HARVARD | BUSINESS | SCHOOL

G.I. Joe Phenomena: Understanding the Limits of Metacognitive Awareness on Debiasing

Ariella S. Kristal
Laurie R. Santos

Working Paper 21-084

G.I. Joe Phenomena: Understanding the Limits of Metacognitive Awareness on Debiasing

Ariella S. Kristal
Harvard Business School

Laurie R. Santos
Yale University

Working Paper 21-084

Funding for this research was provided in part by Harvard Business School.

G.I. Joe Phenomena:

Understanding the limits of metacognitive awareness on debiasing

Ariella S. Kristal[1]* and Laurie R. Santos[2]

*Corresponding author: akristal@hbs.edu

[1] Harvard Business School

[2] Yale University

Abstract

Knowing about one's biases does not always allow one to overcome those biases— a phenomenon referred to as the G. I. Joe fallacy. We explore why knowing about a bias doesn't necessarily change biased behavior. We argue that seemingly disparate *G. I. Joe phenomenon* biases fall into two main categories based on their cognitive architecture. Some biases are *encapsulated*— knowledge cannot affect a bias because the representations or emotions that give rise to the bias are informationally encapsulated. Encapsulated biases are the hardest to overcome since they are cognitively impenetrable by their very nature. In contrast, *attentional biases* are cognitively penetrable enough to be overcome through awareness, yet people still fall prey to such biases under conditions of distraction or limited attention during the moment of decision-making. We conclude by discussing the important implications these two discrete categories have for overcoming these biased and for debiasing efforts generally.

Keywords: biases; debiasing; illusions; judgment and decision-making; nudge

"It's not a case of: 'Read this book and you'll think differently.' I've written this book, and I don't think differently"- Daniel Kahneman on Thinking Fast and Slow (2011)

1. Introduction

Many of our most common self-regulatory failures have a shared feature— that is, we often fail to behave in a certain way even though we know the positive consequences of that behavior. Many of us know that we should limit our sugar and fat intake, yet we still stop at the donut shop too often. Many of us know we need to save for retirement, yet we fail to put enough money away each week. Many of us know we should exercise to reduce our disease risk, but we still hit the snooze button rather than heading for the gym. This disconnect between our knowledge and action does not just occur in the domain of self-regulation. There are examples throughout the field of cognitive science and decision-making where our behavior systematically deviates from rationality. These examples— which we refer to as a "biases" throughout this paper— include cases where people deviate from a standard that is objectively and normatively defined, as well as more subjective cases in which people deviate from a standard behavior they themselves would reflectively express wanting to follow (either by failing to take an action they would want their rational self to take or by performing an action that they reflectively regret).

The term "G. I. Joe fallacy" was coined in 2014 to deal with these sorts of cases— situations where our behaviors do not match up with our knowledge and beliefs. The G. I. Joe fallacy refers to the misguided notion that knowing about a bias is enough to overcome it (Santos & Gendler, 2014). The name of this fallacy derives from the 1980s television series *G. I. Joe*, which ended each cartoon episode with a public service announcement and closing tagline, "Now you know. And knowing is half the battle." Santos and Gendler (2014) argued that for many cognitive and social biases, knowing is *much less* than half of the battle. They further

argued that many people believe this is not the case, that knowledge is in fact enough to behave better, hence the G. I. Joe fallacy.

But the existence of the G.I. Joe fallacy raises an important question: why is knowing *not* half the battle? That is, why does consciously accessing information relevant to a situation not alter our biased behaviors? Why are there so many *G.I. Joe phenomena*— a term we will use in this paper to refer to empirical cases in cognitive science in which knowing is not enough to change the experience of a bias? And do all of these G.I. Joe phenomena work in the same way, via the same mechanistic processes? By understanding the different processes that give rise to a disconnect between reflective awareness and action, we hope to approach the problem of *debiasing*— changing people's behavior so that they behave more rationally— in ways that are more effective than simply learning about the existence of the phenomena themselves.

1.2 Overview of this paper

The goal of this paper is to review the puzzle of these G.I. Joe phenomena in more detail and explore why knowing is not always enough from a cognitive perspective. We argue that different G.I. Joe phenomena emerge for importantly different mechanistic reasons. In doing so, we develop a novel categorization of cognitive biases, arguing for the first time that there are at least two different kinds of cognitive biases from a mechanistic perspective. First, we discuss cases where informational encapsulation prevents knowledge from affecting our perceptions and emotional states— a set of biases we refer to as *encapsulated biases*. Encapsulated biases occur because our reflective awareness of our own biases is unable to cognitively penetrate the information processing that leads to our biased perceptions or behaviors. As we review below, biases can be encapsulated *representationally*— because reflective knowledge states can't penetrate the specific representations that give rise to a bias— or *emotionally*— because conscious reflection is unable to penetrate a strong emotional state, making it nearly impossible for us to successfully apply that knowledge. We next contrast encapsulated biases with a second class of errors, one that we refer to as *attentional biases*. As we explain below, attentional biases

involve cases in which knowing information could make us less likely to commit mistakes in theory, but in practice we often fail to attend to the relevant information in order to make better decisions in the moment. We then discuss the implications of these two novel categories for debiasing efforts, arguing that the mechanisms underlying these biases necessitate that one category of biases will be much harder to overcome than the other . The recommendations that stem from our bias categories channel Soll, Milkman and Payne (2015)'s "User's Guide to Debiasing," which builds off of Fischhoff (1981)'s classic approach to debiasing by suggesting that one can attempt to modify the decision-maker or modify the environment. We will show that for encapsulated biases modifying the decision-maker will be difficult at best, and futile at worst. We will end our review by explaining why a better approach will involve educating the decision-maker *to modify their environment* differently depending on which class of biases they are tackling.

The heart of the argument we propose in this paper is that the biases that plague human decision-making and are immune to conscious situational knowledge do not fall into a single category. Although scholars have attempted to classify biases into categories previously (Arkes, 1991; Fischhoff, 1981; Stanovich, 2009; Wilson & Brekke, 1994), the current review provides a novel conceptualization based on the mechanisms that drive these biases, arguing that seemingly different biases fall into two relatively distinct natural kinds based on their underlying cognitive architectures. This new framework is important as it allows for some important implications, both for how to understand the way these biases operate mechanistically as well as how to design effective interventions to overcome them.

2. Encapsulated biases

One of the key drivers of G.I. Joe phenomena is the *cognitively impenetrable* nature of certain biases: information from other cognitive domains simply cannot interact with the representations and emotional states that lead to our biased behaviors. To use Fodor's (1990) term, many biases are *informationally encapsulated* from conscious reflection. Encapsulated

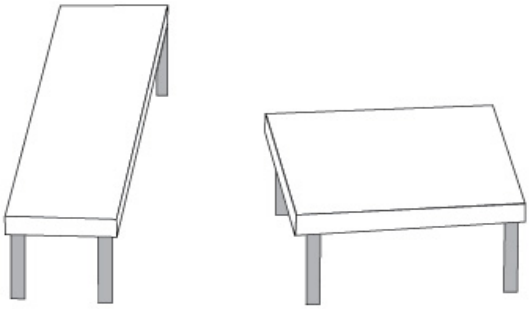
biases have two distinct flavors, which are based on the *content* that cannot be accessed by conscious reflection. Some biases occur because the cognitive *representations* that give rise to a bias can't be penetrated by knowledge from another domain. As we review below, a classic case of *representational encapsulated biases* are visual illusions— even in cases where we know how an illusion works, our knowledge cannot cognitively penetrate the representations that lead us to perceive that illusion (Fodor, 1990). Other biases occur because cognitive reflection is unable to penetrate the affective states that give rise to that bias. These *emotionally encapsulated biases* occur because the emotional states that cause biased behavior are cognitively impenetrable, at least during the heat of the moment.

2.1 Representationally encapsulated biases

We begin by discussing representationally encapsulated biases— cases where we behave in a biased way because the representations that cause our biases are cognitively impenetrable to information from other domains.

2.1.1 Visual illusions. Visual illusions epitomize the phenomena of informational encapsulation because learning how they work rarely affects the percepts we experience when viewing them. In fact, researchers in the field of judgment and decision-making have long argued that cognitive biases sometimes act like perceptual biases (e.g., see Kahneman, 2003; Kahneman & Tversky, 1984). Consider the famous example of Shepard's tabletop illusion (Shepard, 1990), shown below (Figure 1). People tend to see the left table as being longer despite the fact that the shape of the top of the table is exactly the same for both. This mistaken percept remains even when one learns about the size-constancy phenomenon that drives the illusion, or takes a ruler and directly measures the dimensions of the tabletops. Even when you clearly *know* that the tabletops are exactly the same shape, you cannot help but *see* them as different.

Figure 1. Shepard's Tabletop Illusion



The Shepard's tabletop illusion emerges because our brains automatically convert the two-dimensional figures before us into three-dimensional objects we would encounter in space (based on how the legs are positioned), and we therefore perceive the shapes as being different from each other. Knowledge of the nature of the illusion or one's own measurement of the tables is simply unable to penetrate our perception (for a discussion on the lack of evidence for top-down processing in perception more generally, see Firestone & Scholl, 2016[1]). As this example nicely illustrates, even when we consciously understand the factors that play into a visual illusion, we still fall prey to seeing that illusion in a biased way.

Below, we argue that a number of cognitive biases work in this same representationally encapsulated way— even when we understand what factors play into a given bias, our top-down knowledge of those factors simply cannot affect whether we experience the bias. In this paper, we refer to such biases as *representationally encapsulated biases*. As we explain below, the inability of top-down processes to interfere with automatic processes could have developed through deep evolutionary origins (such as in the case of visual illusions) or could instead stem from repeated associations (such the associations that build up via reading, e.g., the difficulty with the Stroop task), but regardless of how the content became encapsulated, the biases described in this section simply cannot be overridden by contradictory knowledge from a

[1] Note that there is still some controversy about how visual illusions come to be so encapsulated. Some scholars argue that visual illusions are built in innately (e.g., Fodor, 1990), while others argue they emerge as deeply ingrained associations that do vary across cultures (Henrich et al., 2010).

different domain. As we review below, many kinds of heuristics and cognitive biases fall into this category, meaning that knowledge of a given bias fails to impact behavior due to representational impenetrability.

