Discovering Implicit Bias

Understanding Conflicts between Our Conscious and Unconscious Beliefs

Honors 320

Great Questions Essay

By the time my fourth semester of college arrived, I felt that I had a good handle on navigating the first day of classes. After visiting each class, I came away with a favorable impression of all my professors, except for one who I was not quite sure about. Professor Anderson[1] was pleasant and friendly, but something seemed off that I could not quite put my finger on. While to my recollection, she did not do a single thing to deserve my suspicion, I sensed she had a sweet exterior but would soon show a nastier side. I began working hard in the class and waited for the day this side of her would surface, but that day never came. As time went on and she was still the same lovely person she had always been, I grew increasingly embarrassed at my initial assessment of her. By the end of the semester, she was one of my favorite instructors and I was left with an experience that was positive, yet tainted by my initial, completely unfounded prejudice. When I asked myself why I had felt this way, I couldn't help noticing that if a male professor had acted in the exact same way on the first day, I probably would have had no issue with him whatsoever.

Did this mean I was sexist? Ever since my first semester when I had managed to take eight classes without even one female instructor, I had strongly believed that my university needed more female faculty. I made a point to take classes from female professors when I could, which became especially important to me after I enrolled in a heavily male-dominated major as a female student. I often defended unpopular female leaders or politicians in conversation, arguing that their seemingly unlikeable traits, such as ambition and assertiveness, were considered admirable in men. Yet here I was, disliking a female instructor right off the bat for no logical reason I could think of. Why, despite my own beliefs, did I still seem to subliminally prefer being taught by a man? It may have been innocent - I knew there could be all kinds of less extreme explanations for why she rubbed me the wrong way. Perhaps she had simply reminded me of someone I knew before who turned out to be unpleasant or insincere. Still, I couldn't help worrying what this incident said about me.

Experiences like this stand out in my mind since in today's society, biases about characteristics such as gender, race, and religion are particularly egregious. However, I have also discovered unexpected subconscious bias in more benign contexts. When discussing bias with a professor recently, he pointed out that while we did not consciously think about it, we had both

[1] Name has been changed.

probably believed there would not be an elephant in the room we just walked into[2] – a reasonable assumption, but an assumption that is not based in any concrete evidence other than past experience, and one that would be dangerous if it turned out to be incorrect. I'm sure I could think of countless other silly examples of assumptions I make every day, but I wonder if I am able to tell when an assumption crosses the line from harmless and inconsequential to more problematic. For example, while not necessarily offensive, incorrect first impressions have led me to pay little attention to people who could have been great friends. As I notice more and more subliminal beliefs in myself that have gone unnoticed for a long time, it makes me wonder how many others I am still unaware of.

My first reaction when I discover a personal prejudice, especially one related to a sensitive subject such as gender, is to worry that I am a terrible person. Does this happen to everyone, or am I in the minority who have trouble acting consistently with what they believe (or think they believe) is right? This is a difficult topic to discuss with others. Of course, some people are blatantly prejudiced, but among those who care about social equity and treating everyone with respect (the majority, I would like to think), many are reluctant to admit any tendency towards bias. Are others simply better or more ethical than I am? Am I the only one who sometimes feels like a hypocrite?

As easy as it is to get caught up in this worrying, my experience suggests that on some level, everyone has some subconscious beliefs that contradict their stated principles. This observation is backed by a plethora of research on what psychologists call "implicit bias." Several studies suggest that people can act on bias "without intending to do so," and that implicit bias against marginalized groups such as women, racial minorities, and members of the LGBTQ+ community is especially common (Brownstein). A person's explicit views may favor an egalitarian society, but his implicit views may be more discriminatory. How this bias develops without his awareness, and why it can contradict his stated beliefs, is less clear. Even if it is hard to recognize, implicit bias certainly exists in many forms. For example, in these last couple of sentences, my use of male pronouns to refer to an unspecified singular person shows some of my own implicit bias.

---

[2] Not to be confused with the metaphorical elephant in the room – my professor was referring to the possibility of a physical elephant sitting in our classroom, or anything else highly improbable that we can safely assume is false. Other examples could include that as I go about my day, I usually don't believe that a rhinoceros is behind me, or that Mick Jagger is in the next room.

If bias is so prevalent, why is it hidden from our view much of the time? Why did it take almost a whole semester for me to consider the possibility of inner bias against Professor Anderson? One possible reason for my lack of awareness is that our biases become "driven underground" as society's definition of what is acceptable changes (Brownstein). Somewhat ironically, bias and stereotypes are often created and perpetuated by social norms, yet these same social pressures can cause people to suppress their bias once it becomes taboo. It can be tempting to think this suppression means that now-frowned-upon prejudices from the past are gone. As a young kid, I remember learning about racism in decades past and feeling happy that I didn't see anything like that around me.[3] In reality, though, this suppression does not usually eliminate prejudice altogether. Instead, it may simply push prejudice to the subconscious, making us susceptible to acting on it without realizing we are doing so. To a child, it may look like the prejudice has been conquered, when it may have just become more subtle.

Even though this naivety colored much of my childhood, I remember at least one early experience that helped me start thinking about bias. My 4th grade social studies class did an activity on the Revolutionary War that involved each of us declaring whether we would have been a Patriot, a Loyalist, or neutral. The class was fairly evenly split between Patriots and neutrality – I of course chose to be a Patriot. Only one person chose to be a Loyalist, and for him the pronouncement seemed more due to contrariness than to any strong personal conviction. Looking down on the cowards who would not stand up for their country, I felt justified in my decision, but even then, I sensed something didn't quite add up. If no one in a group of 30 different people would have sided with the British, then how was it possible that this was the position of 1/3 of colonial Americans? At the time, I wasn't yet ready to question the validity of my own stance, but I started to think there must be some disconnect between what we felt we believed and what we would have thought if we were in that situation.

