Cognitive Biases and Their Importance for Critical Thinking

CONTENTS

Critical Thinker Academy

Introduction

In my five-dimensional model of the most important components of critical thinking I put "knowledge of the psychology of human judgment" under the heading of *background knowledge*:

1. logic
2. argumentation
3. rhetoric
4. background knowledge
    a. *of subject matter*
    b. *of the history of debate on a topic*
    c. ***of the psychology of human judgment***
5. character (attitudes and values)

Knowledge of cognitive biases is a sub-topic that fits within the broader category of the psychology of human judgment.

There has been an explosion of research on cognitive biases over the past forty years, but the critical thinking education community has been slow to integrate this content into the traditional curriculum of critical thinking textbooks. There is no doubt, however, that the study of cognitive biases will be a central feature of critical thinking education in the 21st century.

1. Cognitive Biases: What They Are and Why They’re Important

Everyone agrees that logic and argumentation are important for critical thinking. One of the things that I try to emphasize in my “five pillars” model of critical thinking is that *background knowledge* is also very important.

There are different types of background knowledge that are relevant to critical thinking in different ways. One of the most important types of background knowledge is **knowledge of how our minds actually work** — how human beings actually think and reason, how we actually form beliefs, how we actually make decisions.

There are a lot different scientific fields that study how our minds actually work. These include *behavioral psychology*, *social psychology*, *cognitive psychology*, *cognitive neuroscience*, and other fields. Over the past forty years we’ve learned an awful lot about human reasoning and decision making.

A lot of this research was stimulated by the work of two important researchers, Daniel Kahneman and Amos Tversky, going back to the early 1970s. They laid the foundations for what is now called the “biases and heuristics” tradition in psychology[1].

i. Normative vs. Descriptive Theories of Human Reasoning

To get a feel for the importance of this research, let’s back up a bit. When studying human reasoning you can ask two sorts of question. One is a purely *description* question — how DO human beings IN FACT reason? The other is a *prescriptive* or *normative* question — how SHOULD human beings reason? What’s the difference between *good* reasoning and *bad* reasoning?

When we study logic and argumentation, we’re learning a set of rules and concepts that attempt to answer this *second* question — how *should* we

[1] For a good review of the history of this research, see *Thinking, Fast and Slow*, by Daniel Kahneman (2011, Farrar, Straus and Giroux).

reason. These are elements of a broader normative theory of rationality, of what it means to reason well. Actually what we have is not a one big theory of rationality, but *a bunch of more narrowly focused theories that target reasoning in specific domains or under specific conditions*.

So for example when we're interested in *truth-preserving inferences*, we appeal to **deductive logic** and interpret the rules of deductive logic as norms for reasoning in this area.

When we're interested in reasoning about *chance and uncertainty*, we appeal to **probability theory and statistics** to give us norms for correct reasoning in this area.

If we're talking about *making choices that will maximize certain goals under conditions of uncertainty* then we appeal to **formal decision theory**. If we add the situation where other actors are making choices that can effect the outcomes of our decisions, then we're moving into what's called **game theory**.

ii. The Gap Between How We SHOULD Reason and How DO Reason

So, over time, we've developed a number of different theories of rationality that give us norms for correct reasoning in different domains.

This is great of course, these are very powerful and useful tools.

Now, when it comes to the study of how human reasoning *actually* works, before Kahneman and Tversky's work in the 1970s, there was a widely shared view that, more often than not, the mind, or the brain, processes information in ways that mimic the formal models of reasoning and decision making that were familiar from our normative models of reasoning, from formal logic, probability theory and decision theory.

What Kahneman and Tversky showed is that, more often than not, this is NOT the way our minds work — they showed that **there's a gap between how our normative theories say we *should* reason, and how we *in fact* reason.**

This gap can manifest itself in different ways, and there's no one single explanation for it. One reason, for example, is that in real-world situations, the reasoning processes prescribed by our normative theories of rationality can be computationally very intensive. Our brains would need to process an awful lot of information to implement our best normative theories of reasoning. But that kind of information-processing takes time, and in the real world we often need to make decisions much quicker, sometimes in milli-seconds. You can imagine this time pressure being even more intense if you think about the situations facing our *homo sapien* ancestors; if there's a big animal charging you and you wait too long to figure out what to do, you're dead.

iii. Biases and Heuristics (Rules of Thumb)

So, the speculation is that our brains have evolved various *short-cut mechanisms for making decisions*, especially when the problems we're facing are complex, we have incomplete information, and there's risk involved. In these situations we sample the information available to us, we focus on just those bits that are most relevant to our decision task, and we make a decision based on a rule of thumb, or a shortcut, that does the job.