2.1.2 Implicit Associations. Implicit associations refer to the unconscious ways previous exposures and associations influence our cognition (Greenwald & Banaji, 1995). Such associations stem from affective responses that we learn associatively (Blair, Ma, & Lenton, 2001). If we consistently see one class of stimuli associated with positive outcomes or events, we will be biased to think of these stimuli positively, whereas a different category that tends to be paired with negative outcomes will come to be experienced more negatively. In addition, if we consistently see one class of stimulus associated with another (e.g., men and scientific careers), we will be biased to link those two categories together. These associative processes are problematic because they tend to leak out into our judgments of and behaviors toward different social categories and therefore lead to biases that harm people in the real world. For example, implicit associations often play out in the context of implicit biases about gender and racial identities. In one example, since police chiefs have historically been male, we have come to associate police chiefs with maleness. In a lab setting, Uhlmann and Cohen (2005) provided participants two different candidates for police chief, one was street smart but without much higher education in criminology, while the other had been formally educated, but did not have much in-the-field experience. The researchers then varied which of the two candidates were male. Participants were more likely to select the male candidate for police chief in both conditions. In this way, participants change their explicit hiring criteria to match their implicit associations during hiring decisions (for similar effects in other identities and in real world contexts, see Bertrand and Mullainathan, 2003; Jackson, 2009; Oreopoulos, 2011; Pager, Western, and Bonikowski, 2009; Quandlin, 2018; Tilcsik, 2011; Weichselbaumer, 2003; Wright et al., 2013; for a comprehensive review of 90 audit studies between 2005-2016 see Baert, 2018).

In contrast to some of the other biases reviewed in this paper, there is ample evidence revealing that implicit associations are in fact a G.I. Joe phenomenon, one in which knowing does not change the nature of the bias. Paluck and Green (2009) reviewed hundreds of studies in both the lab and the field and found no rigorous evidence supporting the impact of diversity and unconscious bias training on implicit associations or the biases that emerge because of them. Kalev, Dobbin and Kelly (2006) review thirty years' worth of administrative data from over seven-hundred companies and find that companies that implement mandatory diversity training have fewer minorities and women in leadership positions than those that do not. A recent field experiment testing the causal impact of diversity training found modest positive effects on reported attitudes, but found no impact on any behavioral outcomes aside from an increase in mentoring women in one subset of the sample (Chang et al., 2019). Indeed, to our knowledge there is only one study showing that teaching people about their biases can help them overcome those biases, at least in the case of highly-motivated individuals; this study involved middle school teachers in Italy who took an Implicit Attitude Test (IAT), which measures implicit associations.[2] Half of the teachers received their results before submitting grades for their students, while the other half received their results after they submitted their grades; the performance gap between native Italian and immigrant children significantly reduced when their teachers submitted their grades after receiving feedback on their IAT scores (Alesina et al., 2018). But evidence of training working to overcome implicit bias seems to be the exception, rather than the general rule. Indeed, despite billions of dollars being spent, there is little proven effect of knowledge on reducing the magnitude of these implicit biases (for a review see Bohnet, 2016).

We argue that knowledge usually cannot help us to override our implicit associations and the biases that result from them because of the encapsulated nature of the representations

[2] Note that there is some controversy surrounding the IAT, with some researchers claiming that it has a low correlation with predicting explicit prejudice and real-world behaviors (Brauer et al., 2000; Greenwald et al., 2009).

underlying these associations. Implicit associations are picked up in our experience of the world over time. Constant exposure to certain kinds of statistical connections (e.g., there are more male mathematicians than female mathematicians) can create implicit associations even when we do not consciously endorse the associations we have learned (Arkes, Boehm, & Xu, 1991; Arkes & Tetlock, 2004; Dovidio & Gaertner, 1986; Wheeler & Petty, 2001). As with visual illusions, our associations seem to be immune to outside explicit knowledge even though they automatically shape our behavior. In this way, the mechanisms underlying implicit associations (and the implicit biases that result from them) suggest that they are likely to be representational biases much like the others we have discussed in this section.

It is of course possible to change our current associations by learning new ones. For instance, Blair, Ma and Lenton (2001) showed that counter-stereotypic mental imagery can weaken but not eliminate implicit bias (as predicted by the implicit association test, Greenwald & Banaji, 1995). But note that learning these new kinds of associations takes a long time and comes not from explicit knowledge but from repeated encounters with counter-stereotypic examples (such as male kindergarten teachers or female CEOs). Insofar as implicit associations themselves are not impacted by the knowledge of bias— or the explicit knowledge that such associations are problematic— but rather only through steady updating over time, they can be considered a classic case of a G.I. Joe phenomenon that results from representational encapsulation.

2.1.3 Halo effects. The order in which information is presented can impact one's global evaluation of a choice, despite there being other characteristics we could (and likely should) evaluate independently. The halo effect (Thorndike, 1920; Landy & Sigall, 1974)— in which subjects' first assessment of a product or person influences their subsequent judgments— is one example of this sort of bias. For example, when we see an attractive person enter the room to give a presentation, we may end up judging that person more favorably (e.g., more intelligent,

more competent, etc.) merely because of the halo of their attractiveness (Eagly et al., 1991). But does the halo effect go away when you know about it? Nisbett and Wilson (1977) debriefed participants in one of their studies demonstrating the halo effect and found that participants vehemently denied that their global evaluations influenced their ratings of specific attributes. Wetzel, Wilson, and Kort (1981) then tried explicitly forewarning and teaching participants about the halo effect to no avail. Based on these results, Wilson and Brekke (1994) identified the halo as a classic case of what they called "mental contamination," whereby even if someone was aware they were falling prey to this bias and was motivated to correct it, they would still be unable to do so. We claim that such contamination results from the representationally encapsulated nature of the halo effect. Just as in the case of implicit associations, the halo effect results from the fact that the positive representation we have of a person from one domain (e.g., their attractiveness) affects how subsequent information about that person is processed and represented. One's knowledge of how the halo effect works or that one might be influenced by this bias in the moment simply cannot penetrate the positivity representation once it's been activated. In this way, we argue that the halo effect is another case of a representationally encapsulated bias.

2.1.4. Debiasing representationally encapsulated biases. Recognizing that certain classes of biases are representationally encapsulated in nature has important implications for hypotheses concerning effective ways to get rid of these biases. If our knowledge of a particular bias is indeed representationally impenetrable, then debiasing efforts for these biases cannot rely merely on "raising awareness." Instead, such efforts should instead focus on changing how the problematic representations emerged in the first place. For example, to change our implicit gender associations and the representationally encapsulated biases that result from these, we need to change the very nature of how people represent different social

groups. Such changes require altering the daily associations people notice about different genders and job titles. Indeed, there is evidence the representations that underlie implicit attitudes can be contextually modulated and influenced by interventions that target associative learning over time (Dasgupta & Rivera, 2008; Rydell et al., 2006). For this reason, debiasing our representationally encapsulated biases might require widespread changes to the situations and associations we experience on a daily basis, through a route that is non-conscious, rather than a quick fix. The same types of changes are needed to debias emotionally encapsulated biases, which we discuss in the next section.

Although debiasing efforts may be unneeded or trivial for some representational biases (e.g., visual illusions), these efforts are more critical when it comes to combating biases that cause real world problems, such as implicit associations about different social categories. Frustratingly, billions of dollars are spent annually on training programs aimed at increasing people's knowledge of their implicit biases which have so far been almost completely ineffective (Bohnet, 2016). The realization that implicit biases result from associations that may be representational in nature suggests a more effective approach— organizations can adopt *situational* strategies to restrict the influence of unnecessary information about gender/race that activates these biases in the first place. Approaches like these seem to be far more effective in reducing the classic racial and gender biases that consistently affect hiring than trying to make people aware of their implicit biases. In one simple example of such situational changes, orchestras introduced a screen that "blinded" conductors to the gender of the musician who was auditioning. By preventing a gendered representation from being activated in evaluation context, researchers were able to increase the percentage of women in the top-five orchestras in the United States from 5% to 25% (Goldin & Rouse, 2000)[3].

[3] That said, it is important to be aware of how blinding can have unintended consequences. For example, laws banning employers from asking applicants about their criminal records has been shown to lead to a 5% decrease in probability of being employed for young Black men because the absence of information in one domain leads employers to statistically discriminate based on the information they do have (Doleac & Hansen, 2018).

In addition, understanding how representationally encapsulated biases work can also provide insight into the ways that knowledge of one's biases can prove helpful. For example, recognizing that you are in a situation in which representationally encapsulated biases are at play should prompt you to rely less on your own intuitive judgments and more on other kinds of measures. For example, if you were asked to make a wager over which of the two Shepard's tabletops was longer, instead of relying solely on your own perceptions, you could consciously choose to use a variety of external tools (e.g., a ruler) to supplement your own decision-making. By the same token, awareness can be used in cases of implicit bias to check for the statistical effects of biases before decisions are finalized. Indeed, this is likely how highly-motivated Italian teachers reduced the performance gap between immigrant and non-immigrant students in the one study to show the effect of knowledge on implicit bias (Alesina et al., 2018). Other situational interventions like increasing transparency (whereby one's decisions will be made public) and accountability (whereby one has to justify decisions to a higher and/or external authority) have also been shown to reduce the impact of implicit bias for exactly this reason (Castilla, 2015).

2.2 Emotional encapsulation

Representationally encapsulated biases occur because knowledge cannot penetrate the cognitive representations that give rise to a certain bias. But biases are not only caused by cognitive representations. Many classic biases occur because of a person's *affective state* at the moment of a decision. Take classic cases of self-regulatory failures— in these cases, the feelings motivating us during the moment of decision are so strong (e.g., the temptation to eat a cookie that is right in front of us when we are on a diet) that intellectually knowing about the "irrationality" of that motivator or having a meta-awareness of the feelings themselves simply cannot change how we experience the emotional pull in the moment. We refer to these biases as *emotionally encapsulated biases*— biases that occur because cognitive reflection is unable to penetrate the affective states that give rise to that bias. Just as visual illusions provided a

gateway into understanding representationally encapsulated biases, we believe that the phenomenon of hot/cold states is a good way to illustrate the nature of emotionally encapsulated biases.