Certainly, much of this disconnect can be chalked up to the lack of nuance in a 9-year-old's mind. The group's answers likely would have been different if we did this exercise in a college class – I'm not sure anymore what mine would be. Still, the experience has stayed with me, making me wonder when and why people's behavior might not line up with their stated

[3] Of course, racism is our modern society is not all subliminal; we are still struggling to address the covert racism that people of color experience every day. However, it is certainly true that racism has become less socially acceptable in many circles, meaning that much of our bias has become more hidden.

beliefs. If our actions are not always informed by our conscious beliefs, what are the underlying ideas that *do* inform them? One viewpoint that undoubtedly influenced my 4th grade class was the way history had been taught to us. History contains multiple interlocking narratives, but many curricula try to force all of them into one story. Even if teachers try to be unbiased, they usually end up focusing on the viewpoints of white men and victors in conflicts, while other perspectives are lost (Conway). Knowing that the Patriots were the winners who went on to shape fundamental American values, aligning myself with King George III or failing to take a stance at all seemed unthinkable. However, I later realized that the students who chose differently than me may have been displaying a sophisticated sense of self-awareness. They were able to uncouple their subliminal beliefs from the point of view they had been taught to support.

Even though my education became less one-sided as I got older, the way I learn still depends on context. The underlying beliefs that affect my behavior are likely picked up as I go through life, making observations and living experiences without fully knowing what they mean. However, we develop conscious beliefs like "Patriots (or suffragettes, or the Allies) were the good guys and everyone else was wrong" from a much different vantage point. When we have the benefit of looking back on history or analyzing a situation from the outside, drawing conclusions can be straightforward. However, we do not have this luxury when trying to glean knowledge before we know how a situation turns out.

Perhaps this explains why I continue to see disconnects between belief and action among adults. Many value female representation in the workplace, yet unknowingly find an idea more appealing coming from a man than a woman (Brownstein). Many praise the protests of the Civil Rights movement, yet react to today's protests with annoyance or even anger. Do I do this too? In the most extreme case, I and most others are horrified by tragedies such as the Holocaust or the Rwandan genocide, yet if we were transplanted into one of those environments, there is no guarantee we would stand up to the atrocity or even quietly refrain. This depressing notion is supported by the sobering reality of present-day genocides that go largely ignored, such as the Rohingya in Myanmar or the Nuer in South Sudan (Kranz). In cases where large percentages of the population were perpetrators or bystanders, the odds of having been one of the "good guys" if I was there are not necessarily in my favor. Even with strong principles, am I so susceptible to cultural and societal influences that I would act in violation of those principles? Is it possible to

become fully aware of what my beliefs even are? Perhaps, as social creatures, it is inevitable that society will instill subliminal beliefs that can override the values we believe are guiding us.

In these examples, it seems our understanding of history often involves classifying some people as part of a group entirely separate from ourselves. They are either enemies to be mistrusted or faceless statistics to be treated with indifference. Why is this type of thinking so widespread? Philosopher Georg Hegel argued that human consciousness is being recognized as human by another human consciousness. At first, it seemed peculiar to me that we need other people to even be conscious of our own thoughts. However, other philosophers have pointed out that if a person was never recognized as human by any other humans, they would not know what it means to be human (Hatcher 46). This presents an interesting conundrum regarding my relationships with other people: I need the construct of the Other to understand myself, but classifying people as the Other can cloud my perception of them. This can make it easy to misrepresent or oversimplify the people around me. Not only does this hinder communication, but it causes many societies to treat entire groups of people as subhuman.

In my view, though, philosophers fail to adequately explain *how* my understanding of the Other shapes my worldview. For instance, do we always see ourselves as opposites of the Other, or might we also use this concept to find similarities? In reality, we are often biased towards thinking that people are *more* similar to us than they really are, not more different. We tend to project our view of ourselves onto people around us. I see two issues with this type of thinking: First, if I draw conclusions based on my internal thoughts and the limited information I have about others, I am not using a random sample. In formal research, using a large randomized data set is one of the most important ways to avoid bias because it allows researchers to best represent the group they are trying to understand (Taniguchi et al). Since projection is not based on a random sample, it likely gives us a highly inaccurate view of other people, even if it feels intuitive. Second, we do not apply our projection equally – we generally ascribe good intentions to our own actions, and we are far more likely to project our own good intentions onto members of an in-group than members of an out-group (Alicke 31). In other words, I am more inclined to give people the benefit of the doubt if I see them as similar to me. This helps explain why it is so easy to demonize people we think are on the wrong side of history, but it does not provide a fully satisfying explanation for my own bias. After all, being a woman myself, Professor Anderson was not part of an out-group, yet I still may have misjudged her due to her gender.

Maybe the same forces that divide people into categories can foster negative views of certain groups, even by people within those groups. Some argue that subliminal bias is nothing more than mental associations of certain concepts or attributes. These associations are often shaped by cultural values; for example, in societies with few educated people, the trait "educated" is usually associated with the trait "conscientious" (Zebrowitz 31). Associations can develop without much thought on our part, but they inform our first reactions in many situations. Recently, a friend interviewed me for a class project on regional stereotypes. She listed several regions of the United States and asked me what I thought about people from those regions. Essentially, my answers were the subliminal associations I had developed through some personal experience, but mostly through exposure to cultural stereotypes. For example, I associate the Pacific Northwest with outdoorsy, laidback people; the Northeast with reserved intellectuals; and the South with conservative farmers who rarely venture out of their hometowns.

Mental images such as these seem fairly innocent, especially when based in some truth, but I think they can become problematic when we attach moral connotations to them. "Reserved and intellectual" can quickly become "pretentious and snobby," while "rural" can become "ignorant and provincial." Depending on who you ask, a "laidback and outdoorsy" person could be seen as a caring, easygoing environmentalist or derisively dismissed as a bleeding-heart tree-hugger. Before we know it, we can make character judgments about people based simply on the subconscious associations our minds have created. When I meet new people, it is often hard to figure out why I have a good or bad first impression of them, but maybe the answer is that they have somehow triggered one of my subliminal associations.