These rules of thumb are the "heuristics" in the "biases and heuristics" literature.

Two important things to note: One is that we're usually not consciously aware of the heuristics that we're using, or the information that we're focusing on. *Most of this is going on below the surface.*

The second thing to note is these heuristics *aren't* designed to give us the *best* solutions to our decision problems, all things considered. What they're designed to do is give us solutions that are "good enough" for our immediate purposes.

But "good enough" might mean "good enough in our ancestral environments where these cognitive mechanisms evolved". In contexts that are more removed from those ancestral environments, we can end up making systematically bad choices or errors in reasoning, because we're automatically, subconsciously invoking the heuristic in a situation where that heuristic isn't necessarily the best rule to follow.

So, the term "bias" in this context refers to this systematic *gap* between how we're *actually disposed* to behave or reason, and how we *ought* to behave or reason, by the standards of some normative theory of reasoning or decision making. **The "heuristic" is the *rule of thumb* that we're using to make the decision or the judgment; the "bias" is the *predictable effect* of *using* that rule of thumb in situations where it doesn't give an optimal result.**

iv. An Example: The Anchoring Effect

This is all pretty general, so let me give you an example of a cognitive bias and its related heuristic. This is known as the **anchoring heuristic**, or the **anchoring effect**.

Kahneman and Tversky did a famous experiment in the early 1970s where they asked a group of subjects to estimate the percentage of countries in Africa that are members of the United Nations. Of course most aren't going to know this, for most of us this is just going to be a guess.

But for one group of subjects, they asked the question "Is this percentage more or less than 10%?". For another group of subjects, they asked the question "Is it more or less than 65%?"

The average of the answers of the two groups differed significantly. In the first group, the average answer was 25%. In the second group, the average answer was 45%. The second group estimated higher than the first group.

Why? Well, this is what seems to be going on. If subjects are exposed to a higher number, their estimates were "anchored" to that number. Give them a high number, they estimate higher; give them a lower number, they estimate lower.

So, the idea behind this anchoring heuristic is that when people are asked to estimate a probability or an uncertain number, rather than try to perform a complex calculation in their heads, they start with an implicitly suggested reference point — the anchor — and make adjustments from that reference point to reach their estimate. This is a shortcut, it's a rule of thumb.

Now, you might think in this case, it's not just the number, it's the way the question is phrased that biases the estimates. The subjects are assuming that the researchers know the answer and that the reference number is therefore related in some way to the actual answer. But researchers have redone this experiment many times in different ways.

In one version, for example, the subjects are asked the same question, to estimate the percentage of African nations in the UN, but before they answer, the researcher *spins a roulette wheel in front of the group*, wait for it to land on a number so they can all see the number, and then ask them if the percentage of African nations is larger or smaller than the number on the wheel.

The results are the same. If the number is high people estimate how, if the number is low people estimate low. And in this case the subjects couldn't possibly assume that the number on the roulette wheel had any relation to the actual percentage of African nations in the UN. But their estimates where anchored to this number anyway.

Results like these have proven to be really important for understanding how human beings process information and make judgments on the basis of information. The anchoring effect shows up in strategic negotiation behavior, consumer shopping behavior, in the behavior of stock and real estate markets — it shows up *everywhere*, it's a very widespread and robust effect.

Note, also, that this behavior is, by the standards of our normative theories of correct reasoning, *systematically irrational*.

v. Why This is Important

So this is an example of a cognitive bias. Now, this would be interesting but not deeply significant if the anchoring effect was the only cognitive bias that we've discovered. But if you go to the wikipedia entry under "list of cognitive biases" you'll find a page that lists just *over a hundred* of these biases, and the list is not exhaustive. I encourage everyone to check it out.

So what's the upshot of all this for us as critical thinkers?

At the very least, we all need to acquire a certain level of cognitive bias "literacy". We don't need to become experts, but *we should all be able to recognize the most important and most discussed cognitive biases*. We should all know what "confirmation bias" is, what "base-rate bias" is, what the "gambler's fallacy" is, and so on. *These are just as important as understanding the standard logical fallacies*.