2.2.1 Hot/Cold States. Researchers have long used the terms "hot" and "cold" states to describe times when people are in the "heat of the moment" and experiencing a "visceral reaction" versus thinking "rationally and coolly," respectively (Loewenstein, 1996; Mischel, 1974). Indeed, the notion that "cool" reason is needed to control hotter "passions" can be found hundreds of years earlier in the works of philosopher David Hume (1998/1751). A large body of work shows that these two states are in some rich sense *emotionally encapsulated*— it is very difficult for people in a cold state to appreciate how they will feel when they are in a hot state and vice versa. This bias has come to be known as the *hot/cold empathy gap*, a bias that leads people to mispredict the choices they will make when they are in the alternative state. For example, college men in an unaroused cool state mispredict their likelihood of engaging in risky sexual behavior when they are in a hot sexually aroused state (Ariely & Loewenstein, 2006); the knowledge you have about the risks of sexually transmitted diseases and pregnancy when you are in a cool state simply do not matter when you are making a decision to have sex in a hot state. Similarly, Read & van Leeuwen (1998) found that participants who are satiated mispredict the amount of junk food they will want when they are in a hot state of feeling hungry. Likewise, Christensen-Szalanski (1984) found that pregnant women and women who were at one month postpartum mispredicted the amount of anesthesia they would want when they were in the hot state of actually being in a painful labor. The hot/cold empathy gap also has consequences for real world behavior. Just as we mispredict what we will want in our own hot state when we are in a cold state, so too do we mispredict other people's needs when they are in a hot state. This can allow us to inadvertently cause pain to other people who are in a hotter state than we are. For example, Bernabei and colleagues (1998) argued that cancer patients experiencing chronic pain are regularly not given enough pain medication in part because nurses in a cold state do not

recognize how much pain relief is needed. Similarly, participants are less likely to say that painful cold exposure is an acceptable form of torture when they themselves are placed into a painful cold pressor task (Nordgren, McDonnell, & Loewenstein, 2011). Interestingly, this effect is short-lived; patients go back to their baseline preferences a mere ten minutes after their own experience of the pain.

Hot/cold state decisions represent a G.I. Joe phenomenon almost by definition; even if one appreciates in a cold state that their thinking and decision-making is altered when in a hot state, that knowledge alone does not change one's lived experience when actually in a hot state. As noted above, women who are only one-month post-partum change the values they expressed about using anesthesia in labor only one month before (Christensen-Szalanski, 1984). In addition, our tendency to think cold is unacceptable form of punishment changes a mere 10 minutes after feeling cold ourselves (Nordgren et al., 2011). In this way, when we switch from a hot painful state to a cold comfortable one, we simply no longer have emotional access to the preferences and decisions we had in the opposite state. Hot/cold state biases are impenetrable to our knowledge states because the only true way to cognitively access an emotional state is to be in that state at the time— knowledge of that state alone is not enough.

2.2.2 Situationism. A second case of emotionally encapsulated biases occurs when the emotions we experience in a particular situation affect our decisions. The term situationism has been used by philosophers to describe cases in which the context of and emotions generated by a given situation affect our decision-making outside of our conscious awareness (for a review, see Doris, 1998; Ross & Nisbett, 2011). In this section, we argue that the pull of such emotions can be so impenetrable that merely understanding that the emotion is influencing us may not be enough to prevent us from acting in a "biased" way.

In many classic situationist studies, people are biased to behave differently depending on the emotions they feel in a given situation. For example, classic early work in social psychology demonstrated that experiencing a pleasant situation or feeling— finding a dime in a phone

booth, randomly being given a tasty food, smelling freshly baked cookies, or listening to happy music at the gym— makes people more likely to volunteer to help others (Isen & Levin, 1972; Isen & Simmonds, 1978; Baron, 1997; North et al., 2004). In contrast, the feeling of being in a rush makes people less likely to help others (Darley & Batson, 1973) or recycle (Whillans & Dunn, 2015); indeed, the former effect is so strong that theological seminary students made to feel rushed failed to help an injured person on their way to deliver a sermon about the Good Samaritan (Darley & Batson, 1973). In another example, university admissions officers— whose sole job is to fairly evaluate candidates— change their decisions depending on the cloudiness on the day of their decision, with nicer weather that makes officers feel happier increasing probability of admission by an average of up to 11.9% (Simonsohn, 2007). Negative emotional states can also have an effect on our decision preferences (see review in Keltner & Lerner 2010). For example, Knapp and Clark (1991) found that individuals who were induced to feel sad performed sub-optimally in a resource management task because they optimized for immediate gratification as opposed to long-term success. Similar behavioral changes occur in situations that cause people to feel more aroused. In one classic study, men who had just crossed a rickety 250-feet-high suspension bridge were more likely to provide sexually vivid responses on a survey and call to ask out a female experimenter than men who crossed a 10-foot-high sturdy wooden bridge (Dutton & Aron, 1974). Taken together, these results show that we often behave differently — sometimes in biased ways— when induced to feel aroused or emotional (e.g., feeling sad about the weather, happy about finding a dime, etc.).

But do situations like these continue to affect our decisions even when we are aware of them? That is, is situationism a G.I. Joe phenomenon? Relatively few studies have tested this directly, but it is worth noting that the studies we reviewed above suggest that situationist effects can occur in supposed experts (e.g., admissions officers, Simonsohn, 2007; theological students, Darley & Batson, 1973) and thus may continue to occur even among knowledgeable individuals. Some studies have explored this issue more directly in the domain of emotional effects on

decision-making. While there is some work showing that awareness of emotional situations can mitigate their effects on decision-making (e.g., reminding people of the weather reduces ambient weather effects on judgments of subjective wellbeing, see Schwarz and Clore, 1983[4]), there are also counterexamples, cases in which knowing about the effect of emotions on decision-making does not affect the extent to which a person is affected by his emotions. For example, Han and colleagues (2012) found that teaching individuals about the carryover effect of disgust did not reduce the impact of incidental disgust on participant behavior, in this case, getting rid of their possessions. Indeed, Lerner and colleagues (2015) have argued that instances of successfully teaching individuals to overcome the impact of incidental emotions may be the exception rather than the rule. They and others point out how difficult it can be to become aware of situational effects on one's own decision process because such effects are often invisible to our conscious awareness (see Wilson & Brekke, 1994 for a review of situational effects as a form of what they refer to as "mental contamination"). For these reasons, we argue that many situationist effects, particularly those involving strong emotions and affective experiences, are likely to be G.I. Joe phenomena— the situation will continue to change our behaviors even when we know that a situation is liable to affect us.

But why do situational effects continue to occur even when we know about such effects? Again, we argue that emotional encapsulation may be at fault. Specifically, even if people explicitly know that a given situation could affect their behavior, this knowledge is encapsulated from the emotional states that are unconsciously affecting their behavior during the moment of decision. For example, knowing that a bad mood can affect hiring decisions is unlikely to change the fact that we often feel worse when the weather is bad; to the extent that this negative feeling automatically affects our decisions, knowing about the bias should not change its effect. By the same token, merely knowing that being on a scary bridge can make us more aroused is unlikely

[4] New research by Yap et al. (2017) fails to replicate the findings from Schwarz and Clore (1983) and they continuously fail to find evidence that mood affects judgments of subjective wellbeing altogether.

to change the extent to which higher arousal levels automatically increase our sexual attraction. In this way, we hypothesize that our knowledge of situationism is unable to penetrate the arousal and emotional states that affect our behaviors directly in these situations. Consequently, we argue that situation effects are yet another class of G.I. Joe phenomena that occurs for reasons of emotional encapsulation, though this is a domain in which more direct empirical tests are needed.

2.2.3 Present-bias, self-regulation, and willpower. The problem of willpower or self-control underlies many pressing societal issues such as poor health, low savings, low academic achievement, and beyond (see Duckworth, Milkman & Laibson, 2018 for a review). Psychology and behavioral economics have provided various models to account for these phenomena— such as "present bias," "time inconsistent preferences," and "quasi-hyperbolic discounting" (Ainslie & Haslam, 1992; Laibson, 1997; O'Donoghue & Rabin, 1999)— but essentially all of these models describe how we often come to want a smaller reward during the moment of decision at the expense of a larger reward later.

Self-regulatory failures fall in the domain of G.I. Joe phenomena because, like being in a hot state, explicitly knowing that self-control is challenging is insufficient in itself to successfully exert willpower. Take the case of dieting. Even if I know about present bias and recognize that my desire to eat the ice cream in the freezer is inconsistent with my long-term weight-loss goals, this knowledge is insufficient to halt the craving I experience in the moment for the ice cream. Indeed, some have argued that explicit knowledge of these cases can even have the opposite effect since we often use our explicit reasoning to rationalize succumbing to temptations in the moment (e.g., explicitly rationalizing that just one scoop of ice cream will not affect one's diet, see Wood, 2019).

But why do self-regulatory temptations continue to occur even when we know about them? We contend that emotional encapsulation may be at fault in the domain of self-regulation as well. Like in the case of situationism, even though people explicitly know that a

temptation could affect their behavior, this knowledge is encapsulated from the emotional or arousal state that emerges at the moment of decision-making when the temptation itself is present. Merely knowing that a cookie will look tempting when we have been dieting does not change the fact that cookies do in fact arouse strong cravings in the heat of the moment. Conscious reflection simply cannot penetrate the strong cravings we often feel at the time of a decision, and in the absence of such penetration, our cravings are likely to naturally lead us into self-regulatory failures. In this way, we hypothesize that our knowledge of self-regulatory failures is unable to penetrate the arousal and craving states that affect our behaviors directly in these situations. For this reason, self-regulatory failures are a further example of emotionally encapsulated biases.

2.2.4 Loss aversion. Loss aversion occurs in cases in which we put more work into avoiding losses than we do to seek out equally sized gains (Tversky & Kahneman, 1979). Much work has shown that humans tend to make different decisions when the same outcome is framed as a loss as opposed to a gain, with people changing their typical preferences (e.g., becoming more risk-seeking) in order to avoid what seems like a loss (see Tversky & Kahneman, 1981). In one famous example, participants must choose between a policy that can save 200 lives or a second policy where there is a one-third probability that 600 people will be saved. Most people prefer the first option in this case, yet their preferences switch when the same exact decision is framed as a choice between letting 400 people die, or a one-third probability that no one will die (Tversky & Kahneman, 1981). Frisch (1993) presents subjects with this framing example and several others, side-by-side, to see if participants are still motivated to treat the situations differently, and participants with this knowledge still experience the situations as different, supporting our view that loss aversion is resistant to explicit knowledge in the same way as other G. I. Joe phenomena. [5]

[5] We debated whether the bias of loss aversion was better described as an emotionally encapsulated bias (as we argued here) or more of a representational bias. Indeed, some work has argued that loss aversion has both representational and emotional components (see, Ariely, Huber, & Wertenbroch, 2005; Arkes, 1991; Kahenman &

Loss aversion can also play out in the real world. Businesses often inadvertently frame monetary transactions as either losses or gains in order to get customers to behave differently. For example, consumers behave differently when given "rebates"— which people frame as a return of money that was previously lost, — versus "bonuses", which seem instead like an unexpected gain (Epley, Mak & Idson, 2006). Similarly, customers are more motivated to avoid the loss of having to pay a "surcharge" for using a plastic bag than they are to seek out a "discount" for using a reusable bag. Loss aversion also affects choices in the stock market. People who frequently check their portfolio are more likely to notice occasional performance dips and due to loss aversion tend to sell their stocks at a low price, a phenomenon that Benartzi & Thaler (1995) christened myopic loss aversion.