As I answered the interview questions about regional stereotypes, I realized I had never really asked myself where these associations came from. I cannot pinpoint any exact moments, but I know I have seen stereotypes in a lot of popular media. For example, I recently discovered the sketch comedy series *Portlandia*, which makes fun of hippie culture in Portland, Oregon (Armisen). I have never watched a full episode of the reality series *Duck Dynasty*, but its portrayal of backwoods evangelical "rednecks" seems to reflect common views of the American South (Gurney). Another example that comes to mind is the way the film *Legally Blonde* contrasts different regional stereotypes. The protagonist, Elle Woods, is a bubbly, appearance-driven girl from southern California who attends law school at Harvard, where the students are much more serious and always dressed in drab clothes that contrast with her pink wardrobe

(Platt). However, the focus of this film is not regional stereotypes, but gender stereotypes: in particular, the image of the dumb blonde. Due to her wealth, appearance, and cheery personality, Elle struggles with being seen only as an unintelligent sorority girl.

Gender stereotypes are common in other films as well. The first time I saw the movie *Mean Girls*, I thought the antagonist Regina George was the perfect embodiment of a "fake" girl, one who is nice to your face and cruel behind your back. Regina, the most popular girl in her high school, uses friendliness as a manipulation tactic (Michaels). In the scene that stood out to me most, a girl walks by Regina at school and Regina compliments her skirt. As soon as the girl walks away, Regina says to her friend Cady, "That is the ugliest skirt I have ever seen." Cady then looks at her bracelet, remembering when Regina sweetly complimented it the first day they met. Portrayals like these cement the stereotype of an outgoing, popular girl who can't be trusted. While Professor Anderson didn't have much in common with Regina, maybe something about her friendliness reminded me of a devious "mean girl" I had seen on TV.

To understand how stereotypes from the media and other cultural influences become entrenched in my mind, I found it useful to more deeply consider how we learn. Human thinking and learning can generally be categorized into two separate processes: controlled and automatic. Controlled processing demands our attention and has a limited capacity, but automatic processing happens constantly and involuntarily as our brains react to the stimuli around us (Payne and Gawronski 2). When I am talking to someone, thinking of what words to say is often a conscious process, especially if it is a particularly important conversation such as a job interview. However, the physical act of speaking is done automatically; once I know what I want to say, I do not have to think about how to move my mouth to form those words. An important role of automatic processing is filtering through our conversations, the media we watch, and other outside influences. If watching movies like *Mean Girls* made me more suspicious of other women despite my explicit egalitarian beliefs, then my automatic processing was probably responsible. Associations like these may be biased, but they are difficult to avoid since they are often based in some concrete observation. For example, failing to associate women and science may not be an accurate reflection of women's abilities or interests, but it reflects the reality that most scientists are in fact men (Kelly 528).

It is disconcerting to think about everything my mind is doing without me knowing, but I don't think I need to worry too much unless my automatic processing is influencing my behavior

in ways I don't like. In some cases, implicit bias doesn't affect behavior much; if we have enough time to think through a decision, we can circumvent most of our bias by using controlled processing instead of automatic. However, in situations where we must react quickly, we rely on our automatic processing to shape our reflexive decisions (Brownstein and Saul 48). This means that even if some of my mental associations are usually suppressed, they can reveal themselves in my knee-jerk reactions. For example, I never tried to treat Professor Anderson any differently, but it is possible that in certain moments, I did not always react to her as warmly as I wanted to. This reminds me of the film *Zootopia*, in which the protagonist, Judy Hopps, holds tolerant explicit beliefs but displays bias in some of her impulsive reactions (Howard). Judy is a rabbit who defies stereotypes by becoming the first bunny police officer; after seeing her overcome so much opposition, I did not expect her to ever perpetuate stereotypes herself.

However, Judy shows her inner bias when she tells her fox friend Nick Wilde that she doesn't see him as "one of them," and instinctively reaches for her fox repellent when he becomes angry. (I also found it significant that she felt uncomfortable with her parents giving her the fox repellent, but then carried it with her anyway.) It seems that because of external influences such as news media and her family's opinions, Judy was carrying a subliminal mental association that made her wary of foxes. Watching this, I grew uncomfortable as I wondered if I have any similar implicit attitudes that manifest in my knee-jerk reactions. When walking alone in public, I can't deny that I sometimes feel more nervous around some people than others, based solely on their physical appearance or even their race.[4] These moments are usually surprising to me because I am not used to thinking of others as untrustworthy due to their appearance. If I am startled or on edge, though, my automatic processing can kick in, sometimes betraying an ugly bias that usually stays hidden.

In moments like these, when my gut responses betray some deep-seated bias, am I showing what I truly believe, or am I simply reflecting the influences around me? While philosophers disagree on this question, many seem to favor the latter explanation. For example, philosopher Tamar Gendler asserts that implicit attitudes such as subliminal racism cannot truly be considered beliefs (Schwitzgebel). Rather, they are habitual lines of thinking that are

[4] This particular instance of implicit bias is difficult to navigate, since it is still important for us to feel safe in our own person. I do not believe people should put themselves in situations where they feel unsafe simply to feel less guilty. However, this form of bias is still prevalent and can be harmful.

culturally imposed. Since they are not based on systematic deliberation or evidence, in her view the development of these attitudes does not constitute the type of learning that shapes a person's worldview. It seems that in a world of over-stimulation, we may sometimes revert to a default mindset instead of developing our own beliefs for every situation, and unfortunately this can perpetuate oppressive cultural attitudes.

Humans are not the only entities who display Occam's Razor-like tendencies when learning from experience and drawing conclusions.[5] In machine learning, computers are programmed with a learning algorithm that allows them to study a set of training data and use reasoning to apply what they learn to other data sets. This is useful for tasks such as handwriting recognition, which may require a computer to work without explicit instructions. Programmers have a variety of learning algorithms to choose from, such as decision trees and neural networks, and almost all of them are biased in favor of simpler solutions[6] (Mooney 2). Bias can also result if an algorithm is not detailed enough. Sometimes algorithms have insufficient features, meaning that they are not equipped to identify a large number of characteristics for each object they consider. When this is the case, they may infer false relationships between characteristics, which can create bias (Arrieta et al 38). This seems similar to the tendency of human nature to leap to conclusions, taking an observation and deciding what it means without sufficient evidence. For example, I may observe a friend treating me a little more distantly than usual and quickly conclude she is mad at me. However, if I took time to consider more factors, I might realize that her current demeanor is not actually related to her feelings towards me. Instead, there could be a myriad of other things that had put her in a low mood.