Why? Because as critical thinkers, we need to be aware of the processes that influence our judgments, especially if those processes systematically bias us in ways that make us prone to error and bad decisions.

Also, we want to be on the lookout for *conscious manipulation and exploitation of these biases* by people who are in the influence business, whose job it is to make people think and act in ways that further their interests, rather than your interests. We know that marketing firms and political campaigns hire experts in these areas to help them craft their messages.

vi. A Hypothetical Example: How to Use the Anchoring Effect to Manipulate Public Opinion

Let me give you a hypothetical example (though I know some people who'd say it's not hypothetical). Let's say you're a media adviser to a government that has just conducted a major military strike on a foreign country, and there were civilian casualties resulting from the strike.

Now, if the number of civilians killed is high, then that's bad for the government, it'll be harder to maintain popular support for this action. Let's say our intelligence indicates that the number of casualties is in the thousands. This is not a good number, it's going to be hard to sell this action if that's the number that everyone reads in the news the next day.

So, as an adviser to this government, what do you recommend? I'll tell you what I'd do, if all I cared about was furthering the government's interests.

I'd say "Mr. President, " (or whoever is in charge) "we need to issue a statement before the press gets a hold of this. In this statement, we need to say that the number of estimated casualties resulting from this strike is low, *maybe a hundred or two hundred at the most*."

Now, why would I advise this? *Because I know about the anchoring effect*. I know that the public's estimate of the real number of casualties is going to be anchored to the first number they're exposed to, and if that number is low, they'll estimate low. And even if data eventually comes out with numbers that are higher, the public's estimates will still be lower than they would be if we didn't get in there first and feed them that low number.

That's what I would do, if all I cared about was manipulating public opinion.

Now, this is a hypothetical example, but trust me when I tell you that decisions like these are made every day, under the advice of professionals who are experts in psychological literature.

So there's a broader issue at stake. **This is the kind of background knowledge that is important if our ultimate goal is to be able to claim ownership and responsibility for our own beliefs and values.**

2. Cognitive Biases and the Authority of Science

In this lecture I want to look at another reason for learning about cognitive biases. This reason has to do with a more theoretical question: **What is science?** And more specifically: **What is the essence of modern scientific methodology?**

This is an old question. It's a philosophical question, it's a question about the basis and rationale of the scientific method.

It's also a contentious question. People disagree about what science is and what its methods are. If you're familiar with the philosophy of science literature, or the science studies literature, then you know that there's a lot of debate about these questions. And there's debate about a whole family of related questions, like "Is science objective?", and "What reason do we have to think that science gives us a more accurate picture of the world than any other worldview or belief system?" and "Is there even any such thing as THE scientific method?".

Now, I don't presume to settle these big questions here. At some point in the future I'll do a whole course in the Critical Thinker Academy on philosophy of science issues, but my goals for this lecture are much less ambitious.

What I want to do is focus on that last question, "is there such a thing as THE scientific method"? And I want to talk about the relevance of cognitive biases for this question, and what this means for the authority of science.

i. Is There a Scientific Method?

Is there such a thing as the scientific method? Scientists seem to think so, or least, science textbooks seem to think so. If you open a high school or university science textbook you'll often see a first chapter that discusses the nature of science and scientific reasoning. You might even see little flowcharts that are supposed to encapsulate the scientific method. Just google "the scientific method" and look at the results under "images", you'll see what I mean.

But if you look at the history and philosophy of science literature, and if you look across the full range of natural and social sciences, the picture is very different. What you see is *a variety of scientific methods used within specific fields, but no single method that is common to all the sciences*. And you also see (not always but often) a considerable gap between the idealized textbook descriptions of scientific methodology and the actual processes, the actual histories that lead up to the adoption or rejection of a theory within a scientific community.

These observations lead some people to talk about "the scientific method" as a kind of institutionalized *myth*, a story we tell ourselves about how science works, but that doesn't really correspond to the reality. And it leads some to challenge the objectivity and authority of science, to challenge the presumption, so common in the developed West, that scientific knowledge is superior to other forms of knowledge, that the modern scientific worldview gives our best truest description of the workings of the natural world.

For those who want to defend the superiority of scientific knowledge, the philosophical challenge is how to argue for this view without defaulting to the naive, mythic view of science and scientific reasoning that has been so thoroughly debunked by scholars.