Loss aversion occurs because losing causes a negative feeling, one that's powerful enough to make us automatically motivated to avoid losses. Like the other emotionally encapsulated biases discussed in this section, such negative feelings seem to automatically lead to a certain kind of behavior (typically trying to avoid losses even when doing so requires going against our previously established preferences) in ways that explicit knowledge usually cannot penetrate. In line with this idea, there are hints that expertise rarely reduces or eliminates loss aversion.[6] Indeed, more experienced professional stock traders exhibit myopic loss aversion to a greater extent than less experienced university students (Haigh & List, 2005). For these reasons, we argue that loss aversion may also be a G.I. Joe phenomenon that arises for emotionally encapsulated reasons.

2.2.5. Debiasing emotionally encapsulated biases. We have argued that knowledge is insufficient for eliminating biases caused by emotional encapsulation because the feelings or

---

Tversky, 1984). And even Frisch (1993) describes the multifaceted nature of framing effects, and likens one aspect, separate from the experienced situation, like visual illusions, further supporting the potential of a representational aspect. Nevertheless, given the emotional impenetrability of this bias, in the end we chose to include it in this section.

[6] However, there is one published case in which experts are less subject to loss aversion than amateurs: expert baseball card traders are less affected by loss aversion than their amateur counterparts (List, 2003). This exception, though, we feel proves the affective encapsulation rule. Amateur traders are thought to be more connected to their cards and thus wind up more motivated at the thought of losing them than professionals who tend to think of the profession only in terms of monetary exchanges.

cravings that automatically affect our choices in these cases are— in a rich sense— *emotionally impenetrable* during the moment of decision-making. One simply cannot know how much pain medication one will prefer when one is not currently in pain. Likewise, one simply cannot represent how valuable eating a delicious cupcake will seem when one is not in the presence of such a captivating stimulus. In this way, emotionally encapsulated biases are hard to overcome because there's no way for us to intervene on the influence of certain emotions or situations when we are directly under the influence of those emotions or situations. Unfortunately, due to the emotionally impenetrability of these biases, conscious awareness simply cannot change the emotional pull we feel (or fail to feel) in the moment. As such, just as with representationally encapsulated biases, we predict that policy interventions aimed at emotionally encapsulated biases will be doomed to failure.

The good news is that the nature of these emotionally encapsulated biases suggests a different and more effective way to help decision-makers overcome the power of their emotions and situations: find ways to help people avoid certain emotions or situations in the first place. If we know getting into a hot state will necessarily lead our good habits astray, we can alter our situations so we do not get into a hot state in the first place. Similarly, we can try to modify potentially tempting future situations before they arise so that we are less likely to succumb to temptation. An early example of just this technique comes from the famous story of Odysseus. On his way home after the Trojan War, Odysseus wanted to hear the song of the Sirens, whose singing was alleged to be so beautiful that men would follow it to their death. Odysseus prevented his crew from hearing the songs ahead of time by blocking their ears (i.e., preventing exposure to the tempting situation). And to prevent himself from being tempted to follow the songs, he tied himself to the mast of his ship (i.e., he physically prevented himself from being able to make a typical response to the tempting situation). Odysseus's self-binding is an example of what researchers call a *commitment device*, a formalized arrangement that people enter into ahead of time to restrict their future choice set in ways that allow them to achieve their goals in

the face of temptation. Bryan, Karlan, and Nelson (2010) reviewed the efficacy of commitment devices and found that these tools allow people to successfully overcome present bias in numerous situations— from increasing savings, to improving academic performance, to promoting healthy eating, to allowing people to quit smoking. From the perspective of our framework, commitment devices work in part because they prevent a decision-maker from either getting into a hot emotional state in the first place or from taking action when they are in that state.

In this way, *situation selection* and *situation modification*— actively preventing emotionally encapsulated biases from emerging in the first place— is likely the most effective way to "debias" these biases (Duckworth, Gendler, & Gross, 2016; Duckworth et al., 2016, also known as "exposure control," Gilbert, 1993). Although it is preferable to act "upstream" before the affective state that leads to the bias is activated in the first place (Kristal & Zlatev, forthcoming), there are ways of engaging in cognitive reappraisal that can effectively change the way the affective stimuli are encoded, and therefore experienced emotionally in the moment (Gross, 1988; Duckworth, Gendler, & Gross, 2014). Note, however, that reappraisal requires more cognitive effort and is less likely to succeed in the face of strong temptations/high intensity emotions (Duckworth, Gendler, & Gross, 2014; Sheppes & Gross, 2011).

2.3. Debiasing conclusions for both types of encapsulated biases

To summarize, representationally or emotionally encapsulated biases are G. I. Joe phenomena that emerge due to a particular constraint of cognitive architecture: knowledge of a particular bias is simply unable to penetrate either the representations or affective states that give rise to that bias. For this reason, both kinds of encapsulated biases will be difficult to overcome through awareness or cognitive reflection alone. For both representational and emotionally encapsulated biases, debiasing efforts should instead focus on preventing the problematic representations or emotions from being activated in the first place. For representationally encapsulated biases, this can involve altering the availability of certain

representations by changing the associations that get learned over time or by preventing a decision-situation from activating the representations at the moment of the decision. For emotionally encapsulated biases, one can use similar situation selection techniques to prevent affective states from arising or by preventing decision-makers from acting when that affective state is active. For both types of encapsulated biases, effective debiasing requires not necessarily teaching people about these biases but helping them to devise situations that prevent the experience of certain representations or emotion in the first place. For example, people who know they will be unable to resist the emotional pull of overeating from a tub of ice cream may choose not to buy ice cream to keep in their home, or they may choose to buy multiple smaller individual portion sizes instead of a one-gallon tub (even if it means paying a premium for the equivalent amount of ice cream, see Wertenbroch, 1998). As these examples illustrate, the best way to overcome encapsulated biases is not through knowledge per se but by using knowledge ahead of time to structure environments that restrict the amount of damage we would otherwise inflict on ourselves when faced with strong representations or emotions.

3. Attentional nature of biases

Although the previous sections discussed cognitive biases which result in G.I. Joe phenomena for reasons of encapsulation, not all decision biases result because of cognitive impenetrability. Our final section addresses certain decision-making biases that are cognitively penetrable enough to be overcome through awareness. Nevertheless, people still fall prey to such biases under conditions of distraction or limited attention. We refer to this final set of biases as *attentional biases,*[7] ones that could in theory be overcome through one's own knowledge, but in practice often are not given the nature of the decision-making situation. As we review below, participants tend to show attentional biases because certain aspects of a situation

[7] Note that we do not use the term "attentional" in the way that vision scientists typically use the term, but instead to mean having the needed mental capacity or computational bandwidth to track certain aspects of the problem space (Pashler, 1998). Attentional biases emerge when the computational demands of the problem at hand exceed the needed attentional resources to accurately account for all necessary relevant information.

naturally capture attention or because it can be difficult to attend to the relevant knowledge during the moment of decision-making.

3.1 Out-of-sight-out-of-mind and in-sight-top-of-mind biases

There are a number of cognitive biases that stem from paying too much attention to salient information that confirms what we were already thinking or from not paying enough attention to highly relevant but less visible information. Kahneman (2011) referred to such biases with the acronym WYSIATI— *what you see is all there is*. As we explain below, WYSIATI biases occur because people either pay attention to the wrong information or fail to attend to relevant information. In this way, WYSIATI biases tend to be G.I. Joe phenomena— these biases are usually unaffected by outside knowledge because even when people are aware of the bias, they often fail to attend to the relevant information at the time of their decision.

3.1.1. Availability heuristics. Consider, for example, the availability heuristic, in which people judge events as more probable based on the ease with which they come to mind (Tversky & Kahneman, 1973; Schwarz et al., 1991). In a famous example, Tversky & Kahneman (1973) asked subjects whether there were more English words that start with the letter K or that have K as the third letter. People mistakenly thought that more words started with the letter K since these words were more available when people thought of potential examples. The availability heuristic also explains why people mistakenly think that high-salience events (e.g., a plane crash as a way to die) are more probable than less-salient but more common events (e.g., heart disease). Consequently, people tend to fear plane crashes and not heart disease, even though dying of heart disease is much more common.

But is the availability heuristic an example of a G.I. Joe phenomenon, that is, can knowledge of the bias override it? In contrast to the representational biases we reviewed earlier, the availability bias does not stem from problems of cognitive penetrability. For example, if people are directly given the statistics on the dangers of heart disease during the study, they will readily update their intuitions. The problem is that people tend not to do the cognitive work

needed in the moment to find these less-than-salient statistics. Instead, they merely use the information that is attentionally salient at the time (e.g., vivid examples of plane crash). In this way, knowing the real statistics or putting in the relevant search time could *in theory* reduce our availability bias, but it often fails to do so because of the attentional salience of the examples we commonly bring to mind. In this way, the availability bias, like other WYSIATI biases, is a G.I. phenomenon for attentional reasons— people simply do not tend to do the cognitive work needed to access the right examples.

3.1.2. Confirmation bias. Confirmation bias arises when we seek out evidence confirming our pre-existing beliefs and ignore or discount disconfirmatory evidence (for a review, see Nickerson, 1998). Again, this bias seems to fall under Kahneman (2011)'s definition of a WYSIATI bias in that people tend to simply not look for evidence that discredits their own hypotheses.

There is a mixed literature regarding whether learning about confirmation bias can reduce it (for a review see Lillienfeld, Ammirati, & Landfield, 2009). In one classic real-world example, expert NASA engineers fell prey to confirmation bias when deciding to launch the *Challenger* (for a discussion of this situation see Bazerman & Moore, 2008). Other researchers have tried to more directly test whether knowing about confirmation biases can help us overcome them. Morewedge and colleagues (2015) developed a video game aimed at helping participants to overcome confirmation as well as several other biases (see discussion in section 3.5). Their video game not only successfully taught participants about confirmation bias, but also provided new contexts in which participants could experience the bias in action and receive immediate feedback to learn from previous mistakes. Morewedge and colleagues found that this training allowed subjects to successfully reduce their confirmation biases over time (Morewedge et al., 2015) and to reduce confirmation bias in a task modelled after the *Challenger* example (Sellier, Scopelliti, & Morewedge, 2019). We suspect that the strategies employed in this video game task work in part because they target the attentional nature of confirmation bias.

