LRP Week 14-1 Fallacies and Biases

1 | Mistakes in Deductive and Inductive Reasoning

This class has been about *deductive* and *inductive* reasoning.

1. Deductive reasoning is about *truth-preservation*:
2. Inductive reasoning is about *probability*, or *likelihood*.

In both cases, we can (and do) make mistakes in reasoning all the time. The classic mistake in each type:

1. In deductive reasoning, the classic mistake is: *thinking an argument is truth-preserving when it isn't.*
2. In inductive reasoning, the classic mistake is *thinking that some evidence supports a hypothesis more than it actually does (or less than it actually does).*

People often call the former mistake a "**fallacy**" and the latter mistake a "**bias**."

2 | Some Cognitive Biases

I don't find *formal* deductive fallacies all that interesting. Most of the time we fall for them because there is actually some *context* that means the premises are *good evidence* for the conclusion, and at least make the conclusion reasonable, even if the argument is not not "truth-preserving". In everyday reasoning, biases in inductive reasoning are much more consequential.

2.1 | *Base-Rate Neglect*

Consider the following case:

> Tom is an American man in his mid-30s. He is great with kids, loves storytelling, and grew up reading classics of literature. Is Tom more likely to be a librarian or a farmer?

Base-Rate Neglect is when you take *new* information very seriously, while neglecting what reasonable *old* probabilities might be. That is, you take your *new evidence* too seriously, and you take your *prior* not seriously enough.

2.2 | *Belief Perseverance*

Belief Perseverance is the *opposite* of base-rate neglect. It's when you take *old* information very seriously, while underestimating (or even ignoring outright) new information. That is, you take your *new evidence* not seriously enough, and you take your *prior* too seriously. This can crop up in neutral cases like the following:

> (Edwards, 1968) There are two bookbags, one containing 700 red and 300 blue chips, the other containing 300 red and 700 blue. Take one of the bags. Now, you sample, randomly, with replacement after each chip. In 12 samples, you get 8 reds and 4 blues. what the approximate probability that this is the predominantly red bag?
>
> $\approx 25\%$ $\approx 55\%$ $\approx 75\%$ $\approx 95\%$

But belief perseverance can be even more powerful when changing our beliefs is *uncomfortable*. Learning something new might create **cognitive dissonance**: incompatibilities between our beliefs and values. This is not fun, and gives us incentive to avoid new information.

2.3 | Hindsight Bias

Hindsight Bias is the tendency to overestimate how similar your old opinions were to your opinions now. It's the "I knew the whole time!" Bias.

1. Example: it's hard for professors to explain things to people who don't yet understand them, because it's hard for the professors to remember what it's like to not know it.
2. Important for "should have known". Example: medical malpractice suits and determination of negligence (Berlin 2000).

2.4 | Confirmation Bias

Confirmation Bias is a bias wherein you tend to search for, remember, or interpret information in ways that *support*, rather than *undermine*, your current opinions.

→ Example: The Stanford news articles study (Lord et al 1979)

2.4.1 Is Confirmation Bias Reasonable?

Suppose you have a hypothesis H that you think is likely to be true (or at least, you hope to be true), like "my crush likes me". You want to find out more about whether or not your hypothesis is true, but you don't want to just ask them. So you could:

1. Search for evidence that they like you.
2. Search for evidence that they *don't* like you.
3. Both

Whichever option you choose, you can go out and get evidence. Suppose sometimes they act flirty toward you, other times they treat you more as a friend. They ask you out for drinks, but the next day seem a bit aloof. If you're just searching for evidence that they like you, you will pay more attention to the evidence *in favor* of that than against. And if you update on your evidence, you'll increase your confidence that they like you. And likewise the other way.

Question: Do you think you should be required to seek out evidence in a "neutral" way, seeking equally evidence for and against hypotheses? How would you tell if you're seeking evidence "equally"?

Question: Is there a way to search for evidence in both directions equally, but *still* have confirmation bias?

2.5 | *Self-fulfilling Prophecies*

Confirmation bias can also lead to self-fulfilling beliefs:

1. Connie is confident that her crush likes her, and gathers evidence in favor of that hypothesis. She becomes sure enough, and then asks her crush out. The crush was in fact on the fence, but is charmed by her forwardness and agrees.
2. Una is unconfident that her crush likes her, and gathers evidence in favor of the hypothesis that her crush does not like her. She figures that there's no point trying to do anything, and so nothing happens between the two of them.
3. (Also: the policing-rates question from the problem set!)