I admit that as a philosopher of science I'm in this camp, this is a challenge for me. I cringe when I hear pro-science types make broad pronouncements about how science works. "Science is about the search for universal laws!" "Science is about about causal explanations!" "Science is about testing hypotheses that make falsifiable predictions!". You can find exceptions to all of these without looking very hard.

Yet I still think that there's something special about scientific knowledge, something distinctive that gives it a privileged status. I don't think this is a trivial position to defend, I think the criticisms of the traditional view of science need to be taken seriously.

What I do think — and this gets us to the main point of this lecture, finally — what I do think is that **an understanding of cognitive biases can**

give us a useful perspective on this question, and can help us think about why science is important and why scientific methods are what they are.

ii. Science: A Tool for Reducing the Systematic Errors Caused By Cognitive Biases

Okay, with all that build-up you'd expect that what I have to say about science and cognitive biases is going to be really deep and sophisticated, but it's not, it's really pretty simple. It can be summarized in two points:

1. **Human beings are prone to biases that lead to error.**
2. **Scientific methodology aims to neutralize the effects of these biases, and thereby reduce error.**

Let's look at these two points in order.

First, human beings are prone to biases that lead to error. What sorts of biases? There are lots, but I'm thinking of two kinds of biases in particular.

(1) We are pattern recognition machines

The first is related to the fact that we're **pattern recognition machines**. Human beings see patterns everywhere. We're bombarded by sensory data and our natural disposition is to seek out patterns in the data and attribute meaning to those patterns.

An obvious example of this is our tendency to see faces in anything that remotely resembles the configuration of features on a human face. An electrical outlet looks like a face, the front grill of a car looks like a face, the knots of a tree look like a face.

Now, it's not surprising that we see faces everywhere. Face recognition is important for social primates like us, it's very important that we learn to read the facial cues of other people to know whether they're friendly or a threat, whether they're content or displeased, and so on. So we've evolved this very specialized and sensitive mechanism for facial recognition, and as a result we're basically hard-wired for it.

But this means that we're prone to a certain kind of error as well, where *we see faces in inanimate things that don't really have faces.*

This is just one kind of pattern recognition mechanism working in our brains, there are lots of different kinds working all the time, looking for patterns that might be meaningful, and imposing meaning on those patterns if they trigger the right cues. Michael Shermer calls this feature of our mental functioning **patternicity**.

Now, in general, our ability to identify patterns is an extraordinarily valuable tool. It's essential for survival, for extracting meaningful, relevant information from our physical environment and from our social environment.

But there's a price we pay for this amazing ability — sometimes we attribute meaning to patterns which they don't have, and sometimes we see patterns in what is actually patternless noise. And it can be hard to not see these patterns when we expect to see them; our expectation, or the subconscious workings of our brain, can sometimes impose a meaningful pattern on data, or on sensory experience, even when there's nothing there.

Psychologists have a bunch of names for this phenomenon. For visual images, like seeing faces in clouds, or the Virgin Mary on a slice of burnt toast, it's called **pareidolia**. The general phenomenon is called **apophenia**, which is defined as the experience of seeing meaningful patterns or connections in random or meaningless data.

Here's a fun example.

Have you ever heard of **backmasking**? This is when you play an audio recording backwards, usually these are songs, and you hear meaningful speech when it's played backwards.

When you do this deliberately to an audio recording it's called "backmasking" or "backward masking". It was controversial in the 1980s when a number of Christian groups in the United States claimed that heavy metal rock bands were intentionally putting Satanic messages into their songs which people could hear if they played the songs backwards.

What's interesting about backmasking is that people will often hear meaningful words or phrases in songs that are played backward, even when they were never intentionally put there. And the effect is amplified when you're given some cues about what to look for in the audio.

The very best place on the internet to experience this effect is Jeff Milner's Backmasking site, at jeffmilner.com/backmasking.

On the video and audio versions of this lecture I play a sample from Jeff's site to illustrate the effect, but I can't do that in a document. You can check Jeff's site for yourself, but what I did is play a sample from the Britney Spear's song "Him Me Baby One More Time", first forwards and then backwards. The backwards version sounds like incoherent noise to most ears. But when you prime the listener's expectations by showing them this set of lyrics:

"Sleep with me, I'm not too young."

… and play the backwards sample again, it is impossible NOT to hear those words. It's a striking effect (again, I urge you to go check it out).