It is interesting that humans and computers share these similarities, but how do the learning algorithms used by computers actually work? Can computer learning be compared to my own learning? One of the most common types of reasoning used by machines is induction, which takes specific observations and uses them to draw more general conclusions that can apply to other problems. Many machine algorithms take the form of a decision-tree; much like a "choose your own adventure" story, the tree contains a series of junctures that each lead to a few

[5] Occam's Razor essentially states that the simplest solution or explanation is usually the best one.

[6] Some of the most well-known instances of bias in computer programming are caused by bias in programmers themselves. For example, programmers have created facial recognition technology that is better at recognizing white faces. I have chosen not to include these examples because I want to focus on bias in the algorithms themselves rather than the programmers, but these forms of bias are still prevalent.

different paths. When induction from one data set is applied to a new set, these decision tree algorithms are sometimes biased towards certain paths (White). Programmers must manage this bias carefully to avoid false conclusions, but I was interested to learn that some form of bias is necessary in inductive algorithms. When computer scientists use the term "bias," they are often simply referring to the assumptions they must make when instructing a machine how to learn. Since induction involves generalizing from a specific example, it inherently involves making some assumption about how the specific example they have already studied will connect to other situations (Rich). This means that part of the task of a programmer is to figure out which assumptions will lead to the most accurate conclusions.

As it turns out, humans use inductive reasoning too, and this makes us susceptible to biased logic. We make observations about our experiences and apply them to other situations, and like computers, this can lead to erroneous inferences. David Hume wrote extensively on the problem of inductive logic (Henderson). Deductive reasoning, which applies general observations to more specific situations, can follow logical rules that make its arguments sound and complete. However, since inductive reasoning starts specific and becomes less precise, there is no standard method for using induction in an accurate way. In other words, induction is essentially guessing based on experience, meaning that reaching probable conclusions is the best we can do. One source of inaccuracy from induction is that we may learn something from one situation and assume it applies to other situations, even if the two contexts are not comparable. Hume used the example of seeing white swans and assuming all swans are white, when in fact some are black. This seems innocuous, but the same thought process could lead someone to falsely believe that all Muslims are terrorists ("Inductive Reasoning").

Unfortunately, ridding myself of false beliefs could be quite difficult because humans are susceptible to confirmation bias, meaning we mainly internalize evidence that supports what we already believe. Once we adopt a belief through induction, we may cling to it even if we come across contradictory evidence ("Inductive Reasoning"). How, then, can I limit bias in my own reasoning and make sure my inductive assumptions lead me to the most probable conclusions? Since computer scientists are experienced in managing inductive bias, perhaps I can adapt their methods. One way for algorithms to correct their bias is through lifelong learning. If machine learning algorithms continuously analyze data and learn new patterns even after they have performed an initial task, they can identify and change the biases that lead to inaccurate results

(Amir and Meir 52). Maybe I could do this more intentionally in my own life. My brain may not naturally seek out evidence that goes against my beliefs, but I imagine I could get better at correcting my bias if I deliberately open myself up to new information.

I now have a better grasp of where my bias comes from, but the explanations I have found thus far are not completely satisfying. Is bias simply something I have reasoned my way into, whether consciously or unconsciously? If I had better reasoning skills and access to more impartial media, would that be enough to remove my bias? I'm sure it would go a long way, but my bias still feels a little deeper than that. Perhaps some of my bias is encoded into my nature; like a computer algorithm, a little bias may be necessary for my brain to work effectively. According to Darwin's theory of natural selection, many traits have evolved in humans and other animals because they help their species to survive in some way (Jabr). Survival traits can lead to bias; for example, a tendency toward xenophobia may have been advantageous because mistrust of unfamiliar outsiders allowed people to avoid conflict or infectious disease (Haselton et al). Like Judy Hopps from *Zootopia* impulsively protecting herself from a fox, some of my instincts may be biased because my genes are trying to protect me from danger.

However, humans and animals have many preferences, especially aesthetic preferences, that cannot be easily linked to any evolutionary advantage (Jabr). For example, many biologists debate whether animals' beauty preferences can be explained by natural selection, especially since the opulence displayed by many animals trying to attract a mate does not seem like it could possibly be practical or energy efficient. Biologists are beginning to study an attribute they call "sensory bias," which describes the way that ecology and evolution can explain seemingly capricious preferences. Sensory bias is believed to come from our genetics just as our survival traits do, but the preferences shaped by our sensory bias have no adaptive purpose other than beauty or pleasure (Jabr).

The concept of sensory bias may provide a simple explanation for many of my innate behaviors or partialities I cannot quite account for. This can include simple preferences in food or music, which often differ among otherwise similar people for no apparent reason. More importantly, though, this could mean that my genes make me feel drawn toward or away from certain people, even if it is not rational. Some beauty preferences stem from cultural standards or evolutionary advantages, but others appear arbitrary and innate; after all, "beauty is in the eye of the beholder" is a common phrase that alludes to the wide variety in what people consider

appealing. These preferences have real implications in society, since people tend to treat others more favorably when they find them attractive (Marczyk). While diversity in preferences makes our world richer, it is worrisome to me that something so random could affect people's treatment of those around them. Could sensory bias be enough to make a future job recruiter pass me over, even if I am otherwise qualified? Psychologists would probably argue that recruiters' biases come from some sort of stereotype, but even if I don't bring any stereotypes to mind, maybe a recruiter could still get a bad impression of me simply because of their genes.

Sensory bias may explain the randomness of my likes and dislikes, but often, nothing seems more capricious to me than the emotions of myself or others. Fluctuation in moods is a major factor in my mental state, adding another layer of complexity to my mindset. I am not alone in feeling moody sometimes; empirically, our emotions play a large role in determining how we interpret our surroundings. Both humans and animals make more pessimistic judgments when feeling fearful or depressed, and more optimistic judgments when feeling safe and content (Clegg 2). This implies that moments of particularly strong emotion could make us susceptible to subliminal beliefs that are ultimately inconsequential. In a book that has dramatically shaped my views on the importance of mindfulness[7], psychotherapist Richard Carlson wrote that people dramatically overestimate the magnitude of their problems when they are experiencing a low mood (Carlson 56). Similarly, the changing emotions of others often make it difficult for us to interpret their thoughts and intentions (Carlson 60).