Specifically, this task succeeds because it provides subjects with cognitively efficient ways of attending to and thinking about evidence, thereby reducing the likelihood that participants fall prey to attentional demands that are typical in situations that tend to lead to confirmation bias.

3.1.3. Base rate neglect. A similar phenomenon occurs in the context of base rate neglect, a bias in which participants who are asked to think about a given phenomenon pay too much attention to salient representative information without accounting for the prior probabilities of that phenomenon occurring (Tversky & Kahneman, 1974). One example of this involves predicting the probability that a forty-year old woman has breast cancer if a mammogram comes back positive, considering that the probability of a woman of her age having breast cancer is 1%, the probability of getting a true positive if she does have cancer is 80%, and the probability of getting a false positive is 9.6% (Gigerenzer & Hoffrage, 1995, adapted from Eddy, 1982). The correct answer, according to Bayes theorem, is 7.8%, but most people estimate the probability of her having cancer to be between 70-80% (Gigerenzer & Hoffrage, 1995). Like the availability heuristic, base rate neglect continues to occur in experts, such as professional researchers (see Tversky & Kahneman, 1971). However, there is research suggesting that base-rate neglect is lessened when information is presented in a "computationally simpler" way: using frequency instead of probabilistic data (see Gigerenzer and Hoffrage, 1995). When probability information is presented in terms of frequencies, decision-makers are more easily able to direct their attention to the relevant information. As a result, Gigerenzer and Hoffrage (1995) show that although only 10% of physicians give the correct answer to a base-rate neglect problem when presented with probability information, 46% arrive at the correct answer when they are presented with the information in the form of natural frequencies. A recent review by Barbey and Sloman (2007) argued that this reduction in base-rate neglect stems from the fact that the researchers inadvertently made the problems easier from an attentional standpoint; they contend that presenting problems in ways that more

automatically reveal the underlying structure reduces the cognitive capacity needed to arrive at the correct solution, thereby again supporting our attentional model of base-rate neglect.

3.1.4. Planning fallacy. Another WYSIATI bias, the planning fallacy, occurs when people systematically underestimate the time, costs and risks associated with completing a task (Kahneman & Tversky, 1973; Lovallo & Kahneman, 2003). Participants, for example, underestimate how long it would take to finish holiday shopping, to get ready for a date, or to format a document in an experiment. There is also suggestive evidence that the planning fallacy is a G.I. Joe phenomenon in that this bias continues to occur even amongst those aware of the planning fallacy. Indeed, the current authors admit to having consistently underestimated how long it would take to produce drafts of this paper despite writing a whole section of this paper on the planning fallacy itself.

Nevertheless, research also shows that people can begin to make more accurate predictions in planning situations when they are first asked to break down each of those activities into specific subcomponents (Kruger & Evans, 2004). This intervention makes the previously invisible but relevant information more available to participants, thereby reducing the attentional demands of a typical planning fallacy task. Again, this result suggests that the G.I. Joe nature of the planning fallacy results from attentional problems that can be overcome by making other information (e.g., the specific steps involved in the process) more attentionally salient.

3.1.5. Hindsight bias— Hindsight bias is a form of after-the-fact thinking in which people assume they would have predicted an already-experienced outcome more than they actually would have before knowing about that outcome (Fischhoff, 1975). Much of the early work on the hindsight bias demonstrated that experts also show this error, suggesting that knowing about the existence of this bias doesn't reduce it.

Like the other biases discussed in this section, hindsight bias occurs because of an attentional failure; events that have already occurred are much more attentionally salient and

thus more easy to imagine. It's worth noting that Arkes (1991) had this insight about hindsight (and a number of related attentional biases) several decades ago. He referred to such biases as "association-based biases," and made the strong claim that people will not naturally search for information that they're not attending to because it's hard to think about what we're not currently thinking about. As Arkes (1991) put it: "It would be difficult for subjects to abort a cognitive process that occurs outside of their awareness. 'Please prevent associated items from influencing your thinking' would be a curious entreaty unlikely to accomplish much de-biasing."

It follows, then, that making counterfactual information more salient should help reduce our susceptibility to this bias, which several studies have demonstrated (see Arkes et al., 1988; Koriat, Lichtenstein, & Fischhoff, 1980; Hoch, 1985). For example, there is evidence that "consider the opposite" strategies can reduce hindsight bias in part because they shift people's attention to neglected but decision-relevant factors (Lord, Lepper & Preston 1984). However, it's worth noting that attentional fixes like these are tough to employ in real world circumstances. For example, Sanna, Schwarz, and Stocker (2002) showed that when individuals try to list too many alternative outcomes, they fail to attend to any of them which causes the intervention to backfire.

3.2 Relative choice biases

Like other WYSIATI biases, relative choice biases occur when our attention is focused on one specific aspect of a decision, which drives us to make a decision relative to that specific aspect of the decision problem. As we review below, many cognitive biases stem from people attending too much to a specific salient reference point from which they make relative decisions. While it is possible to overcome these relative choice biases with cognitive effort, we tend not to use such effort in most real-world decisions. In this way, merely knowing that relative choice biases are possible is rarely enough in practice to overcome our attentional laziness in the moment.

3.2.1. Status quo bias. A prime example of relative choices biases involves the status quo bias— a tendency to default to whatever the status quo or current defaults suggest (Samuelson & Zeckhauser, 1988). There is much experimental and real world evidence suggesting that defaults can exert powerful effects on behaviors; defaults can help people save more for retirement (Madrian & Shea, 2001; Choi, Laibson, Madrian, & Metrick, 2003; Thaler & Benartzi, 2004), choose more environmentally-friendly options (Larrick & Soll, 2008; Sunstein & Reisch, 2014; Egebark & Ekström, 2016), and register to become organ donors (Johnson & Goldstein, 2003). Indeed, a recent meta-analysis of the impact of defaults across policy domains reveals an average effect size of $d = 0.63\text{-}0.68$ (Jachimowicz et al., 2019). Like the WYSIATI biases discussed in Section 3.1, people can in theory override these defaults, but doing so can be attentionally costly, which means that we tend to just use whatever the default is when making decisions in the real world. For example, research has shown that disclosing the presence of a default does not reduce the effectiveness of a default; even when people are told that they are randomly assigned to a default, they fail to take the cognitive effort to generate a new default (Loewenstein et al., 2015). In this way, status quo biases are another case of a G.I. Joe phenomenon due to attention biases; people rarely put the attentional effort in to overcome the status quo even when they know they could be susceptible to this bias.

3.2.2. Anchoring. Another signature case of decisions made relative to an arbitrary reference point is the case of anchoring.[8] When asked to make an estimate, people are initially influenced by extraneous anchors and then insufficiently adjust to the correct answer. For example, Tversky and Kahneman (1974) asked participants to estimate the percentage of African countries in the United Nations. Before making the estimate, a wheel was spun and landed on a random value between 0 and 100. Participants first stated whether they believed the true value

[8] Note that other relative choice biases are thought to be due to anchoring. For example, some have argued that the curse of knowledge— a bias in which we assume others know more because we ourselves know something— is an instance of anchoring with insufficient adjustment, where our knowledge is the anchor and the errors arise from insufficient adjustment (Nickerson, 1999).

was above or below that anchor and then had to estimate the true answer. When people were anchored on the number ten, their median estimate was 25% but when they were anchored on the number sixty-five, their median estimate was 45%. Note that anchoring only operates when people have at least some uncertainty about the correct answer and have to search for a correct answer. If a person knows a specific fact, they are not susceptible to this bias. But despite having more general knowledge about a topic than laypeople, experts in a domain are still susceptible to anchoring errors (e.g., real estate agents: Northcraft & Neale, 1987; judges: Englich, Mussweiler, & Strack, 2006). While we are unaware of research directly testing whether telling people about anchoring makes it go away, we interpret the evidence on experts as indicative of the pervasiveness of anchoring even in the face of explicit knowledge about a domain. Further evidence that anchoring is related to limited cognitive capacity is provided by Epley & Gilovich (2006), who demonstrated that people tested in anchoring experiments continue searching for a more accurate estimate only if they are able to do so easily, such as when they are under low levels of cognitive load. Again, these results hint that anchoring may result due to attentional failures during the moment of decision-making, and thus may be a G.I. Joe phenomenon for attentional reasons.

3.3 Affective forecasting

Because of the nature of what we pay attention to (or fail to pay attention to), we tend to inaccurately predict how we are going to feel in the future; we are bad at what is known as *affective forecasting*, the emotional predictions we make about our future affective states (for a review see Wilson & Gilbert, 2003). In one classic study, professors predicted that they would be extremely happy about receiving tenure and devastated if they were denied, but when those outcomes actually came to pass, participants were much more measured in their responses (Gilbert et al., 1998). There are also hints that knowing about affective errors does not immediately make us better at avoiding them; people who fail the same exam multiple times in a row still mispredict how sad they will feel after failing the exam (Ayton et al., 2007), suggesting

that knowledge of one's previous affective forecasting errors does not immediately allow one to overcome this bias.

There are many hints that affective forecasting errors tend to occur due to attentional failures. First, people fail to attend to the fact that they habituate to good and bad events when making affective predictions (for a review see Frederick & Loewenstein, 1999). Second, people attend to and overweight very salient aspects of a potential future event, what researchers have referred to as *focalism* (Wilson et al., 2000) or *focusing illusion* (Schkade & Kahneman, 1998). Attending too much to one aspect of a future event means we inadvertently neglect other less salient aspects of the future event which can lead to poor decision-making. Schkade & Kahneman (1998) first demonstrated the focusing illusion by showing that students in the Midwest and in Southern California reported similar levels of life satisfaction, even though both groups expected Californians to be happier, because they focused too much on the highly salient factor of weather, which is not all that diagnostic of people's happiness. A third attentional culprit in our affective forecasting errors— particularly when it comes to negative events— is our failure to attend to our psychological immune system (Gilbert et al., 1998). When bad things happen, we tend to employ a host of psychological strategies to feel better. Immune neglect— our failure to attend to these feel-good psychological strategies— leads us to affective forecasts in which we overestimate the amount of time we'll feel badly.

Like other G.I. Joe phenomena, we hypothesize that knowing about our tendency to make affective forecasting errors is not enough to overcome this bias mostly because it is hard to allocate our attention to the right factors during the moment of prediction. Nevertheless, debiasing efforts that help participants attend to the right information while making forecasts do seem to improve performance. For example, Wilson and colleagues (2000) showed that participants are better at affective forecasting when they redirect their attention to other non-salient parts of their lives. In one study, college football fans predicted how the outcome of a game would affect their happiness. Fans tended to make more accurate predictions when they

were first directed to attend to how they would spend their time the day after the game, compared with a control group that was simply asked how they would feel the day after their team won or lost. Critically, this study shows that people must put in work to direct their attention beyond a focal event in order to reduce affective forecasting errors. Because people tend not to do so on their own, merely knowing about affective forecasting is not always enough to override this bias for attentional reasons.