Now trust me, Britney Spears didn't intentionally backmask those lyrics. This is meaningless noise being interpreted as meaningful information, and we're all prone to it, and no amount of mental effort can make you not hear those lyrics.

So, what does this have to do with science? My general point is that we're hardwired to see patterns in data and in observable phenomena. Sometimes these patterns are meaningful and point to underlying regularities. If they're regular, repeatable regularities, then we've got the basis for a predictively useful scientific generalization. Our knowledge of the basic laws of nature begins with identifying these basic patterns in observable data.

But sometimes these patterns are not meaningful, sometimes they're just noise, but we mistake them for meaningful patterns. This is an error, but it can be hard to see the error, or correct for the error, especially if we're really invested in the reality of those patterns.

One thing that scientific methodology offers us is a set of tools for distinguishing meaningful from meaningless patterns. These tools come in different forms, but the obvious ones are the standard protocols for controlled experimental studies, and statistical analysis of data from these studies, the stuff you would learn in a good "research methods" class in a science department. I'm not going to elaborate on those methods here, but ideally that's one of the things you end up learning in those courses, how to distinguish meaningful from meaningless patterns.

So that's the first kind of error that we're prone to.

(2) We suck at weighing evidence

The second kind of error that we're prone to has to do with how human beings are naturally disposed to weigh evidence, and more specifically how we weigh evidence as it bears on the truth or falsity of a hypothesis.

The general error is this: **if we think the hypothesis is true, or would like it to be true, then we tend to remember or focus only on the evidence that would count in favor of the hypothesis, and ignore or dismiss evidence that would count against it.**

The result is that as we build up a set of data that is lopsided in favor of the hypothesis, that is biased toward confirmation. That's why this phenomenon is called **confirmation bias**, and there's a huge psychological literature on this.

Actually, "confirmation bias" can refer to a cluster of related phenomena. You can have:

1. **Biased search for information**: This is where people test hypotheses in a one-sided way, by only searching for evidence that is consistent with the hypothesis that they happen to hold. They ask, "what would you expect to see if this hypothesis was true?", and look for evidence to support this prediction, rather than ask

"what would you expect to see if this hypothesis was false?", and look for evidence that would falsify the hypothesis.

2. **Biased interpretation of evidence**: this is where you give two groups the SAME information, the SAME evidence, but they interpret the evidence differently depending on their prior beliefs. So if you strongly believe some hypothesis, then you'll be inclined to think, for example, that studies that support that hypothesis are well-conducted and convincing, but for studies that don't support your hypothesis, you'll be inclined to think that they're not well-conducted, and not convincing. One way to think of this is that *people set higher standards of evidence for hypotheses that go against their current expectations, and they have lower standards of evidence for those that support their expectations*.
3. **Biased memory**: Even if someone has gone ahead and tried to collect evidence in a neutral, unbiased way, they may still remember it selectively, so *you end up recalling more of the confirming evidence than the disconfirming evidence*, and it skews the evidence in favor of your expectations.

I want to emphasize that these biasing effects I'm describing are well documented in the cognitive science literature, it's part of the cognitive biases literature that goes back several decades now and it's an ongoing area of research.

(3) Science is what we do to keep us from lying to ourselves

Okay, so what's the upshot of this for our understanding of science? The upshot is that, left to our own devices, human beings are prone to error in the weighing of evidence. What's the error? The error is thinking that we're making a judgment based on a complete body of evidence, when the body of evidence we're considering has actually been filtered and skewed by confirmation bias. And that's going to lead to error in judgments about how well supported a hypothesis actually is.

And this is where scientific methodology comes in. The physicist Richard Feynman once said that “science is what we do to keep us from lying to ourselves”, and I think there’s a lot of truth to this. **One of the functions of scientific methodology is to neutralize the effects of confirmation bias by forcing us to search for and weigh a complete body of evidence, one that includes not only confirming evidence, but also disconfirming evidence, evidence that would count against the hypothesis in question.**

iii. Example: When is A correlated with B?

Let me give you a simple example to illustrate. Let’s say you walk into a health care clinic and you see a flyer for a psychotherapist’s practice. The flyer says that **Dr. Jones has a 90% success rate in treating mental health problems**. You read a little more closely, and it says that of all the patients he sees who come to his clinic complaining of psychological or mental health problems, 90% of them report an improvement in their condition within two weeks of beginning treatment with Dr. Jones.