Additionally, extended periods of time dominated by one emotion could presumably skew someone's mindset towards a few specific tendencies. In the 1999 comedy-drama series *Freaks and Geeks,* protagonist Lindsay Weir is a well-behaved straight-A student who starts acting rebellious during her junior year of high school (Apatow). Initially, this behavior change merely seems to be a case of teenage disillusionment, much like Holden Caulfield in J.D. Salinger's *The Catcher in the Rye*.[8] Later in the series, though, viewers learn that Lindsay's rebellion comes from her grief and existential confusion surrounding the death of her grandmother. This blinds her to the ways in which her behavior is ultimately self-destructive. However, when her friend Millie starts to go down the same rebellious path, Lindsay tries to stop

[7] I would define mindfulness as the ability to acknowledge your thoughts and emotions without letting yourself become defined or controlled by them.

[8] Holden Caulfield is a cynical 15-year-old boy who thinks everyone else is a "phony," yet cannot see that he himself often acts like a phony as well.

her. Since she is less distracted by emotion when looking at the actions of someone else, Lindsay is able to achieve a clarity in her view towards Millie that she cannot achieve for herself.

Clearly, our emotions influence our thinking, but this relationship goes the other way as well; our cognitive processes affect the way we experience and interpret emotions (Clegg 1). While it may seem painfully obvious that our thoughts influence our emotions, this carries the important implication that bias can be extremely helpful if we want to feel a certain way. I often worry about trying to eliminate as much bias as possible, but up until now I have failed to consider the ways I can actually utilize my bias. Some self-help gurus claim biased beliefs may not always be bad, stating that a true belief does not necessarily equate to a helpful belief. At first it seemed odd to me that anyone would want to seek out untrue beliefs, but I realized my behavior demonstrates that I already understand this principle on some level. One of the most recognizable examples is the common wisdom that confidence can boost performance, even if it is not based in objective truth. In fact, a bias toward overconfidence is often ideal. When I am in the last stages of preparing for a job interview or a musical audition, I certainly find it more useful to focus on telling myself how talented and capable I am, whether or not this is really true.

As I thought about the ways in which sacrificing exact truth could be beneficial, I found that philosophers have a lot to say on this topic, especially as it relates to ethics. Just as there is sometimes a tradeoff between truth and a confident attitude, many beliefs contain a tradeoff between truth and moral objectives. This tradeoff can be seen in racial bias, which often causes recruiters to underestimate the true merit of black candidates for jobs or spots at a university. According to philosophers, the way we address our bias should depend on whether we think it is causing an ethical problem or a knowledge problem. If recruiters simply make themselves aware of their bias and award a couple more points than they think they should, they are solving a knowledge problem by trying to make sure they see each person's true merits. This would be equivalent to learning that people on average underestimate how long a certain task will take by five minutes, and then adding five minutes to my own estimate. There are no moral implications other than simply trying to find the most precise measure of candidates' objective merit.

However, affirmative action often goes beyond this by giving certain minorities extra consideration beyond their objective merit, usually in an effort to compensate for systemic obstacles they may have faced. This is solving an ethical problem rather than a knowledge problem: recruiters may believe that giving a step up to people who have been disadvantaged is

the right thing to do even if some other candidates are technically more qualified (Kelly 533). Because of this, affirmative action itself is often seen as a form of bias, but many argue that this doesn't matter if it fulfills a more important ethical duty.

This means that if I want to address my own bias, I should try to decide whether I want to pursue the exact truth, or whether a slightly biased belief would serve me better. For example, if I am conducting a research experiment, I should strive for exact accuracy and objective truth. Many social scientists have ironically found that even when trying to conduct behavioral and cognitive research, they themselves are biased in their own behaviors and cognitive processes (Nelson 212). This is undesirable because it holds them back from the knowledge and understanding they seek. However, my bias may play a different role in my relationships. In close relationships, most people prefer to err on the side of thinking the best of someone, even if this does not always accurately represent the person's character (Egan 77). In this case, we judge a sense of goodwill and trust to be more important. If I feel someone has acted rudely, I try to give them the benefit of the doubt as much as possible, since I believe it is the right thing to do and I have found it makes me feel happier. This may mean that I will occasionally misjudge someone's true intentions, but this sacrifice is worth it to me.

Because of examples like these, it could be easy to think that bias can help in subjective, interpersonal matters, while it should be avoided in empirical analysis. However, I have already learned that bias in machine learning algorithms is not only useful, but often necessary. How do programmers make sure that the bias in an algorithm is helping and not hurting their results? Since certain biases are useful for different types of problems, programmers must become familiar with the strengths and weaknesses of different algorithms, and then choose the right bias for the right problem (Mooney 2). When utilized well, bias saves processing power and solves problems more efficiently. However, programmers have a tradeoff to consider: if a bias toward a certain solution is useful for one type of problem, it will likely perform poorly if applied to a different type of problem (Montañez et al 2). In my own life, then, maybe this means I shouldn't expect to solve every type of problem in the same way. I may find that a certain mindset helps me to get along well with one person, but an entirely different mindset may be better for interacting with another person.

This potentially has exciting implications for me if I can learn to harness the power of my cognitive tendencies to solve the problems I experience. However, I imagine this could require

an exorbitant level of self-awareness that may be difficult to achieve. It seems that there must be a difference between holding subliminal bias that sometimes happens to find a valid solution, and becoming aware of my bias in a way that allows me to deliberately utilize it. Acknowledging that bias is not always bad can open the door to making the most of my cognitive processes, but how am I to navigate this when the relationship between implicit and explicit beliefs seems so complicated?

When I attended an Especially for Youth[9] program (EFY) sponsored by my church as a teenager, the speaker I remember most told us that our beliefs are the main determinant of our behavior. In both religious and secular contexts, many of us try and fail repeatedly to become healthier, more productive, or kinder people. According to the speaker, our bad habits have become habits for a reason; any changes we try to make can never be sustainable if we do not address the underlying beliefs behind our actions. He argued that if we find ourselves constantly using social media or staying in bed instead of exercising, we must believe on some level that this will help us somehow.[10] At the time, this was a fascinating idea for me; I had never fully considered the power of my mindset before. I did not yet have the terminology to articulate this, but it seems to me that the speaker was referring mostly to our implicit beliefs. For example, if someone is struggling with a goal of yelling at their family less, they must already believe explicitly that breaking this habit will benefit them, but they may still implicitly believe that yelling will get them what they want.

