that have a clearly quantitative angle, like a person's height. But it becomes harder to think about and compare lies that don't have a clearly quantitative bent. For example, which of the two lies is a larger deviation from reality: (1) claiming that you took a trip to Guatemala (when you once had a layover in an airport there on a familiar vacation) or (2) saying that your brother committed suicide last week (when you brother died in a motorcycle accident years ago)? Or is it

a bigger distortion of reality to (1) deny the Holocaust occurred (despite believing it happened) or to deny the paternity of an "illegitimate" child (that you clearly know is yours).

It's not entirely clear how to make these determination. Lies are essentially counterfactuals that people are trying to pass off as real. Some scholars have tried to outline what makes some counterfactuals come to mind more easily, and perhaps seem more probable than others. Such theories might also be useful in determining which lies are larger deviations from reality than others. Kahneman & Miller (1986) *Norm Theory*, for example, tries to explain how people compare "reality to its alternatives". They outline a theory whereby different counterfactuals arise as post hoc constructions, where alterations to various attributes of a might come more easily to mind than others. Interesting examples are provided. But it's less clear how to one might apply their theory to novel cases, to delineate between two existing counterfactuals in terms of which deviates more from reality. Likewise, they reference Hofstadter's (1979, 1983) seemingly relevant notion of "slippability", which concerns how some counterfactuals are more "natural than others" and how some "attributes are particularly resistant to change", p. 142-143). But again, it's harder to use this notion to delineate between the unreality of two counterfactuals, aside from to note that your judgment of which is a bigger deviation from reality is likely based in some intuitive judgment of "slippability". This also highlight the fact that these theories attempt to explain and describe *how* people come up with and judge various counterfactuals, whichdiffer in the extent to which they deviate from reality. They are not necessarily normative or prescriptive theories about which counterfactuals *are* rightly judged as bigger deviations from reality (if such a notion even makes sense).

As we do with "big brazen lies", it is probably easier to distinguish—or at least compare and argue about—these lies on a moral basis. Which is morally worse: lying about a trip you never took or the cause of your brother's death? The latter seems more "wrong". What's worse, denying the Holocaust or your fatherhood of a child? This is perhaps harder to answer, but at least we have a basis for argument and comparison. What were the liar's intentions in each case? In what ways are people affected by the lie?

Big Consequential Lies

Thinking about lies in moral terms also leads us to realize these previous definitions have so far ignored another important feature of lies—their consequences. Indeed, one might measure the "bigness" of a lie by its consequences.

Some lies that might be otherwise hard to distinguish in terms of "size" are easier to distinguish in light of their consequences. Consider a somewhat artificial and morbid case. A deranged and suicidal man tosses a policeman a quarter. He says that the policeman should flip it. If it lands on tails he will shoot himself, and if it lands on heads he'll check into a psychiatric care unit. Consider two possible lies. (1) The coin lands on tails, but the officers lies and says it landed on heads, leading the man checks himself into psychiatric care. (2) The coin lands on

heads, but the officer lies and says it landed on tails, leading the man to shoot himself. A coin landing on heads or tails is equally probable, so its hard to distinguish these lies on the mere probability of the content being communicated (heads or tails). The distortions of reality also seem equal (claiming that a heads flip landed on tails seems about the same distortion of reality as claiming that a tails flip landed on heads). The two lies are more easily distinguished on their consequences. One lie causes a man to kill himself, while the other in effect saves his life. On this basis, the lie that resulted in death seems obviously bigger.

This is another useful feature of lies to point out. But it doesn't seem to capture the full story. For example, consider the original case of Dan Mallory lying about having brain cancer. Is it the consequences per se that make this lie so obvious "big"? He likely violated the trust of others and perhaps inconveniences them or caused them extra work or confusion, however the most grievous consequences are probably those he brought on himself once the lie was found out. His career and reputation were destroyed. And this consequence depending on him being found out. It seems weird that the bigness of this lie should depend on whether he was found out or not. It seems equally big regardless of whether he was found out or not, suggesting that something other than pure costs and consequences is informing our tendency to label it "big".

Economic Perspectives

Joel Sobel is one of the foremost experts on lying in economics (he's authored dozens of papers and in fact teaches an entire class devoted to the "Economics of Lying and Deception"). In one paper, he and his co-authors explore factors affecting people's willingness to lie, and in doing so, also formalize several notions of the "size" of a lie (Gneezy, Kajackaite, & Sobel (2018), which corresponds some of the notions of "bigness" that we have just explored. They outline three dimensions along which the "size" of a lie might vary: (1) something like the "distance" of the lie from the truth (kind of confusingly, they call this the "outcome" dimension, and which corresponds to the notions I explored in the earlier section "Big Lies As Large Deviations from Reality"); (2) the likelihood of the lie (they call this the "likelihood dimension", and further clarify that they are referring to the "ex ante probability that the agent's report is true"; this closely tracks what I discussed in the earlier section "Big Unlikely Lies"), and finally (3) the potential payoff from the lie (the "payoff dimension", which corresponds to the "monetary gains of lying"; these have to do with the consequences of a lie, a topic I explored when discussing "Big Consequential Lies", although this explores the "other side of the coin", so to speak, the gain that can be had from a lie).

They illustrate each with an example of lying about the outcome of an n-sided die roll. Lies on the outcome dimension occur correspond to larger deviations from the actual outcome of the roll (e.g. if you rolled a 1, saying that you rolled a 6 is a bigger lie on this dimension than saying rolled a 2, because 6 is farther from 1). Variation on the likelihood dimension can be imagined, by imagining alternations to probability of various outcomes (e.g. claiming you rolled

a 10 on a fair 10-sided die is a bigger lie than saying you rolled a 6 on a fair 6-sided die, because the former outcome is “ex ante” less probable). And finally, we can imagine different payoffs assigned with the outcome of each roll (e.g. rolling a 1 gains you $10, rolling a 2 gains you $5, and all other numbers gains you $0). In this case, if you roll a 4 ($0 payout) but claim that you rolled a 1 ($10 payout), you would be telling a bigger lie than if you lied and said you rolled a 2 ($5 payout). This is a somewhat useful recapitulation of some of the concepts we explored above, when discussing what constitutes the “bigness” of a lie. Again, I would like to hem for the most part to our notion of “big brazen lies”, discussed in the first part of this section and explored in the earlier section (“Core Response”).

PART 2: Finding Empirical Undergirding

7: Trust & The Default Assumption of Truthfulness

A big part of my explanation for why big lies might work (i.e. in Section 3, "Core Response") is that they violate the trust that both forms the basis of our relationships and holds our communities together. There are many actions and expectations that form the bases of this trust. One that is particularly relevant here is the expectation we have that people will for most part communicate with us honestly[1]. I present some "empirical" evidence that such a norm exists.

A good place to start is Timothy Levine's (2014) *Truth Default-Theory*, which attempts to provide a framework for understanding human deception. The central element of this thesis, as Levine summarizes, is the "idea … that as a default, people presume without conscious reflection that others' communication is honest" (Levine, 2014, p. 381). That is, there is an expectation of honesty by default.

Lending support to this claim is the robust finding robust finding in lie detection literature of a "truth bias" (see: Bond Jr & DePaulo, 2006; Levine, Kim, Sun Park, & Hughes, 2006; Levine, Park, & McCornack, 1999; Zuckerman, DePaulo, & Rosenthal, 1981)—which is the consistent tendency people exhibit, when judging the truthfulness or untruthfulness of statements, to expect a significantly higher portion of those statements to be true. That is, after making say 100 guesses about 100 statements, most people most of the time guess there are more truths than lies. This finding holds, even when the base ratio of truths and lies is varied (Levine et al., 2006, 1999). And note this finding occurs in the context of lie detection studies, where participants are explicitly aware that they are going to encounter lies.

Why would does such a bias exist? One reasonable guess might be that people are calibrated to the speech behavior they experience in their everyday lives[2], where people for the

---

[1] As Sissela Bok (1999, p. 30) puts it "trust in some degree of veracity functions as a *foundation* of relations among human beings; when this trust shatters or wears away, institutions collapse … If there is no confidence in the truthfulness of others, is there any way to assess their fairness, their intentions to help or to harm? How, then, can they be trusted? *Whatever* matters to humans beings, trust is the atmosphere in which it thrives."

[2] Of course, people's calibration should really be related to the number of lies they are able to detect. If lying is highly prevalent, but undetected most of the time, people might be calibrated to this low rate of detection, rather than the high rate of lying. This doesn't seem like a huge problem, however, because as we will see the actual prevalence of lying seems to be pretty low. Thus, even with perfect detection, people would discover a small number of lies (assuming we believe the prevalence statistics are roughly accurate). Furthermore, even if lying were more prevalent, people still have their own lying behavior, which they can observe completely, and use as a basis to estimate the prevalence of lying among others.

most part do abide by a norm to not lie. Indeed, research which tries to estimate how often people lie in "everyday" life does seem to suggest that the answer is not very often. The most commonly referenced figure is that, on average, people lie once or twice a day. This figure comes from a diary study on college students from DePaulo, Kashy, Kirkendol, Wyer, & Epstein (1996). A more reliable estimate, however, comes from Serota, Levine, & Boster (2010), who recruit a national sample of 1,000 individuals, stratified and re-weighted by age, gender, income, and region to match U.S. demographic base rated (obtained, by the authors, from the *Current Population Survey* from the U.S. Census Bureau).

Participants in Serota et al.'s (2010) study were then given a broad definition of lying, that was also worded to be non-judgmental (e.g. "We are interested in truth and lies in people's everyday communication … some lies are told for a good reason. Some lies are selfish; other lies protect others. We are interested in all these different types of lies."). Respondents were then asked to record the number of times they lied to each of five different categories of people (family, friends, business contacts, acquaintances, and total strangers), in each of two different mediums (face-to-face, and then any type of mediated format, such as over the phone, text, or over the internet), over the last 24 hours. After summing across categories and re-weighting to match national demographics , the authors find that people on average people tell 1.7 lies per day. A figure that is not especially high if one considers how many statements we make and messages we send to family, friends, coworkers, acquaintances, and strangers in a given day. And even this average oversells the matter, because most people (59.9%) report not having lied in the last 24 hours at all.[3]

[3] Some questions might be raised as to whether any of these results these constitute a reliable estimate of lying, given that lying is scorned upon, giving people a reason to underreport their occurrence. While this cannot be ruled out, it seems unlikely to me that any underreporting is resulting in some massive concealment of a much higher rate of lying. First, these results are anonymous, nor did the respondents communicate any specifics about the nature of the lie—hardly leaving anyone exposed to be being found out. (Although, some lies, people might not even want to admit, remind themselves, or even remember).

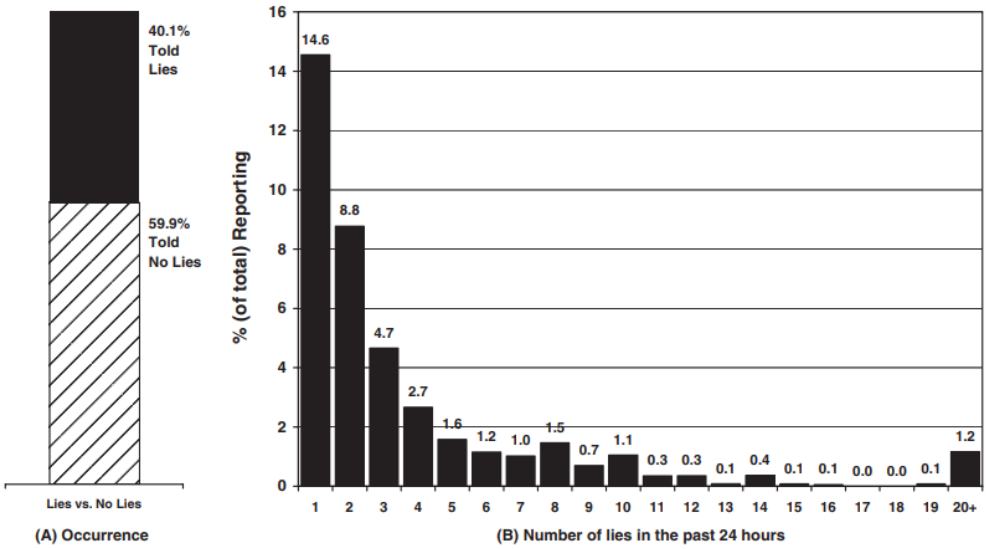


Figure 3 (A) The majority of Americans reported they did not lie in the past 24 hours (59.9%). (B) Percentage distribution by number of lies told; 32.2% told one to five lies and 7.9% reported telling six or more lies.

All of this suggests that people don't expect others to lie and that indeed most people, most of the time aren't lying.

8: When Do People Lie?

People of course do lie. And we are not naïve to this fact. However, there are certain ways that we lie. And the ways that we lie, I suspect, inform people's expectations about how and when and the ways in which people will lie. Big lies succeed in part because they violate some of these expectations.

Here, I will review empirical evidence on what I believe are some of the main factors influencing whether and how egregiously people lie. Three particularly critical factors, I believe, are the following: (1) a person's incentives and motivations, (2) considerations of harm and costs to whom the lie is told, and (3) individual differences (the existence of a few "prolific liars").

Incentives and Motives

One of the most straightforward, yet useful, studies on the factors that influence lying is another paper by Timothy Levine, titled *People Lie For a Reason* (Levine, Kim, & Hamel, 2010). The paper is in part an effort to verify the "principle of veracity", which the philosopher Sissela Bok (1999) lays out in her renowned moral treatise on the ethics of lying, *Lying: Moral Choice in Public and Private Life*. This principle states in essence that there is something like a moral imperative to speak honestly. And people should only violate this imperative for a just cause. As Bok (1999) prettily puts it "lying requires explanation, whereas truth ordinarily does not". People lie for a reason.

To show this, Levine, Kim, & Hamel (2010) present participants in a standard survey experiment with various scenarios, and in their first study, ask them to make a binary choice about whether they would lie or not. They vary, between subjects, which version of the scenario participants get—a version where there is an incentive or reason that one might lie, and a version where there is not. For example, in the first scenario, about a gift, participants are asked to imagine a romantic partner gets them a sweater for their birthday. In the scenario where they have an incentive to lie, they are told they hate the way the sweater looks. In the scenario where they don't have an incentive to lie, they are told they love the way the sweater looks. They are then asked to make a choice about what they would tell the person who got them the sweater, when that person asks if they like it. he other scenarios are similar. The next scenario, for example, asks whether participants would be honest with a friend about whether they look overweight. In the scenario where there is an incentive to lie, the friend does look overweight; while, in the scenario where there is no incentive to lie, the friend looks "clearly … trim" (Levine et al., 2010, Appendix, p. 285). The remaining scenarios are similar, asking whether participants would communicate honestly or dishonestly about: the taste of cooking, they either do or don't like; expressing interest in a second date after a first one that either went well or didn't; giving an

honest opinion about a movie to a romantic interested who they knows like the movie, if they themselves either do or don't like the movie (where their interest might also serve as the basis of a date); telling a friend the truth about whether they mailed a package which the friend had asked them to mail, depending on whether one did or did not forget to send the package.

Their findings are clear. When people have no incentive to lie (e.g. they like the gift, they don't think their friend is overweight, they like the cooking, they enjoyed the date, they like the movie, they mailed the package), they tell the truth. As can be seen below in Table 1 from the paper, in all the scenario (except the last), there is 100% honesty in the absence of any incentives to lie.[4]

Table 1 Frequency of Honest and Deceptive Messages by Motive and Situation in Study 1

	No deception motive		Deception motive			
Situation	Honest	Deceptive	Honest	Deceptive	χ^2	Φ
Gift	32 (100%)	0 (0.0%)	9 (26.5%)	25 (73.5%)	37.88*	.76
Weight	32 (100%)	0 (0.0%)	16 (47.1%)	18 (52.9%)	23.29*	.59
Cooking	34 (100%)	0 (0.0%)	4 (12.5%)	28 (87.5%)	51.67*	.88
Date	32 (100%)	0 (0.0%)	2 (5.9%)	32 (94.1%)	58.46*	.94
Movie	34 (100%)	0 (0.0%)	15 (46.9%)	17 (53.1%)	24.33*	.61
Post office	29 (90.6%)	3 (9.4%)	29 (85.3%)	5 (14.7%)	0.44	.08
Overall	189 (98.4%)	3 (1.6%)	75 (37.5%)	125 (62.5%)		

*$p < .001$.

On the other, when there are some incentives or reasons to lie (which in this study often involve avoiding hurting someone's feelings), people are much more likely to do so. For example, over 70% indicate they would lie about liking the gift (likely to make their partner feel good), and over 90% would indicate lie about liking a movie to a romantic interest when they know that person likes the movie (and, as is hinted in the scenario, might serve as a basis for a date with them).

The only scenario where rates of dishonesty are not above 50% is lying to a friend about mailing a package, when one forgot to do so. Lying, I believe, is lower here because it is the only scenario that is not a "white lie". The other scenarios only involve misrepresenting preferences, a generally benign offense with little lasting consequences.[5] Lying about delivering a package is a

[4] Why there is any lying in the post office scenario is a mystery to me, and I assume must be some misunderstand of the prompt on the part of the participants. This rises to 100% honesty as well, in later studies.

[5] Although see (Bok, 1999; Chapter V: White Lies) for an extended discussion on the ethics of telling white lies and what they can be more harmful than people give due.

more serious offence, because a friend will expect that a necessary task has been completed when it hasn't, changing their future course of actions.

(In further studies, Levine et al., 2010) vary how the responses are recorded—for example, in one study, asking participants to write out how they would respond to these scenarios and then coding these responses for whether they are lies or not. Their results unsurprisingly hold across these variations.)

Studies from economics lend even further support to the idea that people need special reason to tell a lie. There is an entire literature on what is termed "lie aversion , which explores people's general reluctance to tell lies, even when there are moderate incentives to do so (e.g. see: Gneezy et al., 2018; Gneezy, Rockenbach, & Serra-Garcia, 2013; Hurkens & Kartik, 2009; López-Pérez & Spiegelman, 2013; Lundquist, Ellingsen, Gribbe, & Johannesson, 2009). In one representative finding for example, only 26 percent of participants lied in a single-shot experiment where they could gain up to 10 euros by simply misreporting a random number they were assigned by a computer.

Consequences/costs/harm to receiver

When deciding whether to lie, people also take into account how and to what extent this lie will affect others. A nice demonstration of this comes from Uri Gneezy (2005), in the American Economic Review, in paper appropriately titled *Deception: The Role of Consequences*. He creates a one-round game in which there is a sender and a receiver. The receiver will receive a payment, depending on whether they pick "Option A" or "Option B". Option A always earns them more money, but they do not know this. Everything they know about the payments associated with these options comes from the sender. This sender's payment in the experiment is entirely contingent on what message they send to the receiver. They must pick between one of two messages: (1) the truth, which corresponds to Message A, where they tell the receiver "Option A will earn you more money than option B, or (2) a lie, which corresponds to Message B, where they tell the receiver "Option B will earn you more money than option A. (Pretests confirmed that the vast majority of receivers (78%) tended to go with the option suggested by the sender, and that the vast majority of senders (82%) expect that receivers would indeed pick the option they suggested.)