3.4 Intention attribution

The fundamental attribution error is a bias in which people interpret the behaviors of others using inferences about their internal dispositional traits while neglecting the role of situational constraints. A classic method for demonstrating the fundamental attribution error involves having people write an essay in favor or against a specific opinion. New participants are then informed about the writers' instructions, asked to read their essays, and are then asked to what extent the writers agreed with the positions they expressed in the essay (see Jones and Harris, 1967). Despite being aware that the essay writers were assigned to their position, participants attributed a high level of correspondence between what was written and the writer's true beliefs. Unfortunately, the fundamental attribution error is not merely an intellectual problem, but also has real world consequences. For example, Moore and colleagues (2010) showed that this bias can affect university admissions and job application processes, favoring students from institutions with grade inflation over successful students from institutions that did not engage in grade inflation. In this way, decision-makers do not sufficiently discount high grades that are due to lenient institutional grading practices.

Given the real-world consequences of this bias, many have hoped that telling people about the fundamental attribution error could help participants to overcome it. Unfortunately, like the other biases reviewed in this section, merely knowing about this bias does not seem enough to reduce its effect due to a specific kind of attentional failure: given the myriad demands on our attention, we often neglect to attend to how other individuals are influenced by their situation.

In this way, the fundamental attribution error is yet another G.I. Joe phenomenon that stems from attentional problems. Indeed, when people devote more attention to the inference task, they are often able to overcome this bias. For example, Tetlock (1985) demonstrated that people are less likely to fall prey to the fundamental attribution error when they know ahead of time that they will have to justify their impressions. This result suggests that getting people to focus their attention in new ways makes them less likely to fall prey to the fundamental attribution error; however, it is rare that we make such a concerted effort and expend the attentional capacity to do so.

3.5 Debiasing attentional biases

Understanding that some cognitive biases stem specifically from attentional rather than encapsulation constraints provides some hints about how we can help people to prevent these biases from negatively impacting outcomes. In contrast to the encapsulated biases we reviewed earlier— where knowledge can never help due to problems of cognitive penetrability— attentional biases are in fact accessible to information from other domains and thus *can* respond effectively to knowledge or conscious reflection. The problem though, is that it is challenging for people to harness that reflection and attend to the right domain of information during their moment of decision.

As we've reviewed in the sections above, there is evidence that participants can reduce or eliminate their attentional biases in situations that help them attend more easily to the relevant information. For example, Wilson and colleagues (2000) found that participants show less of a robust focalism bias when they are asked to think about daily activities that are irrelevant to a salient positive or negative event. Similarly, Kruger & Evans, (2004) observed that people show less of a planning fallacy if you help them attend to the specific subcomponents of events that matter for how long an action will take. In addition, Epley & Gilovich (2006) found that subjects more readily adjust from salient anchors if you make the adjustment process attentionally easier, such as by reducing cognitive load. In all of these cases, attentional biases are reduced

when the attentional demands of the task are reduced— when relevant information is made more salient or easily available.

There are other hints that attentional biases can be reduced through repeated presentations and feedback. Morewedge and colleagues (2015) have developed an elegant video game task that aims to debias people's performance on a number of the biases we discuss in this section: confirmation bias, fundamental attribution error, anchoring, and representativeness. In these games, participants not only learn about these biases, but they are also given a set of mitigation strategies that— we hypothesize— enable participants to counteract their attentional biases in cognitively efficient ways. Although there is less robust evidence to date that these training techniques work in real world situations (although see Sellier, Scopelliti, & Morewedge, 2019) or that the effects hold long-term, we argue that such techniques could hold promise for reducing attentional biases to the extent that the trainings make the relevant information easier to access by helping subjects repeat effective strategies often enough that seeking relevant information becomes cognitively less demanding. That said, we are also less optimistic about the extent to which training like this will always lead to generalizable and persistent success in avoiding these biases, particularly in real world contexts when attention is naturally divided (Arkes, 1991; Fischoff, 1981; Soll, Milkman, and Payne, 2014). Instead we argue that a better strategy for reducing attentional biases will involve "outsourcing" the necessary inputs to decision-making, whether through deploying decision-making aids, relying on reminders/benchmarks at key stages in decision-making, or by building in accountability mechanisms that prompt people to focus on the non-salient aspects of a decision.

4. Conclusion

In this paper, we have explored why cognitive biases are G.I. Joe phenomena— why awareness of a bias does not make individuals any less susceptible to that bias (see Table 1 below for a summary). We have argued in this review that cognitive biases result in G.I. Joe phenomena for two distinct reasons, each of which tends to shield the bias in question from

rational thought processes. Some biases are encapsulated in nature— the representations or emotions that give rise to these biases are informationally encapsulated from cognitive awareness. In contrast, we argue that other biases are attentional in nature. Knowledge could in theory improve our performance on attentional biases, but does not do so in practice because we often fail to attend to the information we need in order to make better decisions in the moment. Recognizing the differences in these two kinds of biases and understanding their underlying mechanisms, we argue, has important implications for best practices for debiasing different kinds of errors.

Before we return to the implications of the categories we reviewed above for debiasing efforts, we first wanted to review how the categories we have identified here link to previous attempts to classify biases and the ways that our approach differs from previous advances in the literature. The computational and processing demands that lead to what we refer to as attentional biases have received the most discussion in previous reviews. Indeed, similar issues have been discussed by Fischoff (1981), Arkes (1991), Shah and Oppenheimer (2008) in their effort reduction framework for heuristics, Stanovich (2009) in his categorizations of cognitive misers and mindware gaps, as well as in Wilson and Brekke's (1994) cases that involve failures of rule knowledge. However, few of these previous reviews have explored *why* some biases are resistant even when participants have more processing power and are behaving less cognitively miserly. Put differently, few of these previous reviews have made the distinction between biases that could be penetrable with the right effort on the part of the decision-maker and those that are simply impenetrable to outside information.

The errors we have referred to as encapsulated biases are ones that tend to engender the phenomena that Wilson and Brekke's (1994) christened "mental contamination"— encapsulated biases result from perceptions and affective responses that are by their very nature resistant to explicit knowledge. Wilson and Brekke's mental contamination concept was one of the first attempts to articulate that some biases are likely to be structurally resistant to explicit goals and

knowledge, but they stopped short of the stronger claim that we make here— such contamination occurs because these representations and affective states underlying these biases are informationally encapsulated from other information that could lead to more rational judgments. Such biases are cognitively impenetrable and immune to debiasing in a way that attentional biases are not.

We argue that the novel distinction we've made between these two mechanistically-different classes of G. I. Joe phenomena has meaningful implications for research efforts to prevent these errors. The first implication— one that applies to both kinds of biases— is that in most cases raising awareness simply is not going to help. Much of the debiasing literature fits into the dual system perspective of an automatic, fast, intuitive system of thinking (System 1, where most of the biases take place) and a slower, more reflective system (System 2, which has the power to correct the biases of System 2) (Kahneman, 2011; Kahneman & Tversky, 1982; Sloman, 1996; Stanovich 1999). In fact, there is a school of thought that focuses on leveraging System 2 to overcome the shortcomings and correct the errors of System 1 (Kahneman & Frederick, 2002; Milkman, Chugh & Bazerman, 2009). As we have demonstrated above, though most of our biases may originate in System 1, shifting to System 2 is an insufficient solution. Indeed the very nature of the G.I. Joe phenomena we describe are such that one cannot simply reason one's way out of these biases.

Our analysis also suggests that a more effective approach to debiasing will involve recognizing that different categories of biases require different types of solutions. Attentional biases are likely to be the easier to debias since they result not from encapsulated emotions or representations but instead from a lack of the right kind of cognitive effort during the moment of decision-making. Under this view, there is a case to be made for educational approaches to debiasing attentional biases (Stanovich, 2009)— researchers should be able to observe improved performance through interventions that make information needed for successful decision-making more salient at the moment of decision-making or by giving decision-makers practice

cognitively reflecting during the moment of choice. Promising work using techniques like this has come from Morewedge and colleagues (Morewedge et al. 2015; Sellier et al. 2019; Yoon et al. 2020); as reviewed above, they have used video games that allow subjects to practice appropriate solutions to successfully debias a number of different attentional biases such as anchoring, the availability heuristic, and base rate neglect.

Unfortunately, we do not expect these types of approaches to show the same level of success for eliminating encapsulated biases. Because encapsulated biases result from representations or affective states that are informationally impenetrable during the moment of decision, they— by their very nature— cannot be overcome through practice, additional cognitive resource, or conscious reflection. These biases simply cannot be improved in the same way as attentional biases— by exerting more effort at the moment of decision. For both representationally and emotionally encapsulated biases, debiasing efforts should instead focus on preventing the problematic representations or emotions from being activated in the first place. We thus hypothesize that successful debiasing efforts will look very different for attentional versus encapsulated biases. The latter will be improved not through increased cognitive effort at the time of decision but by using that knowledge or effort ahead of time to structure environments to restrict the amount of damage we would otherwise inflict on ourselves when faced with impenetrable representations or emotions.

It is worth recognizing a few caveats to our argument before we conclude. The first is that there are several places in our review where there is a need for more empirical work. For some of the biases we have covered here, there is clear evidence demonstrating that the bias in question is a G.I. Joe phenomenon— that knowledge does not improve performance—, while for others, we have only an anecdotal sense that knowing about a bias or expertise in a particular domain does not make it go away. To be sure that all the biases we cover here truly represent G.I. Joe phenomena, future work should specifically train subjects about the nature of a given bias and then test whether performance improves. A second limitation of the current review is that we

restricted our focus only to the most prominent biases, ones for which there was usually some evidence concerning the role that knowledge plays. While we suspect that other cognitive biases also fit into the framework we have developed here, it would be worth reviewing other biases in detail to determine whether they do indeed fall into the natural kinds we have described.