I don't see this philosophy as a silver bullet for solving all bad habits, but I have tried to apply it over the years, with some success. However, as time passed, I became more pessimistic. The speaker had not specified *how* we were supposed to change our beliefs, and I wasn't sure it was as easy as he seemed to believe. When I contemplate the different layers of knowledge and belief in my mind, I often come back to my irrational fear of candles. As a child, I would stare in awe as my dad quickly flicked his finger back and forth through a candle flame without being burned. Growing up, I saw many others do the same thing and even learned some science behind this, but I have never been able to bring myself to put my own fingers anywhere near a candle flame. I know it won't hurt me, but for some reason I just don't believe it. I cannot recall any

[9] This is a weeklong religiously-based camp where teenagers stay at a college campus and participate in activities such as workshops, classes, game nights, and dances.

[10] Of course, changing one's habits is often more complex than this, especially when addiction is involved. However, for this paper I am focusing on the role of our inner mindsets in shaping our actions.

traumatic personal incidents with fire or burns, so I am not sure what, if anything, caused this belief or how I could go about changing it.

I suppose if I took a leap of faith and experienced it myself without getting burned, I could rid myself of this illogical fear. However, that just brings me back to the initial problem of how to change; it would hardly be helpful for someone to tell me to just buck up and put my finger through the flame. Is there anything more I can do to change my ingrained beliefs? One might truthfully think that I have not yet overcome this obstacle simply because it is not important. Still, this type of mental block can extend to things that *are* important, such as overcoming destructive habits or attitudes. These situations are often far more difficult to deal with than my candle conundrum. With my fear of candles, I can easily identify a specific belief holding me back, but this is not always so simple. Considering the many ways we can develop subconscious bias, how can I change my beliefs if I can't be sure I understand what they are to begin with? Even if I succeeded in changing a belief, would the change be deep enough to reach my behavior? How can I be sure the external influences that caused my beliefs will not continue to affect me in the same way?

Years later, I read a blog post on changing our habits that was a little more nihilistic than the EFY speaker had been. Essentially, the author said it is all too easy to become insecure when we see self-improvement nuts who wake up early every day and find time to meditate, exercise, journal, and eat only vegan home-cooked meals. Those people, for whatever reason, like doing those things and want to keep doing them. This rang true to me; after all, can you think of anyone you know gets up at 5 am to do an extended morning routine, who doesn't ultimately enjoy it? It seems that significant behavior and lifestyle changes are impossible unless we believe they will make our lives better (and these beliefs are confirmed by experience). However, the author's intention was not to help us less-disciplined readers make changes that are currently hard for us. Instead, she concluded that we should let go and stop trying to become like those people if we do not already have that same desire and enjoyment.

While this was freeing in a way, I couldn't help wondering if there could be some middle ground between accepting my natural tendencies and finding ways to amend my beliefs if they are unproductive or invalid. There is really no reason to force myself to adopt a habit that contradicts my worldview, but I hope that if I felt I had a good reason to change a belief, my implicit bias would not thwart this endeavor. It seems that some level of subliminal self-

awareness is necessary for striking this balance and achieving personal growth that goes beyond the superficial. Helping people find this self-awareness is a large part of a therapist's job. There are no clear-cut answers for changing a harmful belief, but the practice of asking probing questions to encourage self-examination can help people understand why they think and act the way they do (Kim). If I can ask myself why I believe certain things, finding ways to change may come more naturally to me. Even if change doesn't happen right away, therapists can help people see which of their beliefs are worth continued effort, and which they can stop fighting. The key seems to be attaining as much awareness as I can about how my mindset influences me.

However, the question of whether, and to what extent, we are aware of our implicit mindsets and biases is a little more layered than the way I have presented it so far. Philosophers and psychological theorists note that there are several possible ways in which we could be oblivious to our own mindsets: for example, people may understand the content of their own prejudice, but be in the dark about its source or the way it influences their behavior (Brownstein). It is not clear which aspects of bias are most hidden from us, or whether we are truly unaware at all. Psychologist Bertram Gawronski posits that while many aspects of our implicit biases may escape our notice, people usually demonstrate awareness of the content of their own biases when pressed. In several experiments, people show they can accurately report their implicit biases when asked directly, especially if they believe there will be some consequence for inaccurate self-reporting (Gawronski 575). Thus, I may have much greater access to my subliminal mental states than I realize.

Results like these are certainly encouraging for me since I have developed a borderline obsession with self-awareness. I am both fascinated and terrified by books like *You Are Not So Smart* by David Mcraney that highlight the ways people's brains deceive them. However, if people are completely capable of uncovering their own implicit biases, this brings up an uncomfortable question: Is my subconscious bias completely willful? Does it remain in the background of my mind only because I deliberately choose not to confront it? Maybe sometimes we choose self-deception as a means of avoiding an unpleasant truth or convincing ourselves of something we wish were true.

The ways that we think and perceive the world are all part of our minds' efforts to maintain a sense of self. Traditionally, many psychologists do not consider introspection by experimental subjects to be a helpful measure of their true beliefs. While earnest self-

examination may be effective in uncovering bias, people often resist sincere introspection because the dissonance resulting from a discovery of implicit bias can be very distressing (Alicke 17). It was certainly upsetting for me to discover the possibility of bias in myself when I took a class with Professor Anderson. Perhaps I would have noticed this tendency sooner if it was not so important to me to think of myself as a feminist. Since our minds seem to be largely focused on maintaining a consistent and meaningful personal identity, our experiences and beliefs about ourselves play a huge role in shaping our view of the outside world. Our neurological responses to various situations are usually influenced more by our personal fears and desires than by the objective nature of those situations. For example, my brain responds the same way to both physical peril and perceived social threats that pose no real danger (Taylor 49). This implies that like other humans, I see anything that may make me feel bad about myself as a major threat because it can damage my sense of self.