The participants incentives were then put at odds in three different treatment conditions. In treatment 1, the sender stands to gain $1 by lying, and sending a message that, if acted upon, would cause the receiver to lose out on $1. In treatment 2, the sender stands to gain $1 by lying, and sending a message that if acted upon would cause the receive to lose out on $10. In treatment 3, the sender stands to gain $10 by lying, and sending a message that, if acted upon, would cause the receiver to lose $10. The set of payouts that set into place these conflicts is shown in Table 1 below, from the paper.

TABLE 1—THE DIFFERENT PAYOFFS USED IN THE DECEPTION GAME

Treatment	Option	Payoff to Player 1	Payoff to Player 2
1	A	5	6
	B	6	5
2	A	5	15
	B	6	5
3	A	5	15
	B	15	5

Admirably, Gneezy (2005) recruits 450 participants to take part in this experiment. He finds the lowest rate of lying (17%), when liars stand to gain the least ($1) and those to whom the lie is told stand to lose the most ($10), i.e. Treatment 2. Lying increases both when either the cost to receiver is lower ($1) and the benefit to the liar stays the same (i.e. Treatment 1, where lying is 36%), or when the cost the receiver stays the same ($10) but the benefit to the liar stays the same (i.e. Treatment 3, where lying is 52%). This is illustrated in Figure 1, from the paper below.

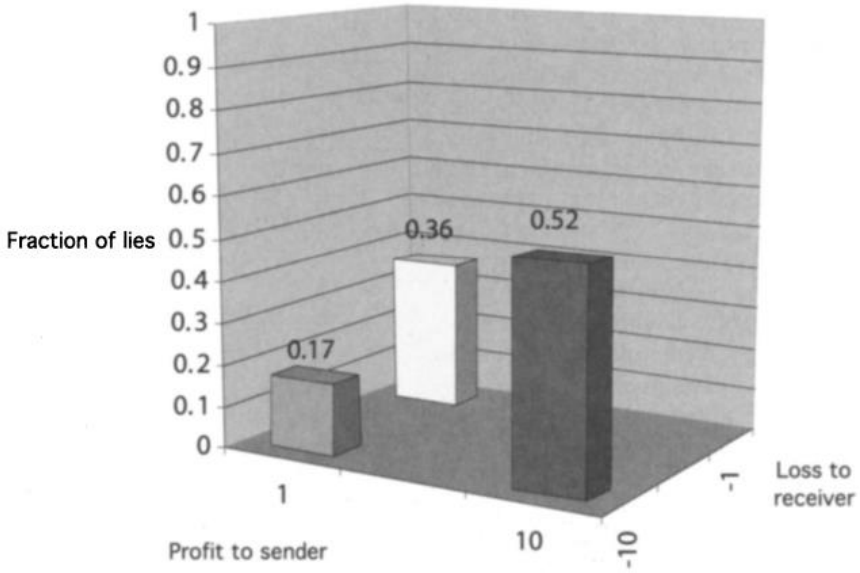


FIGURE 1. FRACTION OF PARTICIPANTS WHO LIED IN THE DECEPTION GAME

Note: The horizontal axis represents the gains from lying for player 1 and the associated loss for player 2.

These results suggest that people take into account the consequences that a lie will have on the person to whom it is told, when deciding whether to lie or not. It also suggest that people have take into account something like "proportionality" when deciding whether to tell a lie. They were more averse to telling a lie that only slightly benefited them, but lead to large loses to their receiver. (Results from the paper suggests that this goes beyond a mere sense of fairness, and

seem particularly strong to considerations of lying. One can set up a dictator game, where the consequences of the dictator's action have the exact same monetary consequences for them and their "subject" as in the three treatment conditions of the sender-receiver game. The authors in fact do this. And when they do, they find, for example, that while lying was only at 17% in Treatment 2 ($1 gain for sender, $10 loss for receiver), the extent to which people pick the option with the same consequences in a dictator game ($1 gain, $10 loss) rises to 42%. So people seem even more careful to not tell lies that results in unequal consequences than they are simply to make decisions or engage in actions that result in unequal consequences[6].)

Individual Differences (Few Prolific Liars)

The final factor that I want to point as having a strong influence on whether people decide to lie or not is the person themselves. People differ. And some people are simply more comfortable with and willing to break the rules, disregard others, and lie.

Evidence for this can be seen in the national survey of the frequency with which people lie in a 24 hour period from Serota et al. (2010). As we saw earlier, most of the people in the survey (59.9%) reported telling no lies in the last 24 hours. Of the remaining 40%, an even smaller portion of "prolific liars" account for most of the lies told[7]. When the authors analyze their distributions, they find that one percent of the sample accounts for over 22% of the lies told, and over half the lies are accounted for by the top 5.3% percent of liars. Serota et al. (2010) also re-analyze data from DePaulo et al. (1996) data as well as several other studies which either directly examine or can be used to examine the prevalence of lying in everyday life (i.e. Feldman, Forrest, & Happ, 2002; George & Robb, 2008). The basic results hold–lying is infrequent, and a small fraction of people seem to account for a disproportionate number of lies. Serota & Levine (2015) follow up on and replicate these results in a paper explicitly devoted to analyzing this subset of frequently liars, appropriately titled (as all the papers in the literature seems to be), *A Few Prolific Liars*.

Big Lies As Violations of Exceptions and Usual Lying Behavior

If the factors above (incentives, consequences, and individual differences) account for substantial variation in the frequency and extent of lying, people are likely somewhat aware of this. For example, how surprisingly was it to you as the reader that (1) people lie when they have

---

[6] The extent to which choosing the self-benefiting "selfish" option increased in each of the three treatment conditions equivalents in the dictator game was about the same in all three conditions (~30% increase in each case). So, this might be better read as further evidence of a general aversion to lying, rather than a particular aversion to unequal lying. Nevertheless, the main results stands that people take into account consequences to others when deciding whether to lie and are particularly averse to telling lies that lead to disproportionate costs and benefits.

[7] Of course, an alternative explanation is that these people, or at least some of them, are not prolific liars, but just have better memories, more inclusive definitions of what constitutes a lie, or are in fact more honest (or most ironically, they may be lying here, by overrepresenting how much they usually lie). It seems reasonable that this might account for at least some of these "prolific liars". How much exactly is unclear and hard to know.

reason to do so, (2) people are less likely to lie if it disproportionately harms someone else, and (3) some people are much more comfortable with and likely to lie than others?

In fact, they can be seen as forming some sort of normative and moral basis that implicitly governs lying. Don't lie unless there is good reason to. Don't lie if its going to disproportionately harm someone. And don't be a pathological or "prolific" liar. Violations of these rules are particularly morally egregious, and also unexpected—so people are not necessarily on high alert for them.

This may be exactly the sort of things that big lies take advantage of. When someone tells a small lie like that they like a sweater that was gifted to them when they don't, we might suspect lying because we are aware of the incentive and motivation to please our gifters. Who however would suspect that when Dan Mallory told his colleagues that he had a spinal tumor his motivation was essentially to excuse his work absences?[8] Likewise in the coinflip scenario in Section 5 ("What makes a lie "big"?: Big Consequential Lies"), who would suspect that a police officer would lie about a coinflip if that officer knew it would result in the death of another person? This would be a big lie. Finally, we might be more suspicious of incredulous claims (like the those that Mallory made) once we know they are made by a person who is a prolific liar. But we're likely to default to assuming they are telling the truth, otherwise.

[8] He now claims that he has bipolar disorder and was too ashamed to admit this. I don't necessarily doubt the diagnosis, and it might actually explain his work absences. Although it's a poor excuse for the lying.

Indeed, the extent to which a lie implicates a person's moral character may be a highly relevant consideration to consider when trying to reason about the psychology of lying. Above and beyond direct incentives and possible consequences, what a various types of lies "say" about a person is, I believe, important both in influencing decisions about whether or not to lie, and people's expectations about which types of statements are likely to be told.

Deciding whether to lie is undoubtedly a moral act both in the mind of the person who might speak a lie and in the mind of person to whom the lie might be spoken. As Uhlmann, (our very own Dave) Pizarro, & Diermeier (2015) point out, in their paper *A Person-Centered Approach to Moral Judgment*, both philosophers and research psychologists often examine and consider moral acts in light of their consequences and in terms of how they violate moral duties. However, the authors go on to explain and convincingly argue that many moral acts and our evaluations of them are better understood in light of their implications on a person's moral character. As the authors nicely put it:

> "Simply stated, when making moral evaluations, it appears as if individuals are often not asking themselves the question "is this act right or wrong?" but rather are asking themselves "is this person good or bad?" (p. 72)

Part of the reason this is so, they argue, is because "social perceivers are fundamentally motivated to acquire information about the moral character of others" (p. 72). In support of this, they marshal evidence from, for example, face perception research that people "easily and quickly" evaluate faces and thus people traits that are essentially goodness and badness (citing Goodwin, Piazza, & Rozin, 2014; Todorov, Said, Engell, & Oosterhof, 2008). And they show how such an account can apparently help explain moral judgments that might otherwise seem puzzling or irrational.[9]

In a particularly noteworthy line, they add:

> "certain transgressions elicit especially negative reactions not because they are unusually wrong in-and-of-themselves, but because they are seen as highly diagnostic of an individual's moral character"

[9] Such as people's tendency to diminish the extent to which they blame acts which are committed impulsively (e.g. crimes of passion) but not reduce their praise for impulsive prosocial acts (Pizarro, Uhlmann, & Salovey, 2003). Although, here I found some of reasoning and reinterpretations a little hard to follow, seeming to need further ad hoc provisos to explain how the person centered approach explains them. Nevertheless, I find the general thrust of the account convincing and this may be a failure of my understanding or reading.

This would certainly seem to align with and help explain the primary reactions that people have upon finding out that someone has told a big lie. The already repeated reaction of one of Dan Mallory's colleagues upon finding out that he lied about having cancer is the perfect example: "*who* would fabricate such a story?" *Who,* as in: what sort of person? *Who*, as in: a person of what kind of morality? Likewise, a crucial component of Hitler's reasoning about why people believe big lies is that those to whom they are told would not believe someone else could have the "imprudence", i.e. negative character, to lie so "infamously".

Thus, people may be reluctant to tell certain types of lies (e.g. lying about spinal cancer or your brother's suicide) because of what they would indicate about a person's moral character. And this in turn likely affects the types of lie that people expect to hear and topics people are expected to lie about.

Some traces of evidence that people take into account the implications that a lie has on their moral character might be seen in a from Mazar, Amir, & Ariely (2008). In several studies, they give people the opportunity to lie by misreporting their performance on a math task and find that they don't do to the maximum degree possible (i.e. don't say they've solved all the math problems). The authors argue that this is because people want to reap some of the gains that can be had from lying, but not so much so that they feel bad about themselves, i.e. maintain their self-concept as a mostly honest and fair person. To support this claim, they report the results of several other studies which seem to indicate that the degree of cheating decreases when self-concept concerns are heightened by having participants sign an (essentially symbolic more than enforceable) honor code before participating. All this suggests an attunement to the implications that a lie has on one's moral character.[10]

Speculations

I would like to end this section with a few speculations. First, this perspective might suggest that one way to get someone to believe a lie is to include somewhere in it an admission or evidence of a moral flaw. This is likely to be particularly effective when the flaw to which you admit is more severe and indicative of poor moral character than the moral transgression that the lie attempts to cover up (e.g. saying "I forgot to pick up the groceries because I was cheating on my wife", when in fact you were just lazy). Of course, such lies make little sense to tell. Why get yourself into more trouble, when you could avoid this by telling the truth? Nevertheless, it is

[10] Gneezy, Kajackaite, & Sobel (2018, p. 448) call into question that Mazar, Amir, & Ariely's (2008) partial lying results are due specifically to concerns about maintaining a "positive self-value", in experiments of their own that allow partial lying—which they examine in conditions to which participant lies are observed to different degree. I actually don't see how their results support this apparent refutations. However, their alternative explanation is that these results are driven by "social identity" concern to be "viewed as honest" (p. 421) by others, a concern that seem to map equally well onto the idea that people are attuned to effects of moral transgressions on one's reputation and character.

precisely because of this logic, that no one would suspect you of lying (they would more readily suspect you of joking, if anything).

A second speculation is that this framework might lend insight into why people of different political orientations can seemingly come to different conclusions about statements from politicians that one side might consider a lie, and about the extent to which these lies are big and brazen.

Which of these same two lies seems bigger and more diagnostic of one's rotten moral character? Donald Trump falsely claiming that his inauguration was the most highly attended inauguration in history (Garber, 2019). Or a future President Elizabeth Warren making the same claim about the attendance at her inauguration? If you are already predisposed to viewing Donald Trump as morally reprehensible, this lie is further evidence that he is not to be trusted and a threat to the communities and institutions you value. It is big and bold, indicative of dishonesty that is likely to carry over into other domains. Meanwhile if you see Elizabeth Warren as a warm, and moral upright person, such a lie seems more indicative of perhaps getting momentarily caught up in the hype and excitement of the moment.

(If you object that Donald Trump *actually* is more predisposed and so you would be right, substitute another person whose moral character you might be wary of, but you don't necessarily consider a pathological liar, imagine say Ted Cruz or perhaps more fairly Mitt Romney making the same claim. Do lies coming from people you like, trust and support seem as big as those same lies coming from people you dislike, dislike, and don't support? For me, any lie I can think of seems bigger and more indicative of moral rot when coming from the mouth of politicians I don't like, whereas when they come from the mouths of politicians I do like, they seem more isolated and ultimately motivated from a good place.)

10: Proposal for a Preliminary Empirical Test

I propose a brief empirical test, that could garner cursory and very preliminary evidence that the general ideas proposed here are in the right direction. This takes the form of a survey experiment. The basic goal is to demonstrate that (1) a strong influence on the types of lies that people tell are the implications that lie would have about their moral character, (2) moral concerns are often more predictive of people's suspicions about whether someone might be lying to them than more "cognitive" aspects of what they might take into account like the probability that the literal content that they are communicating is true.

The basic structure and components of the survey would be as follows. One group of participants would take the perspective of speakers, that is, potential liars. They would be given a scenario (e.g. skipping work), and a range of options about how they might explain themselves in that scenario. This would include several different lies as well as the truth. They would then be asked to make ratings of: how likely they would be to make each statement, classic factors that might be thought to correspond to the bigness of a lie (e.g. the extent to which it deviates from reality), and also the implications that making each statement would have about their moral character. A separate set of participants would then be given this same scenario, from the perspective of the person to whom the statements might be told. These participants would rate: the extent to which they would believe the statement, other classic factors that might be thought to correspond to the bigness of a lie (e.g. the extent to which claim seems ex ante unlikely to them), and also the extent to which telling such a lie would make its teller morally abhorrent.

An example of one such scenario and the ratings that both groups of participants would be asked to make is shown below.

Scenario (Speaker's Perspective): You skipped worked yesterday. You did this mostly because you just didn't feel like coming in. Your boss asks you why you weren't at work yesterday. Please rate your reactions to the following responses you might give.

Lie	Likelihood/Willingness to Say It	Extent of Deviation from Reality	Reflection on Moral Character
	How likely would you be to make this statement? 1 = Not at all 2 = Only slightly likely 2 = Somewhat likely 3 = Fairly likely 4 = Very likely	To what extent does this statement deviate from reality (what is really true)? 1 = Not at all 2 = Only slightly 3 = Somewhat 4 = A fair amount 5 = Very much	To what extent would you be a bad person if you made this statement? 1 = Not at all 2 = Only slightly 3 = Somewhat 4 = A fair amount 5 = Very much
You say that you just didn't feel like coming in.			
You say that you felt sick yesterday.			
You say that your car broke down because a squirrel had lodged itself in your car's engine and you had to go to a mechanic to get it fixed.			
You say that you were diagnosed with brain cancer earlier in the week.			
….			

Scenario (Receiver's Perspective): An employee of yours skipped worked yesterday. You don't know why. Your ask them why they weren't at work yesterday. Please rate your reactions to the following responses, if they gave each.

Lie	Likelihood That Statement Is True	Extent to Which Statement Violates Expectations	Moral Offensiveness
	How likely do you think it is that this statement is true? 1 = Not at all 2 = Only slightly likely 2 = Somewhat likely 3 = Fairly likely 4 = Very likely	How likely is it that something like this would happen (not the likelihood that they would say this, but the chance that such a thing could happen)? 1 = Not at all 2 = Only slightly 3 = Somewhat 4 = A fair amount 5 = Ver much	If this were a lie, to what extent would the person who said it be a bad person? 1 = Not at all 2 = Only slightly 3 = Somewhat 4 = A fair amount 5 = Very much
They say that they just didn't feel like coming in.			
They say that they felt sick yesterday.			
They say that their car broke down because a squirrel had lodged itself in their car's engine and they had to go to a mechanic to get it fixed.			
They say that you were diagnosed with brain cancer earlier in the week.			
…			

Some of the key predictions I would have are that (1) the best predictor of speaker ratings of their likelihood of telling a lie, among the other speaker ratings, would be their rating of a lie's moral offensiveness, and (2) the best predictor of speaker ratings of the likelihood of telling a lie, among the receiver ratings, would be receiver ratings of the moral offensiveness of a lie.[11] More importantly, I would expect (3) the best predictors of receiver ratings of their likelihood of believing a response, among the other receiver ratings, would be their rating of a response's moral offensiveness if it were a lie, and (4) the best predictors of receive receiver ratings of their likelihood of believing a response, among the speaker ratings, would either be speaker ratings of the reflections of each statement on their moral character or their ratings of how likely they would be to make each statement.

Such findings would be useful in establishing that concerns about the moral implications of lies are paramount in both the lies that people tell and what types of lies are believed.

[11] Since speaker and receiver ratings are not linked in this between subject design, I could conduct this analysis by first computing the average rating of each lie, separately for both the speakers and the receivers. I could then link together aggregated speaker and receiver ratings for each lie (i.e. the "rows" in the correlation analysis would be each lie; thus, I would likely actually need to generate and have participants rate a larger number of potential lies in each scenario, ideally up to 25 or more).

11: Rules of Lying

Finally, I would like to note some factors that might aid the believability of certain lies or influence which lies are told, aside from their moral abhorrence. This have to do with certain social and linguistic conventions that we might expect people to adhere to when lying.

The Rules of (Honest) Communication

Let's start by discussing general principles that might govern honest communication. Some are suggested by philosopher Herbert Paul Grice (1975) in his paper about the norms that govern conversation. He suggests that in communication there is an overarching "cooperative principle" whereby the speaker and the receiver generally understand the "common purpose … or at least, mutually accepted direction" of the exchange (which to some extent centers around "maximally effective exchanges of information" although might also include "influencing or directing the actions of others"). And so, towards this end, there are certain rules of conversation that people should and have come to obey. These are categorized into principles of *quantity, quality, relation,* and *manner*. The principle of quantity mandates that a speaker be informative as necessary, without being overly or unnecessary informative. The principle of quality contains the maxims to "not to say what you believe is false"[12] and to "not say that for which you lack adequate evidence". The principle of relation succinctly says "be relevant"—that is, your comments should be relevant to the discussion and goals of the discussion at hand. These first three principles roughly govern *what* should be said. The final principle, the principle of manner, governs *how* one should speak. And it says that one should speak as clearly as possible—avoiding obscurity, ambiguity, and seeking to be brief and speak in a logically ordered manner.