3 | Avoiding Bias?

3.1 | *The Bias Blindspot*

We might appreciate that cognitive biases happen, and that cognitive biases can happen *to us*. But we generally think it isn't happening *right now*:

> **The Bias Blindspot**: we generally think that we are not *currently* subject to a cognitive bias.

It's *really* hard to tell when you're subject to a bias, for at least two reasons:

1. It's tempting to think that biases would be *transparent* to us. That when we are subject to a bias, we sort of know it, and could stop if we wanted to; i.e., that we're turning a blind eye on purpose to our biases. But in fact, the processes that cause biases are often hidden away from us. Mandelbaum (2019) calls this the "Psychological Immune System," helping us keep our current beliefs, keeping us safe from potentially unsettling or paradigm-shifting information.
2. It is reasonable to think that you've already corrected for your biases (even if that is not true). After all, if you have some opinion, you must in some sense think it is the *best* opinion to have, given your evidence. If you thought your opinion were biased in some way, you would correct the bias. In that sense, self-confidence in your own opinions is naturally *baked in*. And it sort of has to be: we couldn't make it through the day if we were constantly doubting every one of our opinions.

3.2 | *Being Open to Belief Revision*

We've been talking about the process of belief revision in abstract terms:

$$\Pr_{old}(H) \quad \overset{\text{evidence } E}{\longrightarrow} \quad \Pr_{new}(H)$$

In *concrete terms*, it can be difficult! In a sense, updating your opinions means *admitting* that your old opinions were incorrect.

But what if your approach was wanting to *have the best opinions, given your evidence?* Then when you get new evidence and change your opinions, your simply doing the most responsible thing given your new evidence. Yes, it means you now think your old opinions were worse. But it doesn't necessarily mean you were a bad reasoner back then: you may just just not have had the proper evidence!

Is there a way to see through the bias blindspot? Here is a strategy (Lord 1984): Suppose you have some opinion about a hypothesis H: you think it's likely true, or you think it's likely false. Then you learn some evidence E. Now *consider the opposite:*

1. Imagine that you had the *opposite* opinion about hypothesis H. Then how would you react to evidence E?
2. Imagine that you got the opposite evidence: *not* E (but you have the same opinions). Then how would you react to the evidence "*not* E"?

Exercise: Consider the Opposite

Consider your *latest* position on the question "Can (the latest) LLMs reason?". Do you it's *likely* or *unlikely* that the answer is "Yes"? **Note your answer.**

If you said "Likely":

1. Suppose you learned evidence E_1: "the latest large reasoning model from OpenAI was able to generate complicated proofs of mathematical theorems proven only by mathematicians in the past decade." How would this change your opinions in whether LLMs can reason?
2. Suppose you learned E_2: "the latest large reasoning model from OpenAI frequently gives contradictory answers to basic probability questions, depending on how the question is framed." How would this change your opinions in whether LLMs can reason?
3. Pretend that you learned E_1, but you previously thought that it was *unlikely* that LLMs could reason. How would this have changed your opinions in whether LLMs can reason?

If you said "Unlikely":

1. Suppose you learned E_2: "the latest large reasoning model from OpenAI frequently gives contradictory answers to basic probability questions, depending on how the question is framed." How would this change your opinions in whether LLMs can reason?
2. Suppose you learned evidence E_1: "the latest large reasoning model from OpenAI was able to generate complicated proofs of mathematical theorems proven only by mathematicians in the past decade." How would this change your opinions in whether LLMs can reason?
3. Pretend that you learned E_2, but you previously thought that it was *likely* that LLMs could reason. How would this have changed your opinions in whether LLMs can reason?

Considering the opposite helps you figure out how much your current opinion is biasing how you're responding to new information, by testing

1. **Same Prior Opinion, Different Info**: Whether you're responding in a biased way to the information you're learning compared to if you had learned information that pointed the *other* direction
2. **Same Info, Different Prior Opinion**: Whether you're responding in a biased way to the information you're learning compared to if you had learned the same information but had had a different current opinion.