I am accustomed to seeing self-centeredness simply as part of human nature, but are there any other explanations for the link between bias and our tendency to focus inward? Some philosophers hypothesize that the emergence of differences between the conscious and the subliminal can be traced to the development of language, especially words that allow us to express individuality (Taylor 77). If I couldn't express any of my abstract thoughts, maybe my belief system would be more unified.[11] Perhaps my conscious beliefs are the ones I articulate, either mentally or out loud, and my subliminal beliefs are the ones that remain intangible. Taking this further, Taylor argues that the word "I" created divisions in our consciousness by allowing people to think about themselves instead of only a group they are part of. While it is difficult to imagine not having the words to express my own thoughts or individuality, I certainly think I would take things less personally if I could only think about myself abstractly. The fact is, though, that for whatever reason, taking things personally comes naturally to most of us, and exposing a personal bias is one of the easiest ways to make someone feel threatened.

Sifting through our implicit biases is a deeply personal and sometimes painful process, and many people find it difficult enough to change their biases that they opt for ignorance rather than coping with cognitive dissonance. How do I decide when confronting my bias is worth the

[11] While deep in an Internet rabbit hole, I recently read that some people have an inner monologue that articulates their thoughts, while some people only think abstractly. Many people in both groups were very surprised to learn about the other group. While beyond the scope of this essay, it would be very interesting to see how the presence or lack of an internal monologue affects a person's sense of self.

energy? Are there times when I have a moral obligation to correct my implicit bias? I certainly want to avoid bias that causes negative outcomes such as discrimination. Aside from tangible consequences, though, some wonder if, for example, it is morally wrong to harbor subconscious racist views, even if we are unaware of them (Kelly 527). This is certainly an interesting question for me – it can be discouraging to have a desire to be a good person but feel that I am coming up short because of bias that I cannot fully see.

Philosophers Kelly and Roedder suggest we begin addressing this question by asking if a certain implicit belief would be an acceptable explicit belief, even if we never acted on it. Most people would probably condemn any vehemently racist person, even if they kept their racism to themselves (Kelly 527). Even though I never disrespected Professor Anderson, I do not think it would be defensible if I was constantly fuming about being taught by an inferior woman. Kelly and Roedder note, however, that there is still much philosophical work to be done before we understand how comparable implicit and explicit beliefs really are.

Despite this uncertainty about the ethics of beliefs, they argue that if we become aware of a bias that is present in the average person, we should try to correct for it in ourselves. Refusing to do so would imply I think I am better than the average person, and this attitude is hardly conducive to the personal growth I want to achieve. I have come to believe that if someone is reluctant to address bias in themselves, this says more about them than the bias itself does. Discovering the possibility of latent sexism in myself almost felt like an attack, but I see now that it was silly for me to feel this way. If I learned that humans are biased in tasks such as estimating the size of circles, I would simply correct for this without feeling bad about myself, and correcting bias towards race, gender, or sexual orientation need not be different (Kelly 533).

Considering all the ways we are wired to categorize people and rely on stereotypes, it seems unlikely I will find an easy way to rid myself of bias. However, I found some encouragement in the surprisingly optimistic conclusion of two philosophers: “While we may have evolved to distinguish between in-groups and out-groups, it is unlikely that we evolved to treat race, gender and ability as grouping criteria” (Brownstein and Saul 59). It appears that hidden bias is inevitable, but heinous moral consequences like systemic exclusion of certain groups may not be. Even if I am not fully aware of my bias, I think that simply becoming comfortable with the fact that I am biased is an important first step in fighting racism, sexism, and other harmful forms of bias. After all, it seems to me that I can attribute much of my bias to

the ease of thinking automatically rather than deliberately. I do not claim to know how I can achieve complete self-awareness and manage my bias perfectly, but choosing to take my mind off autopilot every once in a while will probably go a long way.

Looking back on my experience with Professor Anderson, the most productive question was not, “Am I a terrible person?” It is all too easy to connect my biases to my sense of identity, but this kind of thinking can only make it more difficult to discard my unsavory beliefs. A more constructive question to ask myself periodically may be, “What are the subconscious beliefs affecting my actions, and how could I try to act more fairly?” It has been difficult to admit I am susceptible to biases from my culture, but I realize that my bias really does not say anything about me other than I am human like everyone else. Similarly, there is no need to feel overly self-satisfied about my efforts to overcome bias; just as an implicit prejudice does not necessarily mean I am morally bankrupt, my habit of taking classes from female professors does not mean I am better or more “woke” than anyone. These only mean that I am on a lifelong journey to try to be more ethical, consistent, and self-aware.

While there is a lot of uncertainty about this journey, accepting my vulnerability to bias has made my experience with Professor Anderson feel a little less embarrassing. I do not think that having some implicit and explicit beliefs out of alignment made me a hypocrite. It seems the real danger lies not in bias itself, but in clinging to a certain idea of myself while refusing to acknowledge the times I fall short of this ideal. Acknowledging my bias is uncomfortable, but I would like to think this acknowledgement has made me a better person. My next step is to embrace the challenge of finding ways to grow from this discomfort and use it become closer to the best version of myself.

Bibliography

Alicke, Mark D., et al., editors. *The Self in Social Judgment*. Psychology Press, 2014.

Amit, Ron, and Ron Meir. "Lifelong Learning and Inductive Bias." *Current Opinion in Behavioral Sciences*, vol. 29, 2019, pp. 51–54. *Elsevier*, doi:10.1016/j.cobeha.2019.04.003.

Apatow, Judd, director. created by Paul Feig, performance by Linda Cardellini, and James Franco, NBC, 1999.

Armisen, Fred. *Portlandia*, created by Carrie Brownstein, and Jonathan Krisel, IFC, 2011.

Arrieta, Alejandro Barredo, et al. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI." *Information Fusion*, vol. 58, 2020, pp. 82–115., doi:10.1016/j.inffus.2019.12.012.

Brownstein, Michael. "Implicit Bias." *Stanford Encyclopedia of Philosophy*, Edited by Edward N. Zalta, 2019, https://plato.stanford.edu/archives/fall2019/entries/implicit-bias/.

Brownstein, Michael, and Jennifer Saul, editors. *Implicit Bias and Philosophy*. Oxford University Press, 2016.