The Rules of Lying

When people set out to lie, they nevertheless take into consideration the consequences of their lies on others, as we saw in Section 8 ("When Do People Lie?: Consequences/costs/harm to receiver"). Thus, when lying people might nevertheless, still try to adhere to the "cooperative principle", attempting to maintain a common purpose, as much as possible given one is lying. The ways to do this, when lying might be do adhere to the following rules.

- *minimize the "size" of the lie:* In episode of *The Office*, Kevin starts speaking in very terse, nongrammatical (although understandable) sentences, eventually explaining his logic by saying: "why waste time say lot word, when few word do trick". Similarly, a principle with

[12] Note this is not the same as a mandate not to lie. As a lie is a statement which is not only believed to be false,but communicated to deceive the person to whom it is spoken. Thus, lies are a subset of the types of communications this maxim prohibits.

lying might be: why say big lie, when small lie work? To extent that the cooperate principle tries to ensure that people are on the same page, headed in roughly the same direction, minimizing the "size" of a lie (the extent to which it diverges from and misrepresents reality) might help still keep communicators stay as much on the same page as possible. For this reason, people may tell lies that accomplish their goal but otherwise leave reality as undisturbed as possible. A lie that might suffice to get a student out of a missed homework assignment is to tell their teacher they had a bad headache the night before—a purely internal state, and problem that resolved itself in short order. A "big lie" that would also probably get a student out of a missed homework assignment is to tell their teacher that they were in the hospital all last night because their father had a heart attack. In addition to the moral implications of the lie, it diverges from reality to a large extent, leaving a more indelible mark and lasting consequences. For example, the teacher might spend the rest of the day worrying, and engage in other actions she otherwise wouldn't have, if a smaller, equally sufficient lie were told—she may call up the parents, spend her time visiting the hospital, read articles about how to console grieving students, or try to call in a favor from her brother who is a cardiologist at the hospital; she may bring this up at a future parent-teach conference, leading to further confusion and disarray.

- *avoid details and concreteness:* When speaking honestly, the cooperative principle tries to ensure such outcomes like that speakers only communicate as much as is necessary. It is still possible to adhere to this principle to some extent when lying. Since you are communicating information of little value, the best when to communicate as much as is necessary is to communicate as little as possible. Any details or concrete descriptions are only likely to mislead further, and confuse and prolong the interaction by perhaps unearthing inconsistencies. Of course, such a strategy also helps the liar achieve their ends. Because they are lying, the more details and more concrete they are about their claims, the greater chance they have of giving themselves away, slipping up somewhere, in a way that might lead to a contradiction that someone might notice, leading them to question a liar's claims, and perhaps find further contradictions and eventually unravel the lie.

What empirical evidence can be marshaled to support the possibility that people lie in ways that abide by either of these rules?

To the extent that larger "sized" lies have more negative consequences for others (by misleading them further and possibly leading them to take unnecessary actions), the results from the Gneezy (2005) study the role of consequences might suggest that people would seek to minimize the size of their lies, holding all else constant. An experiment that directly manipulates the size of lie in the way that we mean (the size of distortions from reality) comes from the Gneezy et al. (2018) paper referenced in Section 5 ("What makes a lie "big"?: Economic

Perspectives"). In several experiments, they had a computer assign people random numbers from 1 to 10. Participants were then asked to report what number they had been assigned, giving them an opportunity to lie in ways that resulted in different "sized" distortions of reality (e.g. if you actually were assigned a 1, reporting a 10 in some sense is bigger deviation from reality that reporting you got a 2). However, in all conditions, the payoffs for different lies were always such that "size" of lies in terms of their distortion from reality were either confounded with the "size" of lies in terms of their payoffs and consequences (e.g. reporting being assigned a 1 earned you 1 euro, reporting a 2 earned you 2 euros, and so on up until reporting 10, which earned you 10 euros), or there was a tradeoff off between "size" of a lie in terms of their distortion from reality and their size in terms of associated payoffs. Unsurprisingly, when there was a tradeoffs, the payoffs exerted a stronger influence. Further, it should be noted that these exchanges did not occur in anything like conversational contexts, where there would be a use to communicating in cooperative ways. Nevertheless, it stands to reason and future research could demonstrate that people minimize the size of the lies, when all other factors are held constant. (Indeed a simple survey experiment could simply ask people about their tendency to do this.)

What evidence exists to support the possibility that people might calibrate their lies in order to avoid being too concrete or giving to many details? If any part of the motivation to avoid details and being concrete comes from attempts to minimize one's chances of getting caught, we might turn to Park, Levine, McCornack, Morrison, & Ferrara's (2002) refreshingly simply paper, *How People Really Detect Lie* Park. In it, they report the results of a study in which they asked participants to recall an instance when they discovered that someone was lying and to describe the situations in as much detail as possible. The authors then coded these written responses into various categories depending on how the lie was found out (e.g. contradiction from third party information, physical evidence). The results are presented below, in Table 1 from the paper.

TABLE 1

FREQUENCIES OF RECALLED LIE DISCOVERY METHODS

	Initial Coding		Methods in Combination			
Discovery Method	*f*	%	f_a	f_b	$\%_c$	$\%_d$
Third Party Information	62	32.0%	39	101	52.1%	38.0%
Physical Evidence	35	18.0%	25	60	30.9%	22.6%
Solicited Confession	7	03.6%	29	36	18.6%	13.5%
Unsolicited Confession	16	08.2%	9	25	12.9%	09.4%
Verbal/Nonverbal Behavior	4	02.1%	18	22	11.3%	08.3%
Inconsistent with Knowledge	4	02.1%	8	12	06.2%	04.5%
Inadvertent Confession	4	02.1%	4	8	04.1%	03.0%
Combination	60	30.9%	–	–	–	
Other	2	01.0%	0	2	01.0%	00.8%
Total	194		132	266		

Note. [a] denotes the frequency of individual's discovery methods within combinations, [b] is the total frequency in which the discovery method was observed, including use in combinations, [c] is the percentage of lie discoveries ($N = 194$) involving the method, and [d] is the percentage of total discovery units (266) that were the specific method.

As we can see, the most frequent ways that people were caught, involved some direct contradictions of fact—like a third party saying something else (32%) or some piece of physical evidence belying their claims (18%). The more that people avoid giving details and being concrete, the more likely it would seem they are to being exposed.

Big Lies

As we noted as the very beginning in Section 3 ("Core Response"), in addition to being morally reprehensible many big lies seem to have the feature that they are concrete. Violating this rule of lying may lend further credulity to a big lie. Just as people are unlikely to expect people to tell lies that would render their tellers likely to be judged morally reprehensible if they were to be exposed, it would also seem reasonable that people don't expect others to tell extremely detailed and concrete lies that leave their tellers more vulnerable to being caught.

By subverting this expectation of speech as well, big lies might lend themselves further credibility. We are likely to believe someone if they say that they have cancer. We are even more likely to believe them if they tell us the name of the doctor who diagnosed them, the hospital at which they are being treated, and the times they visit that hospital.

12: Final Summary

Big lies are best understood as “big brazen lies”, where people tell a lie that would bring upon them harsh moral judgment should they be found out—partly because of the ways that misrepresentations on important matters violate the trust that is foundational to relationships, communities, and institutions. Such big brazen are out of line with the usual principles that evidence suggests governs people’s patterns of lying: lie for good reason, do so in ways that proportionally weight the costs and benefits to you and others, avoid behavior that might lead you to be labeled a prolific liar. Future research may best be pointed to examining the particular influence that judgments of moral character have on lying and assessments of lies.

It is because of their subversions of normal and sensible behavior that big lies are both unlikely to be told and more likely to be believed.

REFERENCES

(includes references in appendix)

Allen, V. (1966). Review: Attitude and attitude change. *American Sociological Review*, *31*(2), 283–284.

Bacon, F. (1620). *Novum Organum*.

Bok, S. (1999). *Lying: Moral choice in public and private life*. Vintage.

Bond Jr, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*(3), 214–234.

Cai, H., & Wang, J. T.-Y. (2006). Overcommunication in strategic information transmission games. *Games and Economic Behavior*, *56*(1), 7–36. https://doi.org/10.1016/j.geb.2005.04.001

Chen, S., Shechter, D., & Chaiken, S. (1996). Getting at the truth or getting along: Accuracy- versus impression-motivated heuristic and systematic processing. *Journal of Personality a, d Social Psychology*, 262–275.

Choi, I., Nisbett, R. E., & Norenzayan, A. (1999). Causal attribution across cultures: Variation and universality. *Psychological Bulletin*, *125*(1), 47.

Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. *Econometrica: Journal of the Econometric Society*, 1431–1451.

deceive. (2019, March 25). Retrieved March 25, 2019, from https://www.merriam-webster.com/dictionary/deceive

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology*, *70*(5), 979.

Feldman, R. S., Forrest, J. A., & Happ, B. R. (2002). Self-presentation and verbal deception: Do self-presenters lie more? *Basic and Applied Social Psychology*, *24*(2), 163–170.

George, J. F., & Robb, A. (2008). Deception and computer-mediated communication in daily life. *Communication Reports*, *21*(2), 92–103.

Gneezy, U. (2005). Deception: The Role of Consequences. *American Economic Review*, *95*(1), 384–394. https://doi.org/10.1257/0002828053828662

Gneezy, U., Kajackaite, A., & Sobel, J. (2018). Lying Aversion and the Size of the Lie. *American Economic Review*, *108*(2), 419–453. https://doi.org/10.1257/aer.20161553

Gneezy, U., Rockenbach, B., & Serra-Garcia, M. (2013). Measuring lying aversion. *Journal of Economic Behavior & Organization*, *93*, 293–300.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148.

Grice, H. P. (1975). Logic and conversation. *1975*, 41–58.

Hofstadter, D. R. (1979). *Gödel, Escher, Bach: an eternal golden braid* (Vol. 20). Basic books New York.

Hofstadter, D. R. (1983). Metamagical themas. *Scientific American*, *248*(1), 14–26.

Hovland, C. I., Harvey, O. J., & Sherif, M. (1957). Assimilation and contrast effects in reactions to communication and attitude change. *The Journal of Abnormal and Social Psychology*, *55*(2), 244.

Hurkens, S., & Kartik, N. (2009). Would I lie to you? On social preferences and lying aversion. *Experimental Economics*, *12*(2), 180–192.

Jackman, I. (2003). *Con Men: Fascinating Profiles of Swindlers and Rogues from the Files of the Must Successful Broadcast in Television History*. New York, NY: Simon & Schuster.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*(2), 136.

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

Levine, T. R. (2014). Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection. *Journal of Language and Social Psychology*, *33*(4), 378–392. https://doi.org/10.1177/0261927X14535916

Levine, T. R., Kim, R. K., & Hamel, L. M. (2010). People Lie for a Reason: Three Experiments Documenting the Principle of Veracity. *Communication Research Reports*, *27*(4), 271–285. https://doi.org/10.1080/08824096.2010.496334

Levine, T. R., Kim, R. K., Sun Park, H., & Hughes, M. (2006). Deception detection accuracy is a predictable linear function of message veracity base-rate: A formal test of Park and Levine's probability model. *Communication Monographs*, *73*(3), 243–260.

Levine, T. R., Park, H. S., & McCornack, S. A. (1999). Accuracy in detecting truths and lies: Documenting the "veracity effect." *Communications Monographs*, *66*(2), 125–144.

Little, A. T. (2017). Propaganda and credulity. *Games and Economic Behavior*, *102*, 224–232.

López-Pérez, R., & Spiegelman, E. (2013). Why do people tell the truth? Experimental evidence for pure lie aversion. *Experimental Economics*, *16*(3), 233–247.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098.

Lundquist, T., Ellingsen, T., Gribbe, E., & Johannesson, M. (2009). The aversion to lying. *Journal of Economic Behavior & Organization*, *70*(1–2), 81–92.

Mahon, J. E. (2016). The Definition of Lying and Deception. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016). Retrieved from https://plato.stanford.edu/archives/win2016/entries/lying-definition/

Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review*, *3*(1), 23–48.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*(6), 633–644.

O'Reilly, R. C., Munakata, Y., Frank, M. J., Hazy, T. E., & Contributors. (2012). *Computational Cognitive Neuroscience*. Retrieved from http://ccnbook.colorado.edu

Park, H. S., Levine, T., McCornack, S., Morrison, K., & Ferrara, M. (2002). How people really detect lies. *Communication Monographs*, *69*(2), 144–157.

Parker. (2019, February 4). A Suspense Novelist's Trail of Deceptions | The New Yorker. Retrieved June 6, 2019, from https://www.newyorker.com/magazine/2019/02/11/a-suspense-novelists-trail-of-deceptions

Petty, R. E., & Cacioppo, J. (1986). *The Elaboration Likelihood Model of Persuasion*. 83.

Pizarro, D., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological Science*, *14*(3), 267–272.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II: Current Research and Theory*, *2*, 64–99.

Schacter, D. L., Gilbert, D. T., Nock, M. K., & Wegner, D. M. (2017). *Introducing Psychology* (Fourth edition). New York, NY: Worth Publishers.

Serota, K. B., & Levine, T. R. (2015). A Few Prolific Liars: Variation in the Prevalence of Lying. *Journal of Language and Social Psychology*, *34*(2), 138–157. https://doi.org/10.1177/0261927X14528804

Serota, K. B., Levine, T. R., & Boster, F. J. (2010). The Prevalence of Lying in America: Three Studies of Self-Reported Lies. *Human Communication Research*, *36*(1), 2–25. https://doi.org/10.1111/j.1468-2958.2009.01366.x

Sharot, T., & Garrett, N. (2016). Forming Beliefs: Why Valence Matters. *Trends in Cognitive Sciences*, *20*(1), 25–33. https://doi.org/10.1016/j.tics.2015.11.002

Sherif, M., & Sherif, C. (1967). Attitude as the individual's own categories: The social judgment-involvement approach to attitude and attitude change. In *Attitude, ego-involvement, and change*. John Wiley & Sons, Inc.

Sherif, M., Taub, D., & Hovland, C. I. (1958). Assimilation and contrast effects of anchoring stimuli on judgments. *Journal of Experimental Psychology*, *55*(2), 150.

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455–460.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A Person-Centered Approach to Moral Judgment. *Perspectives on Psychological Science*, *10*(1), 72–81. https://doi.org/10.1177/1745691614556679

Wang, J. T., Spezio, M., & Camerer, C. F. (2010). Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth Telling and Deception in Sender-Receiver Games. *The American Economic Review*, *100*(3), 984–1007.

Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. In *Advances in experimental social psychology* (Vol. 14, pp. 1–59). Elsevier.

APPENDIX

Earlier, in Section 2 ("Thoughts on Question"), I mentioned that I have tried to answer this question several times, and from different perspectives. I wanted to include some of those attempts below, to give a flavor of them and to show that I took seriously and spent some time on very different approaches and literatures. These writings are of course fragmentary and incomplete.

Most of them are attempts at answering or getting a handle on what I delineated as the first question of the two I claimed are contained in the prompt, having to do with how people respond to discrepant information. This section of the prompt (which I highlighted in blue earlier), and it is reproduced below.

"We assume that the more a new piece of information fits with people's prior knowledge and experience, the more likely they will be to learn/believe it. The more discrepant a piece of information (the less it fits, the more surprising, or extreme it seems, etc.), the less likely people will be to learn/believe it. Please review why this is the common assumption, drawing from whatever literatures you think are the most relevant. What evidence is there for it? And, are there any exceptions to this rule? Are there ever cases in which someone is more likely to learn/believe an unexpected (versus expected) piece of new information?"

Word Count (from beginning of next page to end of document): 11,302

1: Discrepancy & Attitude Change (Elaboration Likelihood Model)

PART 1: WHY IS THIS A COMMON ASSUMPTION IN PSCYHOLOGY?

Psychological research suggests the general principle that the more discrepant a piece of information is from a person's knowledge and experience, the less likely they are to accept or believe that information. The empirical origins of this wisdom can be found in several literatures. I review these in turn.

attitude change

One immediately relevant area of research in psychology is the topic of attitude change, which aims to study the conditions under which people change their minds and attitudes. In this literature, we see an emergence of the idea that more discrepant information is rejected almost as soon as the field emerged.

a*n elaborate detour*

Although it is preceded by at least two decades of research, probably the most well-known and highly cited article in the field of attitude change is (Petty & Cacioppo, 1986) Elaboration Likelihood Model (ELM) of Persuasion. Let me first briefly summarize the model since it is important and influential in a lot of the subsequent work on attitude change. And then I will explain where in this model the idea, that the amount of discrepancy is related to the amount of attitude change, comes in. And then I will also review even earlier research in attitude change, which also promulgate the discrepancy idea.

With their ELM framework, (Petty & Cacioppo, 1986) aim to construct a model[13] of attitude change. They construe attitudes broadly[14]—although, most usually their model is best describing a case of attitude change in response to a communicated message (and even more specifically, a written or oral communication delivered by another person; this is certainly the

[13] The authors repeatedly refer to their framework as a model, which I think is a fair characterization. But it should be pointed out that they never get so far laying out any sort of actual mathematical or computer model. So the word model here refers to a verbally described framework (that is not to say that is doesn't offer at least some broad and potentially falsifiable hypotheses).

[14] "We regard attitudes as general evaluation people hold in regard to themselves, other people, objects, and issues" (p. XXX). They add: "These general evaluations can be based on a variety of behavioral, affective, and cognitive experiences" (p. XXX), giving examples of each type: liking a candidate more after donating to their campaigned would be an example of behavior induced attitude change; liking a candidate more because they liked the music they saw in a campaign ad would be an example of affected-induced attitude change; and liking a candidate more because of that candidate's policy positions would be an example of cognitive-initiated attitude change.

empirical setup of most of their studies, on which the theory is based). The model is unfurled as a series of postulates. Simplifying a bit, these can be grouped and summarized as follows:

(1) People want to hold "correct"[15] attitudes, but they do not always have the motivation or ability to arrive at a correct attitude.
(2) Evaluation of a communication is based on both (a) the content and quality of the communication and (b) cues that don't have to do with the communication itself (e.g. the source of the argument). The former is referred to as "central" processing of a communication, and the latter as "peripheral" processing (roughly mapping on the popular System 1, System 2 framework).
(3) Attitude change is thus determined by: (a) argument quality, (b) peripheral cues, (c) the extent to which the communication is evaluated on content and quality or peripheral cues. This final factor can be thought of as the extent to which scrutiny (or "elaboration" in the author's words) is applied when evaluating an argument.
(4) Subsequent postulates aim to specify the factors that influence the likelihood of elaboration.

The theory was notable precisely because of its suggestion of elaboration likelihood, which helped integrate the cacophony of previous theories of attitude change—some of which seemed to only explain attitude change under certain circumstances (inoculation theory: McGuire, 1964; cognitive response theory: Greenwald, 1968; Petty, Ostrom, & Brock, 1981; information integration theory: Anderson, 1981; the theory of reasoned action: Ajzen & Fishbein, 1980; Fishbein, 1980), which they authors proposed to be conditions of high elaboration likelihood, and some of which seemed to only explain attitude change under other circumstances (classical conditioning: Staats & Staats, 1958; mere exposure: Zajonc, 1968, 1980), which the authors proposed to be conditions of low elaboration likelihood.