Our insight that cognitive biases may naturally fall into one of two natural kinds provides a testable framework for exploring what types of debiasing attempts will likely be futile versus promising. We have argued that efforts that attempt to debias all biases in the same way are likely to fail. Instead, policy makers interested in debiasing should remember the G.I. Joe caveat we began this paper with— knowing isn't half the battle. But we argue that researchers need to do more than understand the limits of knowledge on debiasing— policy makers also need to pay careful attention to which category of bias they are dealing with when coming up with solutions. Mechanistically speaking, attentional biases should be the easiest to overcome, as all that is required is to help decision-makers put in the attentional effort needed to find the relevant information. One can do so by putting procedures in place that remind participants to put in additional work or focus their attention on relevant information. Unfortunately, our analysis suggests that the strategies that have to date successfully worked for reducing attentional biases (e.g., videos or video games that teach specific strategies for focusing attention, see Morewedge et al., 2015) are unlikely to work for representational or emotionally encapsulated biases, which involve representations and emotions that are cognitive impenetrable such that debiasing is more difficult. For this latter category of biases, strategies like situation selection and commitment devices are likely to be more effective since such strategies involve preventing certain representations or emotional states from emerging in the first place. Researchers can also make progress on encapsulated biases by changing the nature of problematic representations slowly over time through exposure to new associations (e.g., reducing implicit biases by slowly learning new less biased associations about race/gender over time). Across both of these types of biases, however, solutions require more than merely knowing about a bias.

Effective solutions for these biases require deliberately designing situations and strategically deploying nudges that change the nature of the decision context differently depending on the nature of the bias at hand.

By better examining how classic cognitive biases fall into these two specific mechanistic natural kinds, we hope to help individuals, organizations, and policymakers adopt effective strategies to arrive at their decisions unimpeded by instinctive and ingrained cognitive impediments. Furthermore, it is not enough for behavioral scientists to teach others to understand these biases, but we also must teach them to understand the limits of metacognitive awareness.

Table 1. Summary

Category	Proposed mechanism	Examples	Debiasing Implications
Encapsulated biases	Knowledge cannot affect biased performance because of the informationally and emotionally encapsulated nature of the bias.	Visual illusions, implicit bias, halo effect, hot/cold states, situationism, present-bias, loss aversion.	Direct debiasing is nearly impossible. Rely on external aids for decision-making. Leverage choice architecture to control exposure and modify situations to encourage desired long-term behavior.
Attentional biases	Cognitively penetrable enough to be overcome through awareness, but people still fall prey to such biases under conditions of distraction or limited attention, because certain aspects of a situation naturally capture attention or because it can be difficult to attend to the relevant knowledge during the moment of decision-making.	Availability heuristic, confirmation bias, base-rate neglect, planning fallacy, hindsight bias, status quo bias, anchoring, affective forecasting, intention attribution.	Direct debiasing is potentially possible but cognitively costly. Outsource decision-making when attention is limited, architect reminders and prompts to "consider the opposite" or "adopt an outside view."

References

Ainslie, G., & Haslam, N. (1992). *Hyperbolic discounting.*

Alesina, A., Carlana, M., La Ferrara, E., and Pinotti, P. (2018) "Revealing Stereotypes: Evidence from Immigrants in Schools." *HKS Faculty Research Working Paper Series RWP18-040, November 2018.*

Ariely, D., Huber, J., & Wertenbroch, K. (2005). When do losses loom larger than gains?. *Journal of Marketing Research, 42(*2), 134-138.

Ariely, D., & Loewenstein, G. (2006). The heat of the moment: The effect of sexual arousal on sexual decision making. *Journal of Behavioral Decision Making,* 19(2), 87-98.

Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. Psychological bulletin, 110(3), 486.

Arkes, H. R., Boehm, L. E., & Xu, G. (1991). Determinants of judged validity. *Journal of Experimental Social Psychology*, 27(6), 576-605.

Arkes, H. R., Faust, D., Guilmette, T. J., & Hart, K. (1988). Eliminating the hindsight bias. Journal of applied psychology, 73(2), 305.

Arkes, H. R., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or "would Jesse Jackson fail the Implicit Association Test?". *Psychological Inquiry,* 15(4), 257-278.

Ayton, P., Pott, A., & Elwakili, N. (2007). Affective forecasting: Why can't people predict their emotions?. *Thinking & Reasoning*, *13*(1), 62-80.

Baert, S. (2018). Hiring discrimination: an overview of (almost) all correspondence experiments since 2005. *Audit Studies: Behind the Scenes with Theory, Method, and Nuance* (pp. 63-77). Springer,

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences,* 30(3), 241-254.

Baron, R. A. (1997). The sweet smell of...Helping: Effects of pleasant ambient fragrance on

prosocial behavior in shopping malls. *Personality and Social Psychology Bulletin*, 23(5), 498-503.

Bazerman, M. H., & Moore, D. A. (2008). *Judgment in Managerial Decision Making*.

Benartzi, S., & Thaler, R. H. (1995). Myopic loss aversion and the equity premium puzzle. The quarterly journal of Economics, 110(1), 73-92.

Bernabei, R., Gambassi, G., Lapane, K., Landi, F., Gatsonis, C., Dunlop, R., Lipsitz, L., Steel, K., & Mor, V. (1998). Management of pain in elderly patients with cancer. *Jama*, 279(23), 1877-1882.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review*, 94(4), 991-1013.

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: the moderation of implicit stereotypes through mental imagery. Journal of personality and social psychology, 81(5), 828.

Bohnet, I. (2016). What works: Gender equality by design. Cambridge, MA: Belknap Press of Harvard University Press.

Brauer, M., Wasel, W., & Niedenthal, P. (2000). Implicit and explicit components of prejudice. *Review of General Psychology*, 4, 79–101.

Bryan, G., Karlan, D., & Nelson, S. (2010). Commitment devices. *Annual Review of Economics*, 2(1), 671-698.

Castilla, E. J. (2015). Accounting for the gap: A firm study manipulating organizational accountability and transparency in pay decisions. *Organization Science*, 26(2), 311-333.

Chang, E. H., Milkman, K. L., Gromet, D. M., Rebele, R. W., Massey, C., Duckworth, A. L., & Grant, A.M. (2019). The mixed effects of online diversity training. *Proceedings of the National Academy of Sciences*, 116(16), 7778-7783.

Choi, J. J., Laibson, D., Madrian, B. C., & Metrick, A. (2003). Optimal defaults. *American*

Economic Review, 93(2), 180-185

Christensen-Szalanski, J. J. (1984). Discount functions and the measurement of patients' values: women's decisions during childbirth. *Medical decision making*, 4(1), 47-58.

Darley, J. M., & Batson, C. D. (1973). "From Jerusalem to Jericho": A study of situational and dispositional variables in helping behavior. *Journal of personality and social psychology*, 27(1), 100.

Dasgupta, N., & Rivera, L. M. (2008). When social context matters: The influence of long–term contact and short–term exposure to admired outgroup members on implicit attitudes and behavioral intentions. *Social Cognition*, 26(1), 112-123.

Doleac, J. & Hansen, B. (2018). The unintended consequences of “ban the box”: Statistical discrimination and employment outcomes when criminal histories are hidden. *Journal of Labor Economics*

Doris, J. M. (1998). Persons, situations, and virtue ethics. *Nous, 32(4),* 504-530.

Dovidio, J. F., & Gaertner, S. L. (1986). *Prejudice, discrimination, and racism*. Academic Press.

Duckworth, A. L., Gendler, T. S., & Gross, J. J. (2014). Self-control in school-age children. *Educational Psychologist*, 49(3), 199-217.

Duckworth, A. L., Gendler, T. S., & Gross, J. J. (2016). Situational strategies for self-control. *Perspectives on Psychological Science*, *11*(1), 35-55.

Duckworth, A. L., Milkman, K. L., & Laibson, D. (2018). Beyond willpower: Strategies for reducing failures of self-control. *Psychological Science in the Public Interest*, 19(3), 102-129.

Duckworth, A. L., White, R. E., Matteucci, A. J., Shearer, A., & Gross, J. J. (2016). A stitch in time: Strategic self-control in high school and college students. *Journal of educational psychology*, 108(3), 329.

Dutton, D. G., & Aron, A. P. (1974). Some evidence for heightened sexual attraction under conditions of high anxiety. Journal of personality and social psychology, 30(4), 510.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249- 267). Cambridge, England: Cambridge University Press.

Egebark, J., & Ekström, M. (2016). Can indifference make the world greener?. Journal of Environmental Economics and Management, 76, 1-13.

Englich, B., Mussweiler, T., & Strack, F. (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. Personality and Social Psychology Bulletin, 32(2), 188-200.

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. Psychological science, 17(4), 311-318.

Epley, N., Mak, D., & Idson, L. C. (2006). Rebate or bonus? The impact of income framing on spending and saving. *Journal of Behavioral Decision Making*, *19*(4), 213-227.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. Behavioral and brain sciences, 39.

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. Journal of Experimental Psychology: Human perception and performance, 1(3), 288.

Fischhoff, B. (1981). *Debiasing* (No. PTR-1092-81-3). Decision Research, Eugene Oregon

Fodor, J. (1990) *A Theory of Content and Other Essays.* Cambridge, MA: MIT Press.

Frederick, S., & Loewenstein, G. (1999). Hedonic Adaptation. Well-Being. *The foundations of Hedonic Psychology*/Eds. D. Kahneman, E. Diener, N. Schwarz. NY: Russell Sage, 302-329.

Frisch, D. (1993). Reasons for framing effects. *Organizational behavior and human decision processes, 54(3), 399-429.*

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychological review,* 102(4), 684.

Gilbert, D. T. (1993). The assent of man: Mental representation and the control of belief. In Wegner, D. M., & Pennebaker, J. W. (Eds). *Handbook of mental control.* Prentice-Hall, Inc.

Gilbert, D. T., Pinel, E. C., Wilson, T. D., Blumberg, S. J., & Wheatley, T. P. (1998). Immune neglect: a source of durability bias in affective forecasting. Journal of personality and social psychology, 75(3), 617.

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of" blind" auditions on female musicians. *American economic review*, 90(4), 715-741.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, 102(1), 4.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of personality and social psychology*, 97(1), 17.

Gross, J. J. (1998). The emerging field of emotion regulation: An integrative review. *Review of general psychology*, 2(3), 271-299.

Haigh, M. S., & List, J. A. (2005). Do professional traders exhibit myopic loss aversion? An Experimental analysis. *The Journal of Finance*, 60(1), 523-534.

Han, S., Lerner, J. S., & Zeckhauser, R. (2012). The disgust-promotes-disposal effect. *Journal of Risk and Uncertainty*, 44(2), 101-113.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. *Behavioral and brain sciences*, 33(2-3), 61-83.

Hoch, S. J. (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11(4), 719.

Hume, David (1998 [1751]) *An Enquiry concerning the Principles of Morals: A Critical Edition*, ed. Tom L. Beauchamp. Oxford: Clarendon Press

Isen, A. M., & Levin, P. F. (1972). Effect of feeling good on helping: cookies and kindness.