Carlson, Richard. *Don't Sweat the Small Stuff for Teens: Simple Ways to Keep Your Cool in Stressful Times*. Hyperion, 2000.

Clegg, Isabella. "Cognitive Bias in Zoo Animals: An Optimistic Outlook for Welfare Assessment." *Animals*, vol. 8, no. 7, 2018, p. 104., doi:10.3390/ani8070104.

Conway, Michael. "The Problem With History Classes." *The Atlantic*, Atlantic Media Company, 16 Mar. 2015, www.theatlantic.com/education/archive/2015/03/the-problem-with-history-classes/387823/.

Desmond-Harris, Jenee. "Implicit Bias Means We're All Probably at Least a Little Bit Racist." *Vox*, Vox Media, 15 Aug. 2016, www.vox.com/2014/12/26/7443979/racism-implicit-racial-bias.

Egan, Andy. “Comments on Gendler’s, ‘the Epistemic Costs of Implicit Bias.’” *Philosophical Studies*, vol. 156, no. 1, 2011, pp. 65–79., doi:10.1007/s11098-011-9803-5.

Gawronski, Bertram. “Six Lessons for a Cogent Science of Implicit Bias and Its Criticism.” *Perspectives on Psychological Science*, vol. 14, no. 4, 2019, pp. 574–595., doi:10.1177/1745691619826015.

Grinberg, Emanuella. “4 Ways You Might Display Hidden Bias Every Day.” *CNN*, Cable News Network, 25 Nov. 2015, www.cnn.com/2015/11/24/living/implicit-bias-tests-feat/index.html.

Gurney, Deirdre, and Scott Gurney. “Duck Dynasty.” *Duck Dynasty*, performance by Willie Robertson, A&E, 2012.

Haselton, Martie G., et al. “The Evolution of Cognitive Bias.” *The Handbook of Evolutionary Psychology*, 2015, pp. 1–20., doi:10.1002/9781119125563.evpsych241.

Hatcher, Donald L. *Understanding "The Second Sex."*. Peter Lang, 1984.

Henderson, Leah. “The Problem of Induction.” *Stanford Encyclopedia of Philosophy*, Edited by Edward N. Zalta, 2019, https://plato.stanford.edu/archives/win2019/entries/induction-problem/.

Howard, Byron and Rich Moore, directors. *Zootopia*. Performance by Ginnifer Goodwin, and Jason Bateman, Walt Disney Animation Studios, 2016.

“Inductive Reasoning & Being Wrong.” *News Frames*, 28 Jan. 2014, newsframes.wordpress.com/2013/02/14/inductive-reasoning/.

Jabr, Ferris. “How Beauty Is Making Scientists Rethink Evolution.” *New York Times Magazine*, 9 Jan. 2019, www.nytimes.com/2019/01/09/magazine/beauty-evolution-animal.html.

Kelly, Daniel, and Erica Roedder. “Racial Cognition and the Ethics of Implicit Bias.” *Philosophy Compass*, vol. 3, no. 3, 2008, pp. 522–540., doi:10.1111/j.1747-9991.2008.00138.x.

Kim, John. “How to Change Your False Beliefs.” *Psychology Today*, Sussex Publishers, 20 June 2017, www.psychologytoday.com/us/blog/the-angry-therapist/201706/how-change-your-false-beliefs.

Kranz, Michal. “5 Genocides That Are Still Going on Today.” *Business Insider*, Business Insider, 22 Nov. 2017, www.businessinsider.com/genocides-still-going-on-today-bosnia-2017-11.

Marczyk, Jesse. “The Beautiful People.” *Psychology Today*, Sussex Publishers, 20 Apr. 2018, www.psychologytoday.com/us/blog/pop-psych/201804/the-beautiful-people.

Michaels, Lorne, and Tina Fey. *Mean Girls*. Performance by Rachel McAdams, and Lindsay Lohan, Paramount Home Entertainment, 2005.

Montañez, George D., et al. “The Futility of Bias-Free Learning and Search.” *AI 2019: Advances in Artificial Intelligence Lecture Notes in Computer Science*, 2019, pp. 277–288., doi:10.1007/978-3-030-35288-2_23.

Mooney, Raymond J. “Comparative Experiments on Disambiguating Word Senses: An Illustration of the Role of Bias in Machine Learning.” *Conference on Empirical Methods in Natural Language Processing*, 9 Dec. 1996, arxiv.org/abs/cmp-lg/9612001.

Nelson, Julie A. “The Power of Stereotyping and Confirmation Bias to Overwhelm Accurate Assessment: the Case of Economics, Gender, and Risk Aversion.” *Journal of Economic Methodology*, vol. 21, no. 3, 2014, pp. 211–231., doi:10.1080/1350178x.2014.939691.

Payne, B. Keith, and Bertram Gawronski. *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*. Guilford Press, 2010. *ResearchGate*.

Platt, Marc, and Ric Kidney. *Legally Blonde*. Performance by Reese Witherspoon, MGM Home Entertainment, 2002.

Rich, Alexander. “Using Inductive Bias as a Guide for Effective Machine Learning Prototyping.” *Medium*, Flatiron Engineering, 6 Nov. 2019, medium.com/flatiron-

engineering/using-inductive-bias-as-a-guide-for-effective-machine-learning-prototyping-66e5468407a8.

Schwitzgebel, Eric. “Belief.” *Stanford Encyclopedia of Philosophy*, Edited by Edward N. Zalta, 2019, plato.stanford.edu/archives/fall2019/entries/belief/.

Taniguchi, Hidetaka, et al. “A Machine Learning Model with Human Cognitive Biases Capable of Learning from Small and Biased Datasets.” *Scientific Reports*, vol. 8, no. 1, 2018, doi:10.1038/s41598-018-25679-z.

Taylor, Eldon. *Subliminal Learning: An Eclectic Approach*. R.K. Book, 1992.

White, Allan P., and Wei Zhong Liu. “Bias in Information-Based Measures in Decision Tree Induction.” *Machine Learning*, vol. 15, no. 3, 1994, pp. 321–329., doi:10.1007/bf009933

Zebrowitz, Leslie A. *Social Perception*. Open University Press, 1990.