In several places, the idea that we pay attention to the amount that a message is discrepant comes up. First, in reviewing previous research on attitude change, they point to the work on social judgment theory in the 1960s and 1970s. They explicitly reference discrepancy in the following section summarizing previous models of attitude change:

> Social judgment theory (Sherif & Sherif, 1967) proposes that people evaluate messages mostly on the basis of their perceived position messages are contrasted and rejected if they appear too discrepant (fall in the latitude of rejection), but are assimilated and accepted if they appear closer to one's initial position (fall in the latitude of acceptance; Pallak, Mueller, Dollar, & Pallak, 1972).

2: Discrepancy & Social Judgment Theory

---

[15] What are "correct" attitudes is never defined rigorously—but can be generally read to mean attitudes and opinions which comport with reality.

Sherif & Sherif (1967) present a theory of attitude change (their *social judgment-involvement approach*). They begin by trying to grapple with some of the fundamental building blocks that would need to be defined in order to make clear the objects on which a theory of attitude change would be operating.

First, and most importantly, they ask – what is an attitude? They begin by making a few points, which I agree with, and think are worth mentioning as they are likely to sharpen our discussion and thinking as well. They quote the commentary of Allen (1966) who notes the "excessive preoccupation with changing a response on an isolated (and apparently randomly selected) issue in the laboratory." And they further go on to ask what exactly we mean by an attitude change "number of beans in a jar, or leaves on a tree, or sand pebbles on a square yard of beach? Or are we concerned with his views on his family, on how he sees himself as a person relative to his contemporaries, on the worth his religion, his politics, his profession, his country or his way of life? It is one thing to attempt to change a person who uses a toothbrush to switch from one brand to another. It is quite another thing to persuade someone who has never brushed his teeth that a toothbrush should be used." (p. 111-112) Through this, they come up with an interesting definition of an attitude. They say that it is an internal mental state, with the following properties:

- Attitudes are not innate (crucially, they are "dependent on learning") (p. 112)
- Attitudes can change but are somewhat enduring (they do not fluctuate like "homeostatic" states) (p. 112)
- Attitudes set up some sort of relationship between the person holding them and an object (the authors do not make this reference, but what they are trying to describe seems similar to what philosopher's would call a "propositional attitude")
- Attitudes have motivational or affective properties (this is because many attitudes are formed in the context of highly significant social interactions)
- There is a range of discrete sets of attitudes a person may have on a matter. And a person can vary in the extent to which they have a positive or negative evaluation of these possible attitudes.

They then go on to suggest that on a given topic, a person's attitude can be described by taking the available set of discrete attitudes on that topic, and then placing them into one of three categories: latitude of acceptance, latitude of rejection, latitude of noncommitment. Empirically, they define two ways to do this: the "method of ordered alternatives" and the "own categories procedure".

In Figure 1, we can see the mean number of positions in each of the three categories, where participants positions are ordered by extremity on the x-axis. (The issue concern political issues in the 1960 Presidential election, and position A represents the most extreme Republican position, while position I represents the most extreme Democratic position.) As we can see, as the issues extremity increases, the number of positions in a subject's latitude of acceptance doesn't seem to change, however the number of attitudes in their latitude of rejection increases. Thus, the more extreme a person's view, the less likely they are to accept (the more likely they are to reject) discrepant positions.

Figure 1: Number of Positions in Latitudes of Acceptance, Rejection, and Noncommitment

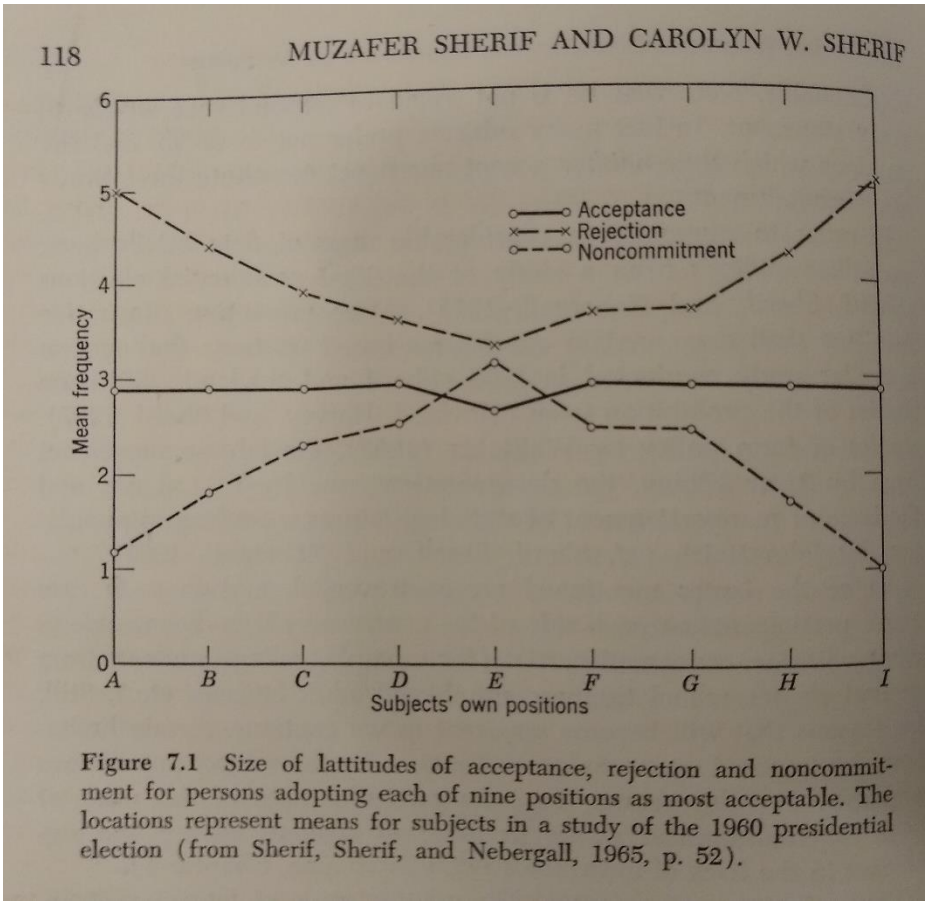


Figure 7.1 Size of lattitudes of acceptance, rejection and noncommitment for persons adopting each of nine positions as most acceptable. The locations represent means for subjects in a study of the 1960 presidential election (from Sherif, Sherif, and Nebergall, 1965, p. 52).

Why is this? The explanation that the authors offer is the core of many of the positions that other attitude change and related psychological theories offer for the why people are likely to reject discrepant information: the *self*. The stronger that one's position is on a topic, the more the "ego" is involved (p. 199). And the more resistance that such a position will face.

assimilation and contrast

Specifically, Sherif & Sherif (1967) argue that attitude change is a function of assimilation and contrast effects where one's own position is used as the focal anchor point. The more general process of assimilation and contrast which the authors rely on is demonstrated by an earlier study by Sherif, Taub, & Hovland (1958). In this study, participants were asked to rate the weight of six weights[16] along an arbitrary six point scale, just by the physical feel of the weights. In the control condition, they made these estimations one by one without any anchor (weights were presented in an arbitrary order). There were then nine experiment conditions, in which participants made the same judgment but with the introduction of an anchor or "reference" weight which they were told should correspond to the highest point on the scale (6). Across the conditions, the weight of this anchor weight increased. In the lightest condition, the reference

[16] With weights: 55, 75, 93, 109, 124, and 141 grams

weight (144 grams) was equal in weight to the heaviest weight in the set of six weights to be evaluated (144 grams). In the most extreme condition, this anchor weight (347 grams) was more than twice the weight of the heaviest weight in the set of six weights to be evaluated (144 grams). The authors then examine the distribution of weight judgements across these different conditions. In the control condition, the distribution of judgments looks more or less uniform across the six scale points.

The key result that the authors obtain is demonstrated in Figure's 2 and 3. When the anchor or reference point is not far away from the stimulus values, the stimuli are judged as closer to the anchor (an assimilation effect). When the anchor or reference point is far away from the stimulus values, the range of the stimulus values are restricted (more clustered) together and farther away from the anchor (a contrast effect). (Of course, how we are to determine when a reference point is "close" versus "far away" is not described.)

> "*In complex social communication the introduction of reference points may produce two opposing effects. Under some conditions the introduction of a reference point or stand beyond S's current position tends to move him towards the new position. Thus, telling him that experts think it will be at least 10 years before peaceful use of atomic power is feasible, may cause an individual to increase his own estimate from one of 5 yr. to one of 6 or 7 yr. Under other conditions the introduction of communication results in a rejection of the new proposal and a stronger entrenchment in his original position. Here on has the frequently mentioned "boomerang effect" (2, 3, 7, 9). To some extent at least, these phenomena may be the result of judgmental processes conceptually closely akin to the phenomena of assimilation and contrast in the judgment of simple stimulus material.*" (p. 150)

Figure 2: Anchoring as a Function of Anchor Extremity

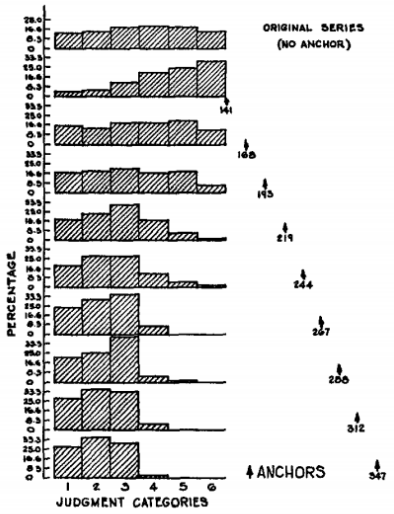


FIG. 1. Distribution of judgments for original series of weights without anchor (top) and with anchors at increasing distances above original series (heaviest anchor at bottom).

Figure 3: Judgment Discrepancy as a Function of Anchor Extremity

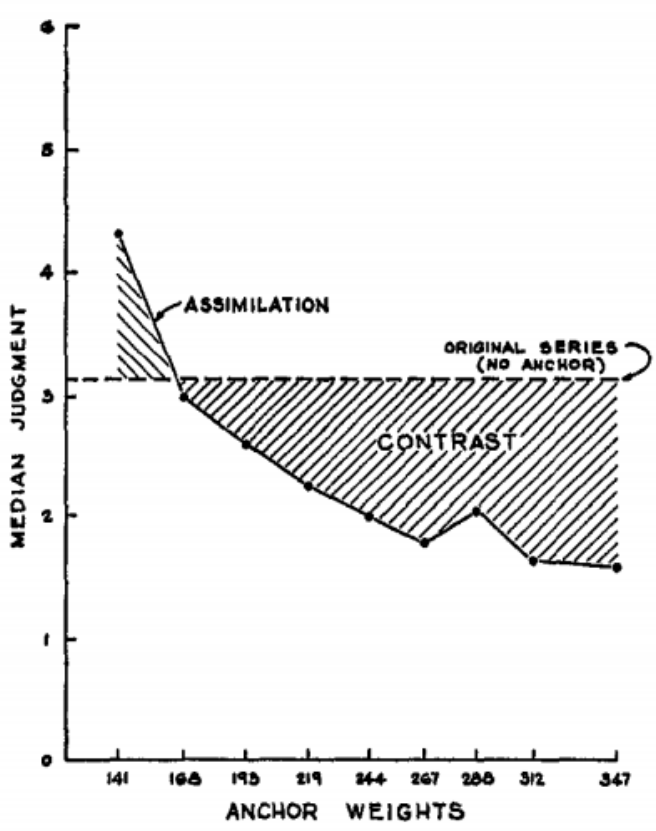


FIG. 2. Median judgments with anchors at increasing distances from the series (abscissa) plotted against the median judgment without anchor.

The authors argue that a similar mechanism is at play in determining people's attitudes and those change in those attitudes in response to a communication, which varies in the extent to which it is discrepant with a person's original attitude. Evidence in support of this idea is presented by Hovland, Harvey, & Sherif (1957). Their key prediction is that "whether assimilation or contrast effects appear would be a function of the relative distance between S's own stand and the position of communication" (p. 245).

They recruited subjects with differing pre-existing attitude on a controversial political issue—the prohibition of alcohol in the participants' state (an issue on which there had just been a vote, which passed by a very narrow margin, keeping the state a "dry" state for the time being). Participants in favor of prohibition ("drys") were recruited from the Women's Christian Temperance Union groups and from a group of workers at the Salvation Army. Hilariously, because those who were opposed to prohibition ("wets") were not as systematically organized or identifiable by public positions or group memberships, those opposed to prohibition were recruited "on the basis of cases personally known to E's [experimenters] or their assistants" (p. 246). A final moderate group was recruited who did not have clear or extreme views on the

issue, consisting of students in classes like journalism, speech, chemistry, and so on. The authors then gather over 500 statements on this issue from members of the local community, these were then sorted by 20 judges into 8 categories. For each category, the authors then came up with a summary statement to represent these 8 categories, which can be ranked ordered from most favorable to being "dry" (position A) to most favorable to "wet" (position H).[17] Over the course of several weeks, "dry" participants were exposed to a 15 minute long argument (actually made in the prohibitions debates by advocates) opposed to prohibitions and a 15 minute argument on the moderate side of the debate; "wet" participants were exposed to a 15 minute argument made by those in support of prohibition as well as the moderate argument; and moderate participants heard all three arguments. Participants then rated the arguments (in terms of how much they liked, agreed, and found them reasonable). And they also categorized their own stance towards the 8 positions, both before and after exposure to the communication (in terms of latitudes of acceptance, rejection, and non-commitment—specifically, they indicated which of the eight positions best represents their view, which other positions they find acceptable, which position least represents their view, which other positions they reject, and the remaining ones which fall into none of those categories were judged to be positions of "non-commitment").

The authors results provide evidence for several tenets of the ego-based account. First, as can be seen in Figure XXX, people tend to evaluate communications consistent with their own position more favorably. And second, as can be seen in Figures XXX and YYY, the more discrepant that a piece of information is with one's position, the less likely that people are to move in the direction of that communication. This notion of discrepancy is explicitly formalized in Table ZZZ, where we see that people are expected to find positions acceptable that are near their own position, and are likely to reject positions that differ from their own view.

[17] **(A)** Since alcohol is the curse of mankind, the sale and use of alcohol, including light beer, should be completely abolished; **(B)** Since alcohol is the main cause of corruption in public life, lawlessness, and immoral acts, its sale and use should be prohibited; **(C)** Since it is hard to stop at a reasonable moderation point in the use of alcohol, it is safer to discourage its use; **(D)** Alcohol should not be sold or used except as a remedy for snake bites, cramps, colds, fainting, and other aches and pains; **(E)** The arguments in favor and against the sale and use of alcohol are nearly equal; **(F)** The sale of alcohol should be so regulated that it is available in limited quantities for special occasions; **(G)** The sale and use of alcohol should be permitted with proper state controls, so that the revenue from taxation may be used for the betterment of schools, highways, and other state institutions; **(H)** Since prohibition is a major cause of corruption in public life, lawlessness, immoral acts, and juvenile delinquency, the sale and use of alcohol should be legalized, and there was also an (although this is not included in all analyses) **(I)** It has become evident that man cannot get along without alcohol; therefore, there should be no restriction whatsoever on its sale and use

Figure 4: Receptivity to Communication as a Function of One's Own Position

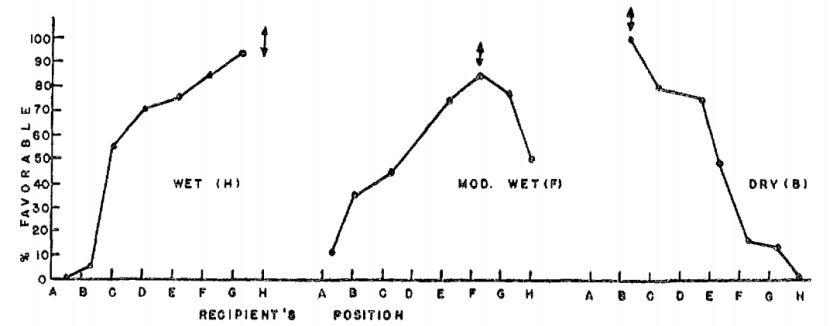


FIG. 1. PERCENTAGE OF FAVORABLE EVALUATIONS OF WET (H), MODERATELY WET (F), AND DRY (B) COMMUNICATIONS FOR *S*s HOLDING VARIOUS POSITIONS ON PROHIBITION (BASED ON MEAN ACCEPTABLE STATEMENT)
(Positions of communications indicated by arrow)

TABLE 2
OPINION CHANGE
Changes in Mean Acceptable Statement for *S*s with Differing Initial Stands

Group	*N*	Before Comm.	After Comm.	Change in Direction of Comm.
Wet Communication (H)				
Drys	69	2.39	2.34	−.05*
Unselected	92	5.10	5.65	+.55*
Dry Communication (B)				
Wets	25	6.70	6.74	−.04
Unselected	87	5.90	5.78	+.14
Moderately Wet Communication (F)				
Drys	114	2.17	2.26	+.09

* Difference between changes:
$p = <.03$ (one tail).

TABLE 3
OPINION CHANGE
Percentage of *S*s Changing in Direction of Communication or in Opposed Direction

Group	*N*	Change in Direction of Comm.	No Change	Change in Direction Opposed to Comm.	Net Change
Wet Communication (H)					
Drys	69	% 27.5	% 49.3	% 23.2	% +4.5*
Unselected	92	52.2	23.9	23.9	+28.3*
Dry Communication (B)					
Wets	25	24.0	56.0	20.0	+4.0
Unselected	87	40.2	33.4	26.4	+13.8
Moderately Wet Communication (F)					
Drys	114	31.6	49.1	19.3	+12.3

* Difference between changes:
$p = <.04$ (one tail).

TABLE 4

HYPOTHETICAL LATITUDES OF ACCEPTANCE AND REJECTION OF *S*S HOLDING EACH POSITION

(Columns show latitude of acceptance (strongly accept plus accept), latitude of rejection (strongly reject plus reject) and positions neither acceptable nor unacceptable to *S*s holding given positions.)

Rating Positions	Own Position								
	A	B	C	D	E	F	G	H	I
A	√√	√	0	×	×	××	××	××	××
B	√	√√	√	0	×	×	×	×	×
C	0	√	√√	√	0	×	×	×	×
D	×	0	√	√√	√	0	×	×	×
E	×	×	0	√	√√	√	0	×	×
F	×	×	×	0	√	√√	√	0	×
G	×	×	×	×	0	√	√√	√	0
H	×	×	×	×	×	0	√	√√	√
I	××	××	××	××	×	×	0	√	√√

Code: √√ = strongly accept; √ = accept; 0 = neither accept nor reject; × = reject; ×× = strongly reject.

A similar result is presented by Lord, Ross, & Lepper (1979) twenty years later. They begin their article with a Francis Bacon (1620) quote which shows just how far back the origins of this idea trace.