Journal of personality and social psychology, 21(3), 384.

Isen, A. M., & Simmonds, S. F. (1978). The effect of feeling good on a helping task that is incompatible with good mood. Social Psychology, 346-349.

Jachimowicz, J., Duncan, S., Weber, E. U., & Johnson, E. J. (2018). When and why defaults influence decisions: A meta-analysis of default effects. *Behavioural Public Policy*, 1-28.

Jackson, M. (2009) Disadvantaged through discrimination? The role of employers in social stratification. *British Journal of Sociology* 60:669–692

Johnson, E. J., & Goldstein, D. (2003). Do defaults save lives? *Science* 1338-1339.

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of experimental social psychology*, 3(1), 1-24.

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *American psychologist*, 58(9), 697.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. Heuristics and biases: *The psychology of intuitive judgment*, 49, 81.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological review*, *80*(4), 237.

Kahneman, D., & Tversky, A. (1982). The psychology of preferences. *Scientific American*, 246(1), 160-173.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, 39(4), 341.

Kalev, A., Dobbin, F., & Kelly, E. (2006). Best practices or best guesses? Assessing the efficacy of corporate affirmative action and diversity policies. *American sociological review*, 71(4), 589-617.

Keltner, D., & Lerner, J. S. (2010). Emotion. *Handbook of social psychology*.

Knapp, A., & Clark, M. S. (1991). Some detrimental effects of negative mood on individuals'

ability to solve resource dilemmas. *Personality and Social Psychology Bulletin,* 17(6), 678-688.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human learning and memory,* 6(2), 107.

Kruger, J., & Evans, M. (2004). If you do not want to be late, enumerate: Unpacking reduces the planning fallacy. *Journal of Experimental Social Psychology*, 40(5), 586-598.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2), 443-478.

Landy, D., & Sigall, H. (1974). Beauty is talent: task evaluation as a function of the performer's physical attractiveness. *Journal of Personality and Social Psychology,* 29(3), 299.

Larrick, R. P., and J. B. Soll (2008), 'The MPG illusion', *Science,* 320, 1593–1594

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual review of psychology*, 66.

Lilienfeld, S. O., Ammirati, R., & Landfield, K. (2009). Giving debiasing away: Can psychological research on correcting cognitive errors promote human welfare?. *Perspectives on psychological science,* 4(4), 390-398.

List, J. A. (2003). Does market experience eliminate market anomalies?. *The Quarterly Journal of Economics, 118*(1), 41-71.

Loewenstein, G. (1996). Out of control: Visceral influences on behavior. *Organizational behavior and human decision processes,* 65(3), 272-292.

Loewenstein, G., Bryce, C., Hagmann, D., & Rajpal, S. (2015). Warning: You are about to be nudged. *Behavioral Science & Policy,* 1(1), 35-42.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: a corrective strategy for social judgment. *Journal of personality and social psychology,* 47(6), 123

Lovallo, D., & Kahneman, D. (2003). Delusions of success. Harvard business review, 81(7), 56-63.

Madrian, B. C., & Shea, D. F. (2001). The power of suggestion: Inertia in 401 (k) participation and savings behavior. The Quarterly journal of economics, 116(4), 1149-1187.

Milkman, K. L., Chugh, D., & Bazerman, M. H. (2009). How can decision making be improved?. *Perspectives on psychological science*, 4(4), 379-383.

Mischel, W. (1974). Processes in delay of gratification. In *Advances in experimental social psychology* (Vol. 7, pp. 249-292). Academic Press.

Moore, D. A., Swift, S. A., Sharek, Z. S., & Gino, F. (2010). Correspondence bias in performance evaluation: Why grade inflation works. Personality and Social Psychology Bulletin, 36(6), 843-852.

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing decisions: Improved decision making with a single training intervention. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 129-140.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of general psychology, 2(2), 175.

Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological Bulletin*, 125(6), 737.

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: evidence for unconscious alteration of judgments. *Journal of personality and social psychology*, 35(4), 250.

Nordgren, L. F., McDonnell, M. H. M., & Loewenstein, G. (2011). What constitutes torture? Psychological impediments to an objective evaluation of enhanced interrogation tactics. *Psychological science, 22(*5), 689-694.

North, A. C., Tarrant, M., & Hargreaves, D. J. (2004). The effects of music on helping behavior: A field study. *Environment and Behavior,* 36(2), 266-275.

Northcraft, G. B., & Neale, M. A. (1987). Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organizational behavior and*

human decision processes, 39(1), 84-97.

O'Donoghue, T., & Rabin, M. (1999). Doing it now or later. *American economic review*, 89(1), 103-124.

Oreopoulos, P. (2011). Why do skilled immigrants struggle in the labor market? A field experiment with thirteen thousand resumes. *American Economic Journal: Economic Policy*, 3(4), 148-71.

Pager, D., Bonikowski, B., & Western, B. (2009). Discrimination in a low-wage labor market: A field experiment. American sociological review, 74(5), 777-799.

Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research and practice. Annual review of psychology, 60, 339-367.

Pashler, H. E. (Ed.). (1998). *Attention*. Psychology Press.

Quadlin, N. (2018). The mark of a woman's record: Gender and academic performance in hiring. *American Sociological Review*, 83(2), 331-360.

Read, D., & Van Leeuwen, B. (1998). Predicting hunger: The effects of appetite and delay on choice. *Organizational behavior and human decision processes*, 76(2), 189-205.

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

Rydell, R. J., McConnell, A. R., Mackie, D. M., & Strain, L. M. (2006). Of two minds: Forming and changing valence-inconsistent implicit and explicit attitudes. *Psychological Science*, 17(11), 954-958.

Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. Journal of risk and uncertainty, 1(1), 7-59.

Sanna, L. J., Schwarz, N., & Stocker, S. L. (2002). When debiasing backfires: Accessible content and accessibility experiences in debiasing hindsight. Journal of Experimental Psychology: Learning, Memory, and Cognition, 28(3), 497.

Santos, L. R. & Gendler, T. S. (2014). What scientific idea is ready for retirement: Knowing is

half the battle. Edge.org. http://edge.org/response-detail/25436.

Schkade, D. A., & Kahneman, D. (1998). Does living in California make people happy? A focusing illusion in judgments of life satisfaction. *Psychological science,* 9(5), 340-346.

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. *Journal of personality and social psychology,* 45(3), 513.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: another look at the availability heuristic. *Journal of personality and social psychology,* 61(2), 195.

Sellier, A. L., Scopelliti, I., & Morewedge, C. K. (2019). Debiasing training improves decision making in the field. *Psychological science,* 30(9), 1371-1379.

Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological bulletin,* 134(2), 207.

Shepard, R. N. (1990). *Mind sights: Original visual illusions, ambiguities, and other anomalies, with a commentary on the play of mind in perception and art.* WH Freeman/Times Books/Henry Holt & Co.

Simonsohn, U. (2007). Clouds make nerds look good: Field evidence of the impact of incidental factors on decision making. *Journal of Behavioral Decision Making,* 20(2), 143-152.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological bulletin,* 119(1), 3.

Soll, J. B., Milkman, K. L., & Payne, J. W. (2015). A user's guide to debiasing. *The Wiley Blackwell handbook of judgment and decision making,* 2, 924-951.

Stanovich, K. E. (1999). *Who is rational?: Studies of individual differences in reasoning.* Psychology Press.

Stanovich, K. E. (2009). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In *two minds: Dual processes and beyond,* 55-88.

Sunstein, C. R., & Reisch, L. A. (2014). Automatically green: Behavioral economics and environmental protection. Harv. Envtl. L. Rev., 38, 127.

Tetlock, P. (1985). Accountability: A social check on the fundamental attribution error. *Social psychology quarterly*, 48(3), 227-236.

Thaler, R. H., & Benartzi, S. (2004). Save more tomorrow™: Using behavioral economics to increase employee saving. Journal of political Economy, 112(S1), S164-S187.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of applied psychology*, 4(1), 25-29.

Tilcsik, A. (2011). Pride and prejudice: Employment discrimination against openly gay men in the United States. American Journal of Sociology, 117(2), 586-626.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological bulletin*, 76(2), 105.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and Probability. Cognitive psychology, 5(2), 207-232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185 (4157), 1124-1131.

Tversky, A & Kahneman, D. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2), 263-292.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481), 453-458.

Uhlmann, E. L., & Cohen, G. L. (2005). Constructed criteria: Redefining merit to justify discrimination. *Psychological Science*, 16(6), 474-480.

Weichselbaumer, D. (2003). Sexual orientation discrimination in hiring. *Labour Economics*, 10(6), 629-642.

Wertenbroch, K. (1998). Consumption self-control by rationing purchase quantities of virtue and vice. *Marketing science*, 17(4), 317-337.

Wetzel, C. G., Wilson, T. D., & Kort, J. (1981). The halo effect revisited: Forewarned is not forearmed. *Journal of Experimental Social Psychology*, 17(4), 427-439.

Wheeler, S. C., & Petty, R. E. (2001). The effects of stereotype activation on behavior: A review of possible mechanisms. *Psychological Bulletin,* 127, 797–826.

Whillans, A. V., & Dunn, E. W. (2015). Thinking about time as money decreases environmental behavior. Organizational Behavior and Human Decision Processes, 127, 44-52.

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: unwanted influences on judgments and evaluations. *Psychological bulletin,* 116(1), 117.

Wilson, T. D., & Gilbert, D. T. (2003). Affective forecasting. *Advances in experimental social psychology*, 35(3), 345-411.

Wilson, T. D., Wheatley, T., Meyers, J. M., Gilbert, D. T., & Axsom, D. (2000). Focalism: A source of durability bias in affective forecasting. *Journal of personality and social psychology*, 78(5), 821.

Wood, W. (2019). *Good Habits, Bad Habits: The Science of Making Positive Changes that Stick*. Pan Macmillan.

Wright, B. R., Wallace, M., Bailey, J., & Hyde, A. (2013). Religious affiliation and hiring discrimination in New England: A field experiment. Research in Social Stratification and Mobility, 34, 111-126.

Yap, S. C. Y., Wortman, J., Anusic, I., Baker, S. G., Scherer, L. D., Donnellan, M. B., & Lucas, R. E. (2017). The effect of mood on judgments of subjective well-being: Nine tests of the judgment model. *Journal of Personality and Social Psychology*, *113*(6), 939-961.

Yoon, H., Scopelliti, I., & Morewedge, C. K. (2020). Decision making can be improved through observational learning. *Organizational Behavior and Human Decision Processes*, 162, 1: 155-188.