> *The human understanding when it has once adopted an opinion (either as being the received opinion or as being agreeable to itself) draws all things else to support and agree with it. And though there be a greater number and weight of instances to be found on the other side, yet these it either neglects and despises, or else by some distinction sets aside and rejects, in order that by this great and pernicious predetermination the authority of its former conclusions may remain inviolate.*

They add: “Thus, there is considerable evidence that people tend to interpret subsequent evidence so as to maintain their initial beliefs” (p. 2099).

The authors maintain that the underlying mechanism driving their effects is the process of “biased assimilation”—in the authors characterization, the tendency of people to ignore, dismiss or diminish evidence inconsistent with their pre-conceptions and existing theories, and their equal tendency to attend and give weight to evidence consistent with their pre-conceptions and existing theories (p. 2099). As for why *that* happens, the authors do not explain.

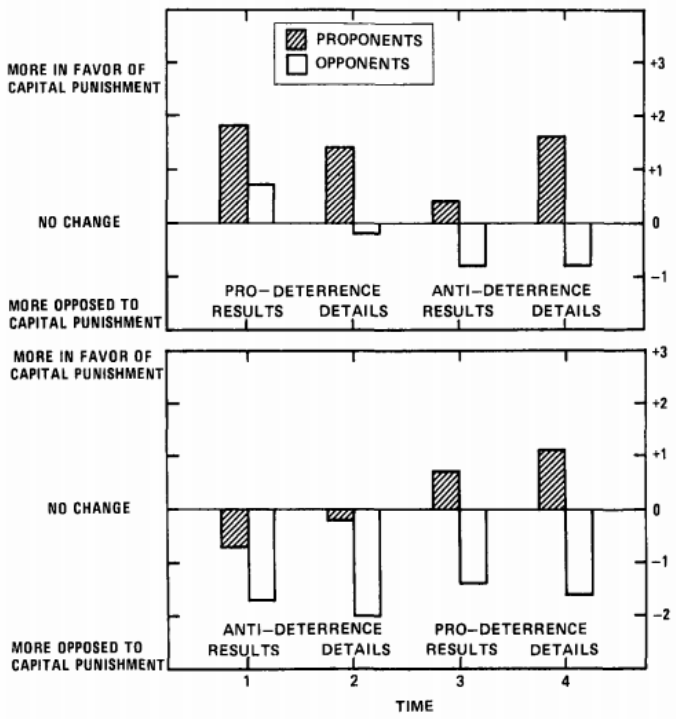


Figure 1. Top panel: Attitude changes on capital punishment relative to start of experiment as reported across time by subjects who received prodeterrence study first. Bottom panel: Attitude changes on capital punishment relative to start of experiment as reported across time by subjects who received antideterrence study first.

3: Discrepancy & Learning

---

learning

Learning is the process by which we acquire new information and update our views and mental models of reality. Thus, it is another natural place to look to understand how we respond to discrepant information.

In the cognitive and neuroscientific literature on learning, a distinction is made between two types of learning: self-organized and error-driven (O'Reilly, Munakata, Frank, Hazy, & Contributors, 2012). These differ primarily in time scales over which updating takes place. As O'Reilly et al. (2012) explain, the self-organized learning is learning that takes places over a longer period of time, and seeks to extract "statistical regularities" to develop an "internal model" of the world. Meanwhile, the error-driven learning takes place over very short time scales, and is drive more directly and immediately by immediate comparisons between expectations and actual outcomes. This type of learning is closely tied to things like curiosity and surprise, seems to be more active in children, and is likely driven by neuromodulators like dopamine, norepinephrine, and acetylcholine.

Learning Rules Across the Brain

	Learning Signal			*Dynamics*		
Area	**Reward**	**Error**	**Self Org**	**Separator**	**Integrator**	**Attractor**
Basal Ganglia	+++	---	---	++	-	---
Cerebellum	---	+++	---	+++	---	---
Hippocampus	+	+	+++	+++	---	+++
Neocortex	++	+++	++	---	+++	+++

Table 7.1 Comparison of learning mechanisms and activity/representational dynamics across four primary areas of the brain. +++ means that the area definitely has given property, with fewer +'s indicating less confidence in and/or importance of this feature. --- means that the area definitely does not have the given property, again with fewer -'s indicating lower confidence or importance.

We can also look at how introductory textbooks breakdown learning. Schacter, Gilbert, Nock, & Wegner (2017) summarize the large buckets into which learning research has been historically categorized: classical conditioning and operant conditioning. Then, they also mention the important categories of: observational learning and implicit learning. It is worth briefly summarizing these and thinking about which category that updating and learning might fall into.

Classical and operant conditioning describe those cases in which we learn from experience.

One of the first (and most influential; although eventually wrong) models used to explain the major results on classical conditions is the Rescorla-Wagner (1972) model of classical conditioning. The key idea behind this model (and many subsequent, more advanced models of learning) is that learning is critically related to surprise. The more surprised that a cognitive system (human, dog, bonobo, sea slug) is, the more it learns. That is, the *more discrepant* a piece of information is with experience, the more likely we are to learn from it. They present a very simply model of learning, which I will go through[18].

The main outcome they are trying to explain is the conditioned response (e.g. drooling) in response to an unconditioned stimulus (e.g. a bell). This outcome is usually denoted with the symbol V. This can be thought of as a proportion of percent (that varies from 0 to 1, or 0% to 100%), where V = 1.0 would mean that 100% of the time, a presentation of the unconditioned stimulus (e.g. bell) elicits the unconditioned response (e.g. drooling in response to the bell). More generally, though, this can be thought of as the current state of learning of the association between the conditioned stimulus (bell) and unconditioned stimulus (food). In most cases, where there is perfect coupling between conditioned stimulus (bell) and unconditioned stimulus (food), complete learning would be when V reach 1.0.

The basic formula for a current level of learned association is

$$V = V_{old} + \Delta V$$

where V_{old} is the previous level of V, and ΔV is the change in V that occurred as a result of the previous trial. The important point, in terms of the question under consideration here, is that learning is a function of previous beliefs (V_{old}) and our reaction to new information (ΔV).

The most important part of the model then explains how we adjust to new information, by giving a formula for how ΔV is determined.

$$\Delta V = \alpha * \beta * (\lambda - V_{expected})$$

The first two parameters are just parameters that are set to vary between 0 and 1, and adjust the extent of learning that will take place, based on the intensity of the unconditioned stimulus (α_x), e.g. how easily noticeable say the piece of food is, and the intensity of the conditioned stimulus (β), e.g. how easily noticeable say the sound of bell is. They are constant across trials, simply adjusting the amount of learning that will take place for each trial by a constant amount and can simply be ignored for our purposes. The important part of the model is the third term ($\lambda - V_{expected}$). This formalizes the notion of surprise; λ represents the maximum possible value of the conditioned stimulus in the experiment (usually either 1 or 100%, meaning a complete pairing between the conditioned and unconditioned stimulus). And $V_{expected}$ is the organism's current expectation, based on all previous trials and experiences up until that trial. If there is only one conditioned and one conditional stimulus, then on the n^{th} trial, $V_{expected} = V_{n-1}$.

[18] These four sources were very helpful in walking for understanding the implementation of this (very simply) model: http://users.ipfw.edu/abbott/314/Rescorla2.htm, http://campus.albion.edu/wjwilson/files/2012/03/RWSimplified.pdf, http://www2.bcs.rochester.edu/courses/crsinf/153/03.pdf, https://www.youtube.com/watch?v=GWKU0bW4vdw.

(However, $V_{expected}$ is not denoted V_{n-1}, because there may be multiple conditioned or unconditioned stimulus, in which case $V_{expected}$ is the summed average across these.) Thus, in the simple case, the current state of knowledge on the n^{th} can be written as:

$$V_n = V_{n-1} + (\alpha * \beta * (\lambda - V_{n-1}))$$

And so, the larger the surprise or discrepancy with previous expectations ($\lambda - V_{n-1}$), the more learning that happens, and the more that the organism's beliefs change.

4: Discrepancy & What We Want To Be True

What do we mean by discrepant?

- Discrepant with our existing attitudes/beliefs (lit.: attitude change)
- Discrepant with what we already know to be true/have learning (lit.: learning)
- Discrepant with what we want (lit.: good news v. bad news literature)

Discrepant from what we want (self-relevant good news v. bad news literature)

One way that information can be discrepant is that it differs in the extent to which we want it to be true.

With regard to self-relevant information, Sharot & Garrett (2016) presents apparently robust evidence that people update more in response to desirable information than undesirable information. It is worth unpacking the exact nature of this task the exact way in which the study is implemented. The task is in fact very simple. Participants are presented with 80 different possible "life events" (e.g. developing a kidney stone, suffering a break-in). They are then asked to asked to estimate the likelihood that such an event will befall them. They are then presented with an actual estimate that such an event will befall them. Finally, they are asked to re-estimate the likelihood that this event will befall them, in light of the information they just received. Desirable information ("good news") is thus either a case where a bad event (e.g. burglary) is less likely to happen than participants originally estimated, or where a good event is more likely to happen than participants originally estimated. Undesirable information is the opposite (negative events turn out to be more likely than thought, good events turn out to be less likely than thought). The key empirical finding is that the absolute magnitude that people update is larger when they receive desirable information than undesirable information.[19] Another notable finding is that when trying to model participants final models, a model which uses two separate parameters for the learning rate (one parameter when the information updated on is desirable, and another when the information updated on is undesirable) performs significantly better than a model with only one parameter for the learning rate (i.e. no differentiation is made between desirable and undesirable information, and the same learning rate parameter is used for both).

[19] Also note that a possible alternative explanation that people have greater memory for desirable information is ruled out in two ways: (1) the effect holds both when (a) final estimates are elicited while the empirical probabilities are still visible to participants and (b) when they are elicited slightly after those empirical probabilities are no longer visible to participants and thus they must rely on memory, and (2) in their paradigm there is no difference in memory for desirable and undesirable information, and it doesn't account for any of the observed effects.

Figure 5: Differential Updating of Beliefs in Response to Desirable ("Good") and Undesirable ("Bad") News

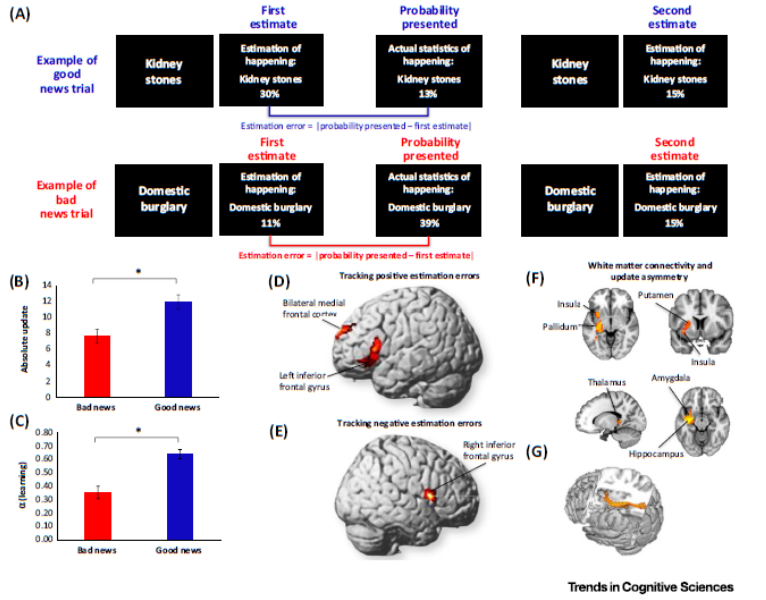


Figure 1. Belief Update Task: Brain and Behavior. To test for asymmetric belief formation and quantify its extent we have recently developed the belief update task. (A) Participants are presented with approximately 80 different life events and are asked to estimate their likelihood of experiencing each event. They are then presented with the average likelihood of the event occurring to someone like them and are asked to re-estimate their likelihood. (B) Participants adjust their beliefs to a greater extent when they receive good news (i.e. that a negative event is less likely to occur than expected) compared to when they receive bad news (i.e., that a negative event is more likely to occur than expected) [30]. The results are evident both when the re-estimate is elicited shortly after information presentation [30] and when elicited while information is still on screen [38], thus eliminating the possibility that the results are mediated by memory differences for the information. Indeed, memory for the information provided does not differ for good and bad news, and the results hold when controlling for participants' prior estimates, past experience, and other stimuli specific features. (C) The use of separate learning parameters, one for good news (α_G) and one for bad news (α_B), leads to a better fit with the data than a model using a single learning parameter. The learning parameter (α) is derived from this equation: second estimate = first estimate + α(estimation error), where estimation error is the difference between the first estimate and the information given. The graphs in (B,C) are generated from the combined data of 30 healthy participants across two studies [30,41]). (D) Estimation errors for good news correlate with blood oxygenation level-dependent (BOLD) signal in the left inferior frontal gyrus (IFG) and the medial frontal cortex (MFG). Figure adapted from [30]. (E) Estimation errors for bad news correlate negatively with BOLD response in the right IFG. Figure adapted from [30]. (F,G) Diffusion tensor imaging reveals that individuals with greater positive update bias have stronger white matter connectivity between the left IFG and left pallidum, left insula, left putamen, left amygdala, left hippocampus, and left thalamus. Figure adapted from [42].

The Sharot & Garrett (2016) shed some relevant light on the question of whether we update less in the face of more discrepant information. For one, the data could be re-analyzed to examine whether indeed larger discrepancies (between initial estimates and empirical rates) lead to less updating (I suspected the opposite). But, more to the core of their claim, their studies make speak to the extent of updating in light of affective, and self-relevant information. And they show that a given piece of information is less likely to be incorporated into our beliefs, the more that it is discrepant with what we want (i.e. undesirable as opposed to desirable). Thus, in fact, discrepancy can be thought of in this sense (emotional discrepancy with what is desired). In this sense then, greater emotional discrepancy leads to less updating.

5: Big Lies, from a Socio-Political Perspective

META INTRO

There are two pits at the core of this question, both of which are interesting and worth spending time to explore and attempt to explain. One question is social and political, asking why a population that is lied to can seemingly come to collectively believe that lie. The other is more purely cognitive, asking: as a piece of information is increasingly at odds with our existing beliefs and experience, how much do we tend to believe it and update in light of it?

INTRO (Socio-political)

Propagandist, it has been said, sometimes advocate for the "big lie." The idea behind the big lie is that people are more likely to believe a deception if is large. (That is, more discrepant with one's knowledge, experience, or beliefs.) The advice is usually presented as a sort of cynical wisdom about the nature of people's beliefs. Even a moments reflection, however, calls this wisdom into question. After being accused murdering Saudi Arabian journalist Jamal Khashoggi after his disappearance upon entering a Saudi embassy in Turkey on October 2nd 2018, Saudi Arabia's crown prince Mohammad bin Salman and his government did not claim that perhaps it was a group of drunk Parisian high school students who snuck into the embassy and murdered Mr. Khashoggi as a prank. They stuck with stories that hemmed to the truth. First, they tried to simply deny it—plausible enough, until more evidence emerged. Then they claimed that the murder was the result of "rogue" agents within their own government, again stretching credulity, but not impossible. Then, as yet more evidence emerged, they claimed that it was the result of a fist fight after an interrogation gone wrong, again veering even closer to the truth. Similarly, when four Russian citizens were found with computer equipment near the Hague in and were accused of trying to hack the Organization for the Prohibition of Chemical weapons in March of 2018, Russian Foreign Minister Sergey Lavrov responded by claiming that these men were on a "routine" trip, not that perhaps they were American spies disguised as Russian nationals who purposefully got themselves caught in an elaborate setup in order to embarrass the Kremlin.[20] Guilty defendants do not resort to outlandish claims when trying to defend their innocence, instead seeking plausible alibies or claiming not to "remember"—all lies of a subtler variety. If one were to developing a general psychological principles rule for getting away with lying, it would seem the opposite of the "big lie" would be a much better rule—make your lies believable, small, and not too discrepant with the what is expected.

---

[20] Yet, in both cases, the international community did not believe these believable lies. So one might argue that they might have had more luck had they gone with a "big lie." Perhaps. But this seems unlikely. Another nuance to these particular cases is that there was pretty damning evidence that each government had committed the act they were accused of. And further, this is perhaps a slightly different case than the one imagined by the propagandists, who are perhaps trying to offer advice about how to convince a citizenry of a broader political point.

Nevertheless, are there cases where a "big lie" might be more successful than a "small one"? What are the general psychological principles that govern how we deal with discrepant information?

theoretical models in political science

Political scientist Andrew Little (2017) summarizes previous work in political science and economics, which seeks to explain why a government would go through the effort of lying to its population, especially one that is often sophisticated enough to be aware that they are being lied to. Some models he shows, for example, explain this as functional because it serves as a signal of the various commitments of the government—even if they do not actually believe the lies. Little (2017) explores an even simpler model. His own succinct summary does it the most justice: "motivated by a wide variety of empirical results spanning several disciplines, I explore the consequences of a simpler explanation: politicians lie because some people believe them."

empirical evidence from political science

Models are useful, but of course, they only go so far. It is important to examine how they actually line up with reality. There are two interesting and relevant lines of research to explore in this regard. One looks at behavior of individual people in lab-type settings, to see if their communication comports with the outlines of the theory. The other examines the effectiveness of lying on the institutional. Namely, does propaganda and lying work?

On the first point, there is interesting evidence that people seem to be overly honest and open in their communications, compared to what would be optimal. Some evidence on this comes from Cai & Wang (2006) and Wang, Spezio, & Camerer (2010) who examine the actual amount of communication between parties in a strategic communicative interaction—the latter study even employing eye-tracking to measure pupil dilation among parties that deceive. The finding here seems to be that people are overly honest in strategic communication; they reveal too much. Further, people seem to "go along" with what others say in communicative contexts to "get along" socially with their communicative parnets (Chen, Shechter, & Chaiken, 1996).

On the institutional front, there is evidence that indeed—yes—propaganda works. For example, Enikolopov, Petrova, & Zhuravskaya (2011) compare electoral outcomes in 1999 parliamentary elections in Russia between areas that had access to a (and the only) TV news channel that was independent from the government, and those that didn't. Under their models and estimates, access to independent media leads to 8.9% less support for the government party and increase in support for opposition parties of 6.3%.

6: Cheap Talk in Game Theory

Before touching the empirical evidence, it is useful to outline the conditions that exist in many communicative contexts, the different configurations of incentives that might exist, and how that all affects communication. For example, what is the relationship between the parties communicating? Does one have more information than the other? What are their incentives and do they align? There is an important literature exploring this latter question, which originates from economics and game theory. In a now classic paper, Crawford & Sobel (1982) outline a model of "cheap talk," which explores and delineates strategies for communication between the simplest bare bones case of two communicators (a send and receiver), where the sender has more information than the receiver, and communication is direct and costless. They show that there are essential three primary equilibrium states: one, which results in simply "babbling" (that is, communicating but not saying anything), and the other two are complete communication, and no communication at all.

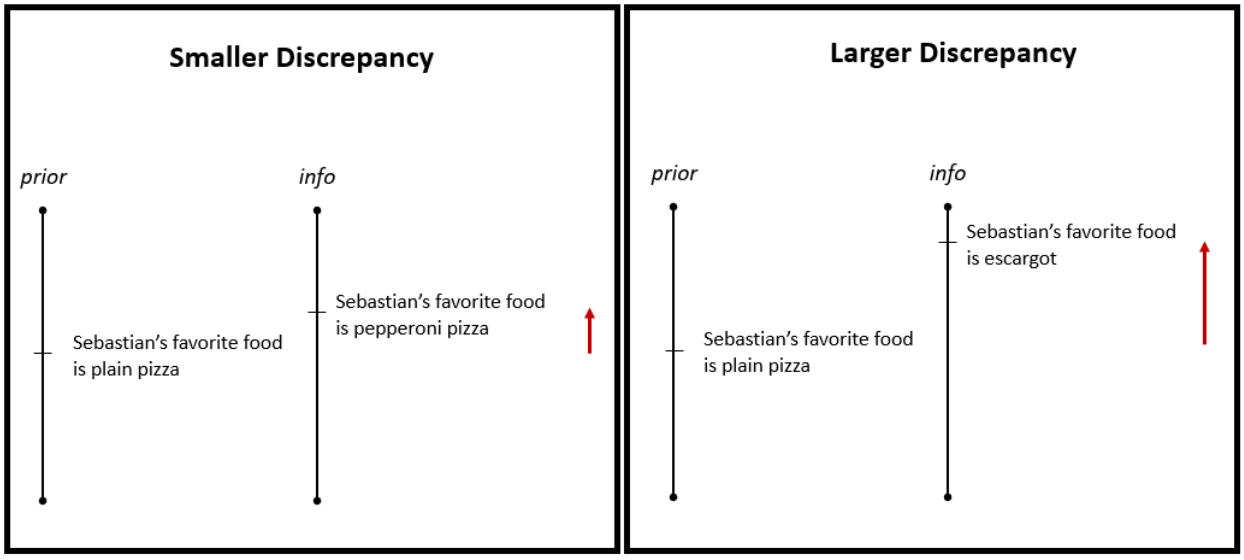
7: Attempted Formal Model of Belief Change (in Response To Discrepant Info.)

Part 1: A Framework for Belief Change (Inputs)

We can think of any model as a mechanism which describes how certain "inputs" are turned into certain "outputs". At the outset, we only want to consider one input, that input on which the question is focused—what we will call *discrepancy.* To start, discrepancy can be thought to describe the size of the gap between two things:

- our prior belief ("prior")
- a piece of new information ("info")

The extent of this discrepancy can vary in size. This is illustrated in the figure below. The size of the red arrow represents the size of the discrepancy.



Part 1: A Framework for Belief Change (Continuous and Categorical Beliefs)

Let's make a sort of distinction at the outset that will help keep things simple later.

Some beliefs can be thought of as "continuous" and some beliefs can be thought of as "categorical".

Meanwhile, the examples in the first figure can be thought of as beliefs that are categorical. That is, the nature of these beliefs are discrete propositions. A person must adopt a single proposition among a set of propositions. Either you think my favorite food is plain pizza, or pepperoni pizza, or escargot (or some other food). This does not mean that the degree of discrepancy between

one's prior belief and new information can't be described continuously—by this notion of discrepancy. Various propositions may differ in the extent to which they are discrepant with one's prior beliefs.

The examples used in the second figure (rollerblading) can be thought of as beliefs that are continuous. That is, the nature of the belief is a quantitative assessment of the amount that I like rollerblading. Someone may believe that I like rollerblading various amounts (e.g. I like it a lot, I like it a little, I don't like, I hate it). Many beliefs are of this nature. Commons examples might be beliefs about one's own skill in a domain (e.g. I am very smart, I am kind of smart, I am dumb, I am very dumb), or an assessment of another person (e.g. that stranger is very untrustworthy, that stranger is somewhat untrustworthy, that stranger is somewhat trustworthy, that stranger is very untrustworthy).

Part 1: Outlining Framework (Outputs)

Now let's turn to the "outputs" we are interested in explaining. We are interested in how two particular outcomes are related to discrepancy. One outcome is a notion we can call *believability*. This can be thought of as a probability or likelihood—something akin to the probability that a new piece of information is true. The second outcome that we are interested in can be called a person's *final belief*. This can be thought of as the belief a person ultimately adopts after discrepancy between one's prior beliefs and new information is resolved in some way.

Part 1: Outlining Framework (Outputs: Believability)

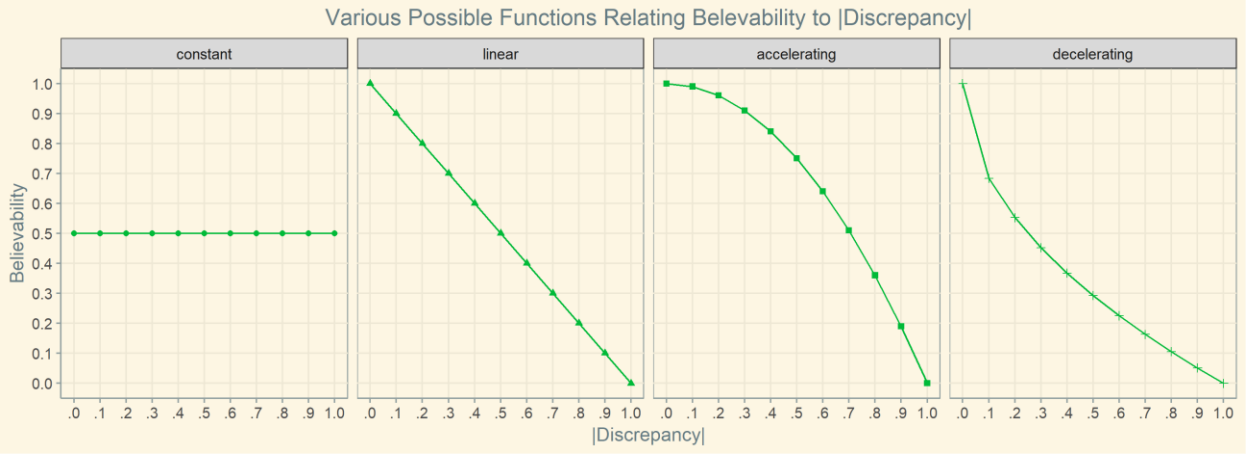
Let's start with *believability*. We have some notion that discrepancy and believability are related and how want to understand how exactly this is. Stated more formally, we are trying to define some function, which can be denoted as f_b, that describes the relationship between *discrepancy* and *believability*:

- $believability = f_b(discrepancy)$

We do not know the nature of this function. But we can begin by discussing various reasonable possibilities. What sort of properties would this function have? This simplest guess might be that such a function would have at least one key property: (1) the more discrepant a new piece of information is with our prior beliefs, the less we believe it. (To keep things simple for now, let's also assume that believability is only related to the size of the discrepancy and not its direction. That is, we only care about the absolute value of the discrepancy.) More formally then, this is the same as saying that believability is a monotonically decreasing function of the absolute value of discrepancy. In notation, this is the same as saying the following about the function f_b:

- $believability = f_b(|discrepancy|)$
- $\forall\, d_1, d_2 \in discrepancy, where\ |d_2| \geq |d_1|, it\ must\ be\ the\ case\ that\ f_b(d_2) \geq f_b(d_1)$

What do some functions that satisfy these basic properties look like? Some simple examples are shown below. These functions can be grouped into categories of "constant", "linear", "accelerating", and "decelarting".[21]



What is the psychological interpretation of these functions? All functions but the first one (i.e. the functions "linear", "accelerating", "decelerating") have the property that as the discrepancy of a piece of new information grows (it becomes more inconsistent with our prior beliefs), its believability decreases. These three functions only differ in the exact manner in which that decrease occurs. In the "linear" function, believability decreases the same amount for each unit of discrepancy. The "accelerating" function makes the claim that as the discrepancy of a new information grows bigger and bigger, it's believability actually drops at a faster and faster rate. The "decelerating" function corresponds to the claim that as a piece of new information becomes more discrepant, it is believed less and less, but the rate at which this decrease occurs actually slows down as the size of the discrepancy increases. (The distinctions between these three basic decreasing functions may not prove too meaningful for empirical purposes, as their empirical distinction would require us to be able to measure discrepancy using an interval scale, which may prove difficult to generate.) Finally, one other possibility for f_b is worth considering—this is a believability function that is "constant". Psychologically, this corresponds to the idea that a piece of information is judged equally believable no matter how discrepant it is from our prior beliefs. This is akin to maintaining a constant state of credulity. An example of such a believability function might describe a student's level of credibility in a classroom. No matter how counterintuitive or surprising it is what the professor says, the student believes it to an equal extent.[22]

[21] Note that the "constant" function is not strictly decreasing, but it fits the formal criteria we outlined (over the interval, as discrepancy increases, believability either increases or stays the same; in the constant case were ae jut in the extreme case where believability always stays the same). More importantly, I included it because it has a psychologically meaningful interpretation, which is described.

[22] e.g. The student believes the professor equally when the professor claims the somewhat intuitive (i.e. less discrepant) laws of classical Newtonian mechanic are true as when the professor claims that the completely counterintuitive (i.e. highly discrepant) laws of quantum mechanics are true.

Part 1: Outlining Framework (Outputs: Believability, Strict Big Lie Models)

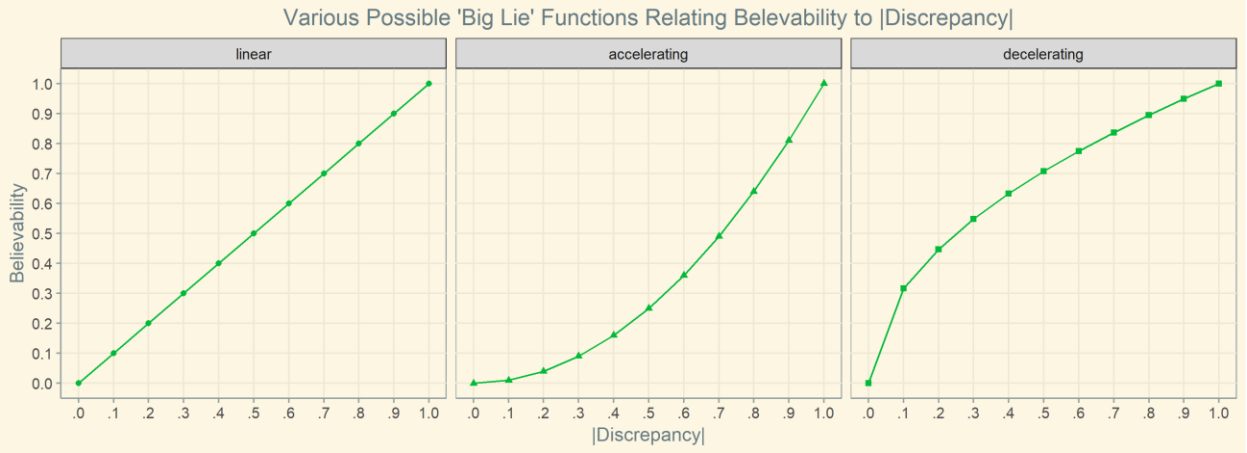
We have not yet discussed the other output we want to examine (*final belief*). However, this is the first point at which notions of the "big lie" come in to view.

Various interpretations of the "big lie" claim can be thought to call into question the very basic (monotonically decreasing) believability functions, f_b, outlined just before.

The most extreme version of the "big lie" claim could be taken to mean that over the full range of discrepancies, believability is always increasing as discrepancy is increasing. In notation, (again sticking with the notion that we only care about the absolute value of the discrepancy), this can be stated as:

- $believability = f_b(|discrepancy|)$
  - where $\forall\ d_1, d_2\ \in discrepancy, where\ |d_2| \geq |d_1|, it\ must\ be\ the\ case\ that\ f_b(d_2) > f_b(d_1)$

In the figure just below some simple functions satisfying this "always increasing" property are shown.



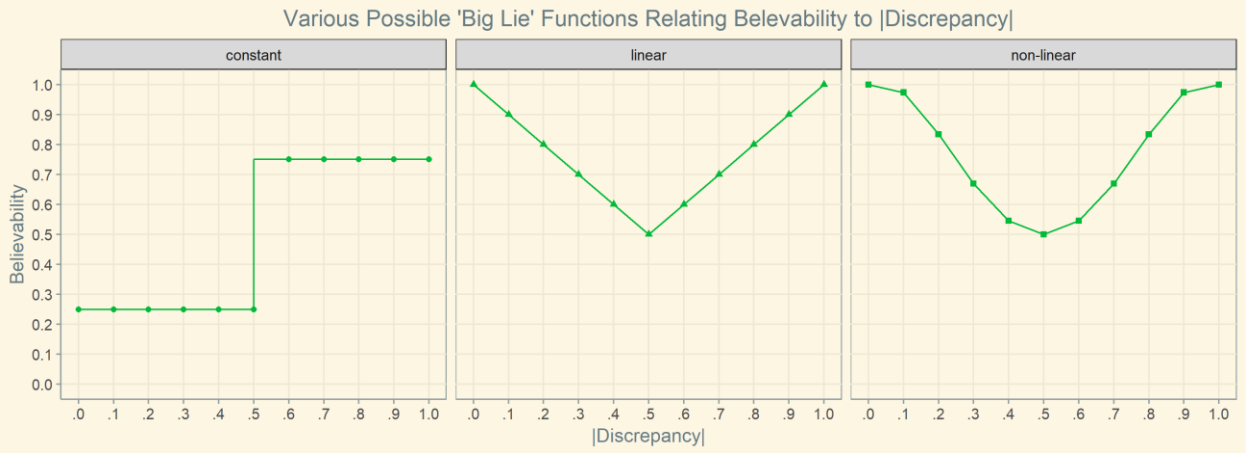
Again, we can ask: what is the psychological interpretation of these functions? These functions make the claim that as discrepancy increases, believability always increases. This seems like an obviously incorrect description of reality. It would correspond to a world where, without exception, teachers are likely to judge a student's claim that "I didn't do my homework because aliens abducted me" as more likely than "I didn't do my homework because my computer broke", which in turn would be judged more likely than "I didn't do my homework because I got sick last night" which again would be judged as more likely than "I didn't do my homework because I procrastinated so long that I didn't have time to finish it". Nevertheless, this is one reading of the "big lie" claim (and perhaps the most literal).

Part 1: Outlining Framework (Outputs: Believability, Strict Big Lie Models)

What might be a more reasonable reading of the big lie claim then? The next most reasonable and simple notion that comes to mind is that, at first, as discrepancy increases, believability decreases; but then at a certain point, things reverse, as discrepancy increases, believability increases. In formal notation:

- $believability = f_b(|discrepancy|)$
  - $where\ \exists\ d_{k,} \in discrepancy, where\ \forall\ d_1,\ d_2 \in discrepancy\ such\ that |d_1| < |d_k|\ and\ |d_2| < |d_k|\ and\ \ |d_2| > |d_1|, it\ must\ be\ the\ case\ that\ f_b(d_2) \leq f_b(d_1)$
  - $and\ where\ \forall\ d_1,\ d_2 \in discrepancy\ such\ that |d_1| > |d_k|\ and\ |d_2| > |d_k|\ and\ \ |d_2| > |d_1|, it\ must\ be\ the\ case\ that\ f_b(d_2) \geq f_b(d_1)$

Some functions that satisfy this property are shown in the figure below.



Psychologically, the notion that such functions aim to encapsulate is that at the beginning, discrepancy and believability interact in the "expected" way (negative relationship), but beyond a certain point, claims are so outrageous that they are actually judged as more likely to be true (positive relationship).[23] How could this be the case? In Hitler's telling, people will believe big lies because "*it would never come into their heads to fabricate colossal untruths* [themselves], *and they would not believe that others could have the impudence to distort the truth so infamously*." To go with the homework example, this is the claim that a teacher may be less likely to believe a student if they say "I didn't do my homework because my computer broke" than if they say "I didn't do my homework because I got sick" (less discrepant with prior beliefs), but there are claims more discrepant than the computer breaking claim that they might yet be more likely to believe, such as if the student claimed "I didn't do my homework because my mother committed suicide last night."

[23] The "constant" case is slightly different, although with similar meaning. Here, there is simply some point at which believability "jumps up".

Two things become apparent from this line of thought. First, it suggests that the inputs in our believability function are impoverished (so far, we only have one input: discrepancy). While descriptively, it is the case that case just outlined could manifest as a certain discrepancy "threshold" (beyond which, believability is judged more likely as discrepancy grows), there is no explanation for why this threshold exists and why the trend reverses beyond this threshold. So what additional factor could we turn to explain this reversal? One very reasonable possibility, suggested by the Hitler quote itself, is that when people hear claims which vary in the extent to which they are discrepant from one's prior beliefs, people engage in some sort of causal or attributional reasoning about the processes that generated those claim (Choi, Nisbett, & Norenzayan, 1999; Kelley, 1973; Malle, 1999). "Why would this student have said this?" the teacher likely asks himself in response to any of the students claims. In either the computer breaking or the getting sick case, the teacher likely has two major reasons come to mind for why the student said this: "either [1] this student is lying and it is not the case that they actually [got sick/had their computer break] and they are just using this to cover up the real reason they didn't do their homework (e.g. procrastination, laziness, or [2] this student is telling the truth and [they got sick/had their computer break] right when this assignment is due". As the claims become more and more discrepant with the teacher's prior beliefs (e.g. people aren't usually sick, but it's not uncommon to get sick; it is even less often the case that people's computer break, although this too does happen), the claims are judged as less and less believable. Likewise, when the student claims they didn't do their homework because their parent committed suicide two major causes appear in mind for the teacher, in this case: "either [1] this student is a horrible person who has absolutely no moral principles and is willing to make egregious (and later refutable) claims to get out of a lousy homework assignment or [2] this student's parent actually committed suicide last night and so they were too distraught and pre-occupied with that entire situation to do their homework." If they choose to believe the student in the suicide case, they have judged the more charitable causal explanation as more likely. Thus, some sort of causal consideration explains they the relationship between discrepancy and believability reverses at some point.

Even putting aside the failure of this model to consider causal reasoning as a factor that determines believability, there is an even simpler problem with this model. It suggests that after a certain point of discrepancy is reached, statements are always judged more believable. This would imply that the teacher would judge the claim "I didn't do my homework because I received a phone call from the President last night requesting that I spend the night consulting the Secretary of State on an urgent international crisis" as *even more* likely than the suicide claim (because it is likely even more discrepant with the teacher's prior beliefs). It seems impossible that things would work this way either.

Part 1: Outlining Framework (Outputs: Believability, Rollercoaster Model)

So where to go from here? It is clear that we need to modify the believability function in some way to account for the possibilities just noted. In making any modification, we should seek to keep things as simple as possible and make as few changes as we can. This suggests two major sets of possibilities.

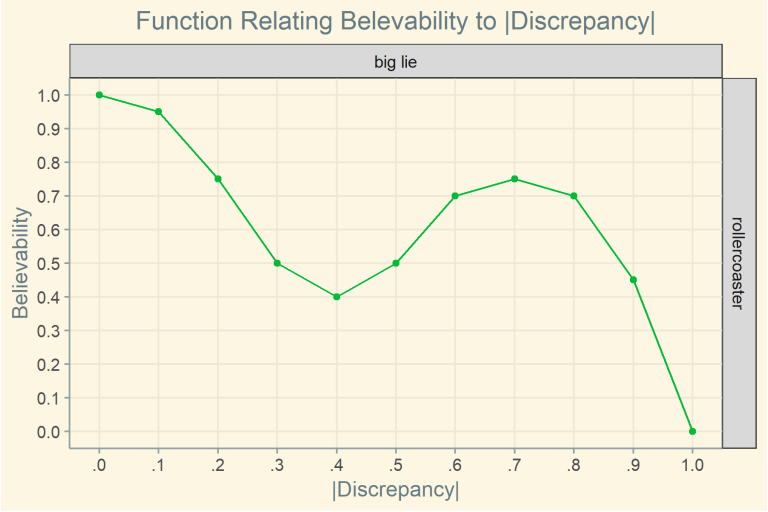
- Do not add any more inputs to the believability function (i.e. keep it only a function of discrepancy magnitude), but modify the shape of the function to at least account for the

fact that it doesn't make sense that there exists a certain point after which believability is always increasing.
- Change our model in such a way that it accounts for the exceptions just noted. This may either take the form of considering other parameters ("inputs") or more fundamentally restructuring the model.

Since it is simpler, let's start with the first possibility—modifying the shape of the believability function. An even more tempered version of the big lie claim in accordance with this approach then might simply be to say that there exists segments over the entire discrepancy range, over which the relationship between discrepancy and believability is positive, (i.e. consistent with the "big lie" notion, that as discrepancy increases, believability also increases), but eventually at the very end of the discrepancy range, it is the case that believability decreases as discrepancy increases. In mathematical notation:

- $believability = f_b(|discrepancy|)$
  - $where\ \exists\ d_a, d_b \in discrepancy, such\ that\ |d_a| < |d_b|, where\ it\ is\ the\ case\ that\ \forall\ d_1, d_2\ \in discrepancy\ such\ that\ |d_a| < |d_1| < |d_b|\ and\ |d_a| < |d_2| < |d_b|\ and\ |d_1| < |d_2|, f_b(d_2) > f_b(d_1)$
  - $and\ where\ \exists\ d_x, d_y \in discrepancy, such\ that\ |d_x| < |d_y|, where\ it\ is\ the\ case\ that\ \forall\ d_1, d_2\ \in discrepancy\ such\ that\ |d_x| < |d_1| < |d_y|\ and\ |d_x| < |d_2| < |d_y|\ and\ |d_1| < |d_2|, f_b(d_2) < f_b(d_1)$
  - $and\ where\ \nexists\ d_i \in discrepancy\ such\ that\ |d_i| > |d_y|$

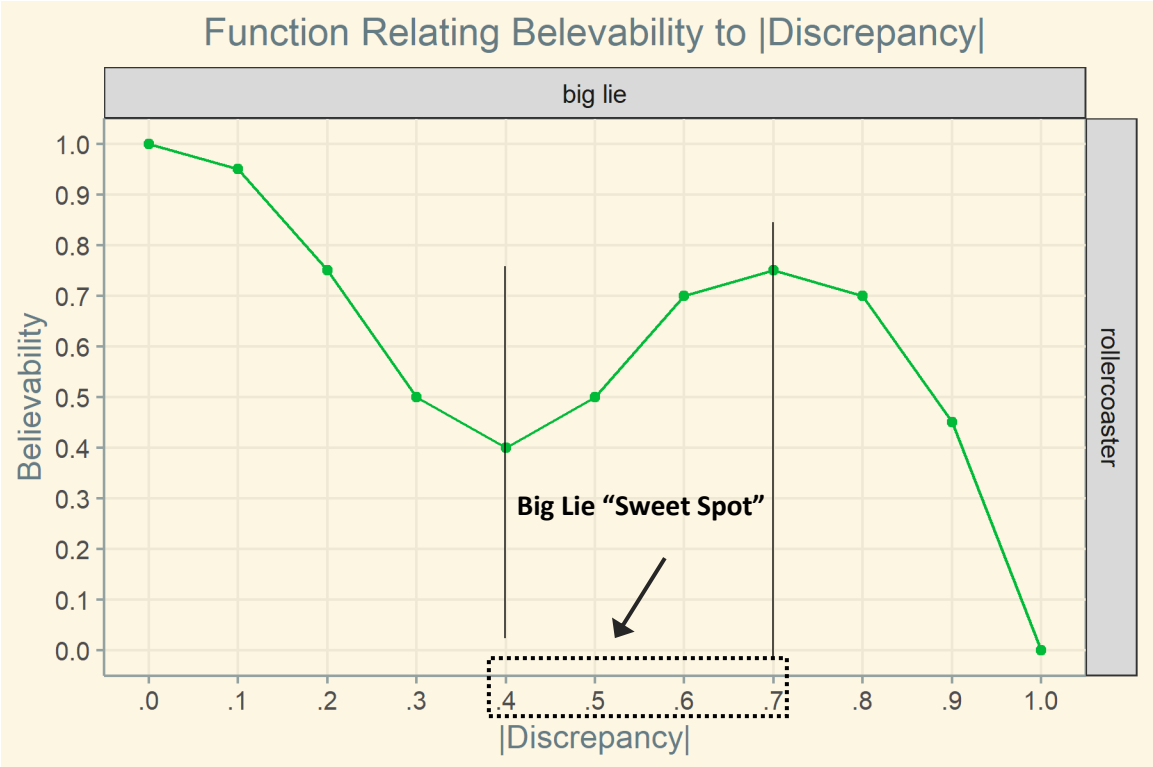
An example of a function that satisfies this property is shown below.



The psychological interpretation of this function could be something like the following. At first, as claims become more discrepant from one's prior beliefs, they are judged as less believable (e.g. in the homework case, moving from the being sick lie, to the computer lie). Then, eventually one's state of disbelieving is still going down as claims are more discrepant, but the rate at which this decrease occurs starts to "slow down" (e.g. where discrepancy is between

about 0.2 and 0.4 in the above figure)[24]. Eventually, the direction of the relationship actually reverses (e.g. where discrepancy is between 0.4 and 0.7 in the figure). After this, we reach a point (discrepancy > 0.7) where the trend reverses again (and discrepancy and believability are again negatively related, and the rate of decrease is increasing at an even faster rate).

We could call this the "medium lie" or "sweet spot" interpretation. Psychologically, the idea is that there is a set of big lies that are actually "big but not too big"—they are not so crazy that they are dismissed out of hand (e.g. in the homework example, the President or aliens excuse) but they are believed more (e.g. the suicide claim) than other less discrepant lies (e.g. the compute breaking or being sick lie).



Part 1: Outlining Framework (Modifying the Believability Function to Account for Counterexample)

The only minor problem with the model on the previous page is that it explains absolutely nothing and is little more than a squiggly line. We confused ourselves by thinking we were outlining a "model" for believability when in fact we were just overfitting a regression to the data points of our memories.

[24] Mathematically, what I am trying to point out is that over the first interval in which discrepancy is decreasing (discrepancy = 0.0 to discrepancy = 0.4), the concavity of the curve changes—from concave to convex. Several more concavity reversals occur later as well.

The shape of this "rollercoaster" believability function may "account" for the previous thought experiment that proved our "kinked" believability functions inadequate (i.e. those class of functions that changed direction once; they started by decreased and then increased until the end of the interval after a certain point). However, the counterexample revealed something more fundamental. Believability seems to be determined by something more than pure discrepancy from one thought. We reasoned that the teacher may be more likely to believe a student if that student said "I couldn't do my homework because my mother committed suicide" than if that student said "I couldn't do my homework because my laptop broke", even though the former may be more "discrepant" in some sense.

So how can we think of the believability function in a way that accounts for this fact? And again, let's aim to resolve this in the simplest way possible.

The simplest solution I can think of merely takes note of the fact that people hold many prior beliefs, not just one. In the thought experiments we have been thinking through, implicitly we are sort of thinking of a case with one prior belief (e.g. "Billy didn't do his homework because he is lazy") and one discrepant piece of information (e.g. "Billy said he didn't do his homework because his computer broke").

In reality, we have many prior beliefs. And thus any new piece of information is not simply discrepant to varying degrees with our one prior belief, but is discrepant to varying degrees with all of our prior beliefs (or at least all those prior beliefs that "come to mind" in some sense).

Before outlining how this might work, we will first need to talk about the final output that we are concerned with explaining: *final beliefs.*

Part 1: Outlining Framework (Outputs: Final Belief)

The final outcome we are interested in is people's final beliefs.

Here it's worth revisiting the distinction we made between what we called continuous and categorical beliefs.

Let me outline the various basic inputs for each. Then I will decompose in more detail the categorical cases. Then I will continue on to the case of continuous beliefs.

The function that determines final beliefs in the case of categorical beliefs takes only one input—*believability*—the output of out believability function, f_b, outlined before. The only difference for the case of continuous beliefs is that the final belief function takes in two inputs—*believability* and *discrepancy*.

For categorical beliefs:

- $final\ belief = f_{cat}(believability)$

For continuous beliefs:

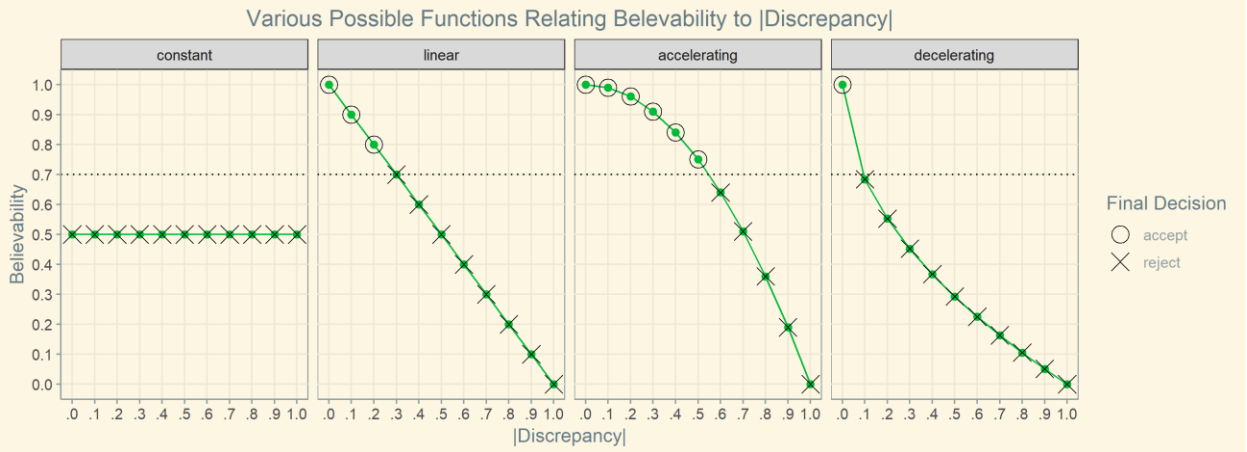
- $final\ belief = f_{cont}(believability, discrepancy)$

Part 1: Outlining Framework (Outputs: Final Belief for Categorical Beliefs)

So how might this function that determines final beliefs work in the case of categorical beliefs? One very basic idea is that there is simply some threshold. If the believability that resulted from the believability function, f_b, (which itself was determined by the level of discrepancy) is at or below the threshold, you stick with your prior belief. On the other hand, if the believability that was output by the believability function, f_b is above the threshold, you adopt the new belief suggested by the discrepant piece of information. In mathematical notation:

- $final\ belief = f_{cat}(believability)$, where
  - $f_{cat}(believability) = \begin{cases} if\ believability \leq threshold, \\ stick\ with\ prior\ belief \\ if\ believability > threshold, \\ adopt\ \ belief\ suggested\ by\ discrepant\ information \end{cases}$

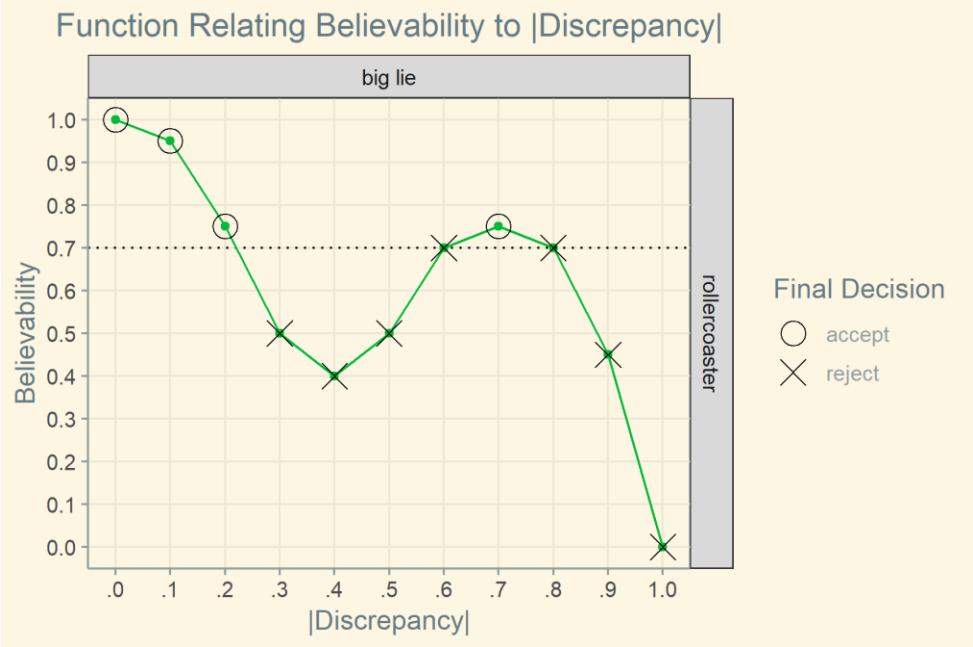
In the figure below, I show the output of an f_{cat} final belief function where the threshold hold is 0.70, as it interacts with the most simple monotonically decreasing believability functions considered above. (The notation is a little inconsistent, but basically, any time the believability function outputs a value above 0.70, we "accept" that the discrepant information is true and adopt the belief suggested by the discrepant information. Any time the believability function outputs a value at or below, 0.70, we "reject" the discrepant information and stick with our prior belief.)



Again in psychological terms, imagine back to the case of the homework excuses. Imagine the teacher finds out here student didn't do their homework, and the teachers prior belief is "The student didn't do their homework because they are lazy." The student then could make any of the following claims (which increase in the extent to which they are discrepant with the teacher's

prior belief): “I didn’t do my homework because I am lazy” (not discrepant at all; discrepancy = 0.0), “I didn’t do my homework because I got sick” (discrepancy = 0.2), “I didn’t do my homework because my computer broke” (discrepancy = 0.3), “I didn’t do my homework because my mother committed suicide” (discrepancy = 0.7), “I didn’t do my homework because I was abducted by aliens” (discrepancy = 0.90). As we see, any claims which lead to a level of believability at or below the threshold of 0.70 lead the teacher to “reject” the students claim and maintain her prior belief that the student didn’t do their homework because the student is lazy. On the other hand, any claims that results in a believability above 0.70, lead the teacher to “accept” the students claim and adopt a new belief suggested by this claim (e.g. the students claim). (Note that the new piece of information which lead to the discrepancy from one’s prior belief is not always necessarily the same as the actual final belief that might be arrived at. For example, if I have the “prior” belief that my friend Amir’s favorite food is plain pizza, but then I am confronted with the “new information” that he chose pepperoni pizza when we went out together to eat, I may adopt the “final belief” that his favorite food is pepperoni pizza.)[25]

Here is the output of this same f_{cat} final belief (threshold = 0.70), as it interacts with the final “rollercoaster” believability function we considered earlier.



Here we see that, because believability is not strictly decreasing, a more discrepant piece of information (e.g. at |discrepancy| = 0.70, such as the student claiming they didn’t do their homework because of a suicide) might lead to an adoption of a new belief while a less discrepant piece of information (e.g. at |discrepancy| = 0.50, such as the student claiming they didn’t do their homework because of a broken computer) might lead us to stick with our prior.

[25] i.e. notice that there are three “objects: prior belief = “Amir’s favorite food is pizza”; new info = “Amir ordered pepperoni pizza”; final belief = “Amir’s favorite food is pepperoni pizza”

Part 1: Outlining Framework (Considering an Additional Input to the Believability Function: Reliability)

Finally, because it will be useful later, let's also consider one additional parameter which might affect believability. This is some notion of reliability. Since information often comes from different sources, this might be thought of as the extent to which information from a source is trusted. This aims to capture the fact that within any given type of believability function, f_b, even though believability might decrease as discrepancy increases, for any given level of discrepancy certain sources are trusted more than others.

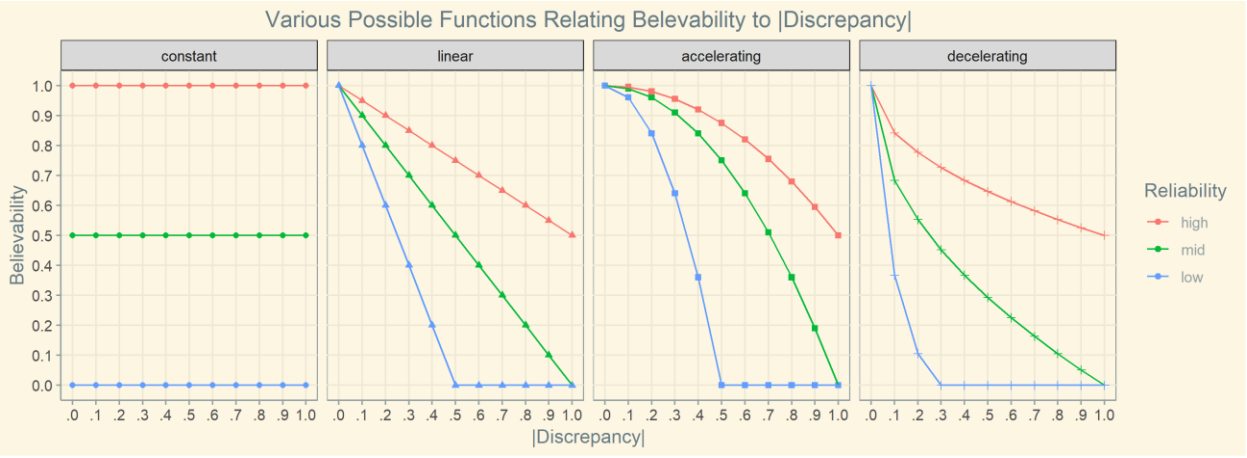
In notation, most abstractly we might simply say that we are adding a parameter to the believability function, f_b, such that believability is now a function of not only discrepancy, but also reliability.

- $believability = f_b(discrepancy, reliability)$

Let's just rely on some rank order notion of reliability, where the only requirement is that different sources can be ranked in terms of how reliable they are (e.g. we can distinguish between "high", "mid", and "low" reliability sources). The figure below aims to illustrate how believability may be affected by reliability for various possible believability functions.

This notion could be instantiated in at least one of two ways:

- As a simple horizontal shift
- As a some sort of change in the nature of the slope



Psychologically, the interpretation of the effect of reliability on believability is about the same for the functions "linear", "accelerating", and "decelerating". Consider the linear panel for a specific example. As a piece of information from a highly reliable source (pink line) becomes more discrepant with our prior beliefs, the less we believe it. Nevertheless, we always believe it more than a piece of information that is equally discrepant but comes from a less reliable

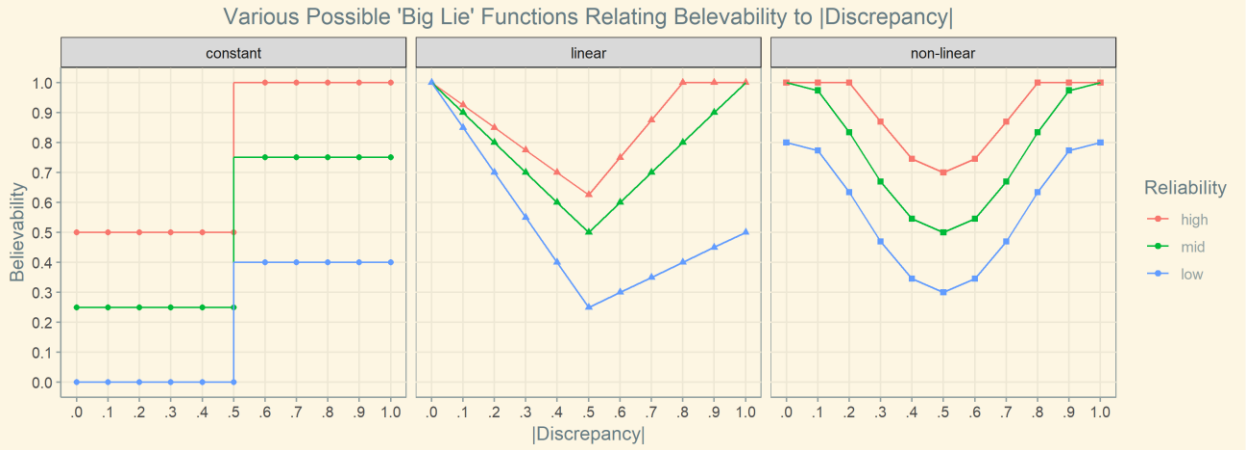
source.[26] To give a concrete example, imagine the claim that there will be a blizzard in Ithaca next week (discrepant from my belief that there won't be a blizzard). I will believe that claim more if I hear about it as an official email announcement from Cornell University (highly reliable, e.g. pink line) than if an undergraduate mentions it to me (e.g. green or blue lines). Nevertheless, for either of these sources, I will believe any given claim less and less, as that claim becomes more discrepant with my prior beliefs. For example, I will believe Cornell University less if they claim there will be an earthquake next week (even more discrepant from my prior beliefs) than if they claim there will be a blizzard next week. And I will also believe the undergraduate less if they claim there will be an earthquake rather than a blizzard. However, again, holding discrepancy constant, I will believe the earthquake claim more if I hear it coming from Cornell University rather the undergraduate. (This also points out the fact that there are cases where I may believe a less discrepant claim from an unreliable source more than a highly discrepant claim from a reliable source.)[27] In the case of a "constant" believability function, f_b, the interpretation of reliability is the same, but its manifestation is simpler. The interpretation is simply that there are different sources which vary in the extent to which people trust them, but they maintain the same level of believability now matter the level of discrepancy. For example, no matter how discrepant the claim from a highly reliable source, it is always believed the same amount (e.g. the pink line in the "constant" panel might describe a young child's belief of various claims their parent might make; no matter how crazy the claim, they believe the claim 100%; while the blue (low reliability) line in the "constant" panel might describe that child's belief of someone they don't like at school, no matter what that person says, even if it is hardly discrepant, they don't believe it at all.)

Before moving on to more reasonable interpretations of the "big lie" claim, one new question is suggested as soon as we start thinking about the idea that discrepancy and believability are positively related. This is the question of how our notion of reliability, or varying degrees of trust in a source, would operate in a context where believability actually increases with discrepancy.

"*It would never come into their heads to fabricate colossal untruths, and they would not believe that others could have the impudence to distort the truth so infamously*."

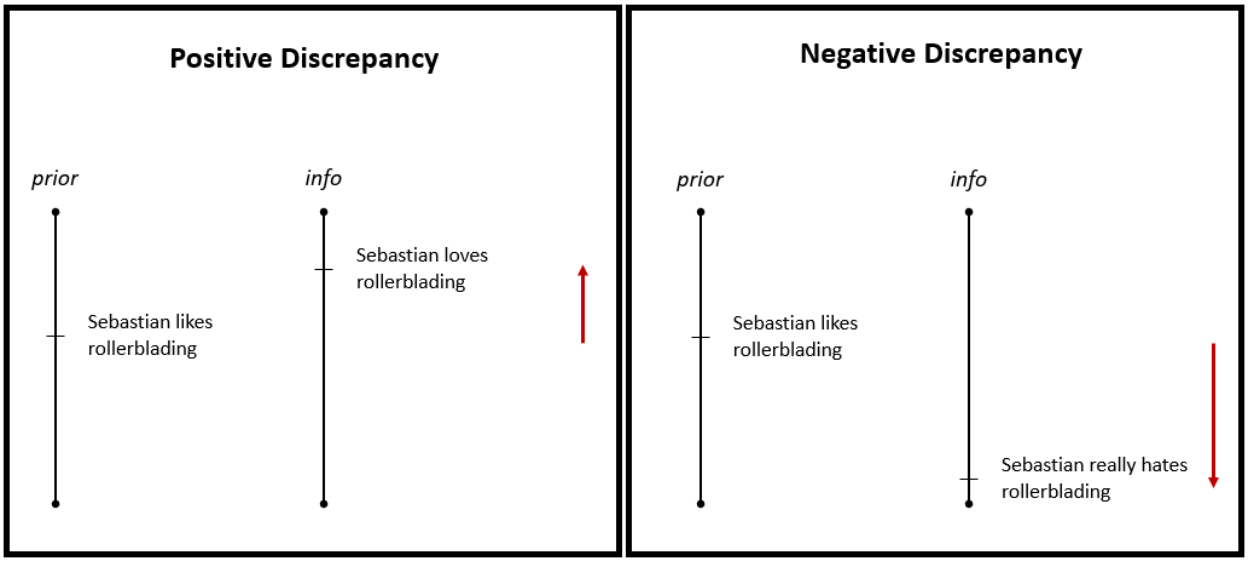
[26] i.e. for any given value on the x-axis (i.e. level of discrepancy), the y-axis value (i.e. believability) is always higher (or certainly not lower) for highly reliable sources (e.g. pink line) than for less reliable sources (e.g. green line)

[27] Imagine a horizontal line at any point along the y-axis in, for example, the "linear" panel. Such a line would intersect, for example, the pink (highly reliable) line and green (middlingly reliable) line at different x-axis values (levels of discrepancy). The two points on the x-axis where the two intersections would occur would indicate the level of discrepancy for which these claims by these two sources would be judged as equally believable. (The distance on the x-axis between where these intersections occur would indicate the amount of additional discrepancy that a more reliable source could "get away with" before their statements would be judged as equally believable as that of a less discrepant claim from a less reliable source.) If the highly reliable source makes a claim at any point on the x (discrepancy) axis beyond where this intersection occurred, that claim would be believed less than the original less discrepant claim from the less reliable source.



Part 1: Outlining Framework (Continuous Beliefs Models)

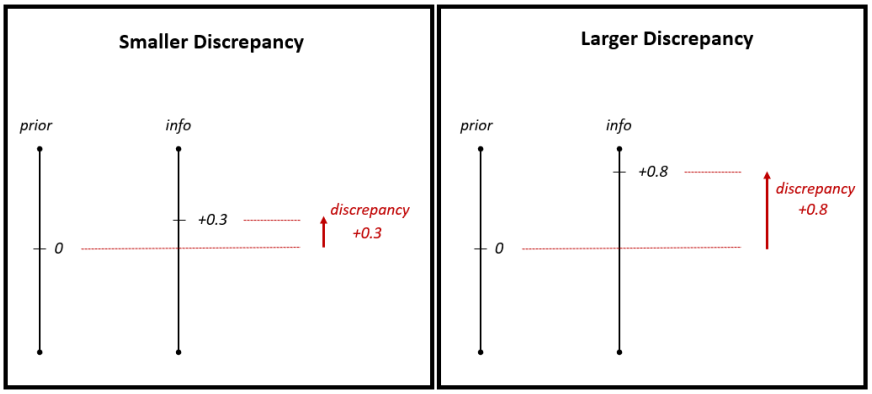
In some cases, we may also consider the fact that discrepancy may also vary in "direction." This is illustrated in the figure below, where whether the red arrow is pointing up or down indicates the direction of the discrepancy. This can be used to characterize domains where beliefs have some notion of opposing poles. The most obvious example of beliefs that could fall in this category are beliefs that have some sort of notion of valence or desirability. In such domains, certain pieces of new information are more desirable or positive, to some extent, than one's prior beliefs (e.g. I think I am more attractive than 60% of people, my "prior"; but I am confronted with new information indicating I am more attractive than 80% of people), and other pieces of information are less desirable or more negative, to some extent, than one's prior beliefs (e.g. I think I am more attractive than 60% of people, my "prior"; but I am confronted with new information indicating I am only more attractive than 40% of people). Positive signed discrepancies represent cases where the new piece of information is more positive or desirable than our prior beliefs, and negative signed discrepancies represents cases where the new piece of informaiton is more negative or undesirable than our prior beliefs.



For the purposes of our model, let's use *discrepancy* to quantify the extent to which a new piece of information is discrepant from one's prior belief (or beliefs), for both categorical and continuous beliefs . And specifically, in mathematical terms, let's take this to mean that *discrepancy* is a real number that varies from -1 to +1. Just for the sake of writing it out in formal notation, this is the same as saying:

- $discrepancy \in \mathbb{R}, where\ |discrepancy| \leq 1$

An example of two different sizes discrepancies is shown below.



And now let's think about belief change in the case of "continuous" beliefs. Belief change is some function of discrepancy and believability. There are many different ways this function might operate.

[note: belief change is not just attitude change. It can correspond to cases of what we would call "learning"]

In general terms, we said before:

- $belief\ change = f(discrepancy, believability)$

A very simple option is to specify:

- $f(discrepancy, believability) = discrepancy * believability$

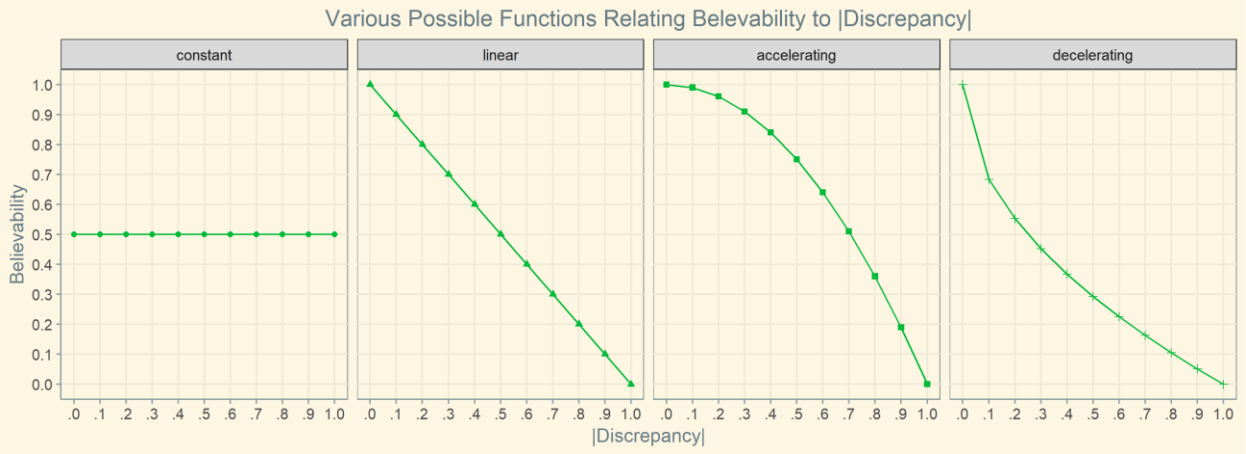
This can be thought almost in terms of the way we think of very basic expected value formulas (expected value = probability * value). Belief change here is a function of discrepancy (the amount by a which a piece of new information varies from our existing belief, which can be thought of as the value term in the basic expected value formula), multiplied by believability (which can be thought of as the probability that the new information is true). Most critically, there is a tradeoff here between discrepancy and believability in determining belief change. For most of our formulations of the believability function, as discrepancy increases believability decreases. Thus, in determining belief change, the forces of believability and discrepancy are acting in opposite directions. As discrepancy increases, believability decreases, and thus whether belief change increases or decreases is a function of which "force" "wins out"—raw discrepancy, or its effect on believability.

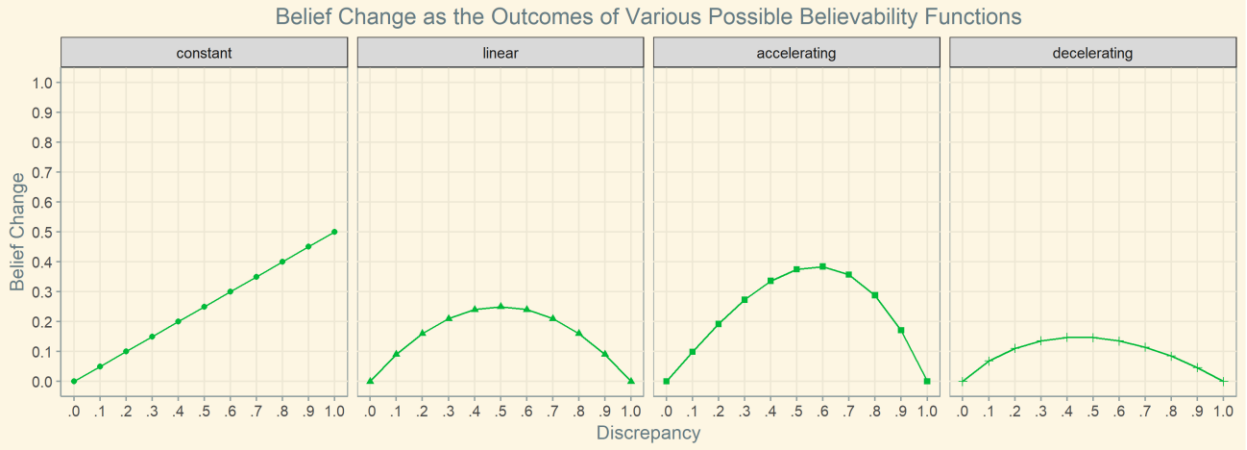
The graphs below capture this tension. In all cases, belief change is modeled as $belief change = discrepancy * believability$. The outcomes are shown as a function of variation in the different functions that may determine $believability$.

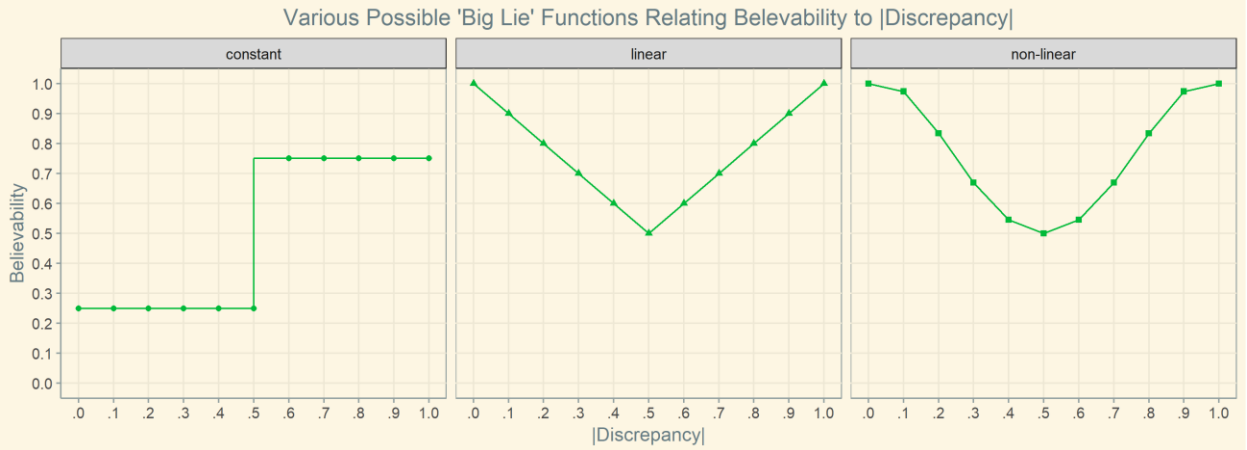
In the graph below, we only need to show discrepancy on the x-axis, even though $belief\ change = f(discrepancy, believability)$.

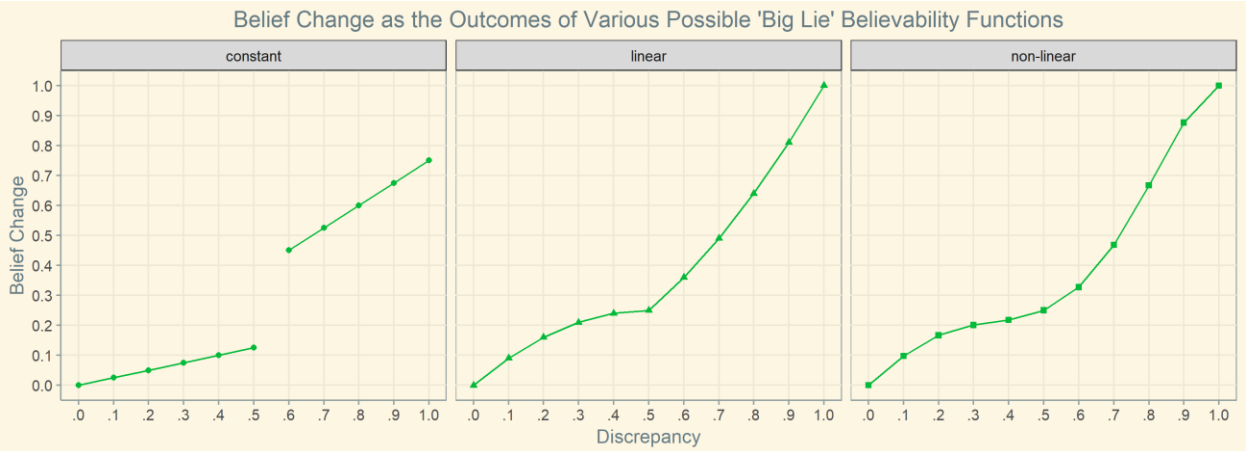
- This is because $believability = f_1(discrepancy)$, and thus:
  - $belief\ change = f(discrepancy, f_1(discrepancy))$, which is the same as:
  - $belief\ change = f_2(discrepancy)$
    - the graph below will show belief change outcomes in this form, as simply a function of discrepancy, for different believability, f_1, models

[didn't model negative belief change. But it's the same as reflecting the results around the y-axis. Imagine literally placing so that the thing part is running along the y-axis.]




Belief Change as the Outcomes of Various Possible Believability Functions
constant
linear
accelerating
decelerating
Belief Change
Discrepancy


Various Possible 'Big Lie' Functions Relating Belevability to |Discrepancy|
constant
linear
non-linear
Believability
|Discrepancy|


Belief Change as the Outcomes of Various Possible 'Big Lie' Believability Functions
constant
linear
non-linear
Belief Change
Discrepancy