Now, for the sake of argument, let’s assume this figure is accurate. If 100 people walk through his door suffering from some mental health problem, then if you survey them two weeks after beginning treatments with Dr. Jones, roughly 90 of them will say that their condition has improved.

Here’s the first question: **Does this evidence support the conclusion that Dr. Jones’s treatment is *causally responsible* for the improvement in their condition?**

I know from experience that if I ask my science students this question, about one third will say that it does support this causal hypothesis, and about two thirds will say “no”, it doesn’t, because they’ve been told many times in their psychology classes that correlation does not imply causation. Maybe there’s some other factor involved that’s responsible for the correlation, maybe its the placebo effect, whatever. I think a survey of the general public would give a much stronger result, with a lot more people thinking that this evidence supports the claim that Dr. Jones' treatments are causally responsible for the improvement in condition.

Now, what if I ask this question? Let's grant that the correlation doesn't necessarily imply causation. But does the evidence given even support the weaker claim of a correlation? **Does the fact that 90% of patients report improvement of condition support the claim that there's at least a correlation between Dr. Jones' treatments and the improvement in condition?** And by correlation I just mean that if you go to Dr. Jones' clinic with a mental health problem, you're statistically more likely to report improvement in your condition than if you didn't seek out treatment.

If I ask this question to my science students, *almost all of them will say that the evidence supports some kind of correlation*. When I ask them how strong a correlation, more than half will say that it's 90% or close to 90% — that you're 90% more likely to report an improvement in your condition if you go see Dr. Jones.

Now, for those of you playing along at home, what do you think the answer is? **Would you be surprised to hear that this evidence doesn't even support a claim of correlation? In fact, it gives us no reason to think there's any kind of correlation whatsoever, much less a 90% correlation!**

Why is this? *It's because we're only looking at confirming evidence, we have no information about potentially disconfirming evidence*. To establish a correlation between Dr. Jones' treatments and improvement in condition, we would need to compare *two numbers*: the probability that a person will improve if they seek out treatment with Dr. Jones, and *the probability that they will improve anyway, on their own, without seeking treatment from Dr. Jones.*

If it turned out, for example, that 90% of people will report improvement in their mental health problems within two weeks *anyway*, without seeking treatment, then your odds of improving are the same whether you see Dr. Jones or not, and the correlation is *zero*, there's no positive correlation at all.

Or maybe there's an 80% rate of improvement without treatment, so if the rate of improvement is 90%, then at best this would support a weak positive correlation of 10%, which is still very different from a 90% correlation.

So, this evidence in Dr. Jones' flyer, even if it's all 100% accurate, not only does it not support the hypothesis that his treatments are the cause of the improvement in his patients' condition, *it doesn't even support a correlation of any kind between his treatments and improvement in his patients' condition.* There just isn't enough information to justify these claims. But almost none of my science students, when they're given this hypothetical case, will see that this is the case; most will say that the evidence supports a very strong correlation. This because they're looking at an incomplete body of evidence and drawing a hasty inference, *but they don't realize until it's pointed out to them that the evidence is incomplete.*

Now, this is a very simple example, but it helps to illustrate the general take-away point, which is that **on our own we suck at weighing evidence**. And that highlights the function and the value of statistical methods and scientific protocols — **they force us to seek out and take into consideration types of evidence that are relevant to the truth of the hypothesis, but that we would otherwise ignore, or dismiss, or downplay, due to confirmation bias.**

Now, I've been focusing on confirmation bias here, because I think it's particularly important, but of course it's not the only source of bias in science, there are lots of them, but they just highlight the general point.

In sciences that deal with living subjects, for example, there are biases that can result from **expectation effects**, like the **placebo effect**, where the mere expectation of a treatment will elicit a physiological response from a test subject. You can correct for this bias by using blind, randomized control groups, where subjects don't know whether they're receiving a treatment or just a placebo, so you can estimate the size of the placebo effect and use that to estimate the effect size due to the treatment itself.

And there are also **experimenter effects**, where the experimenter's knowledge and expectations can unintentionally influence how subjects respond and how observations are interpreted. You neutralize these kinds of biases by using double-blind studies, where the researchers who directly interact with subjects and do data analysis don't know whether a given subject is in the test group or the control group.

The point is that all these protocols are in place because human beings are prone to error, and we need these protocols to neutralize or correct for these errors.

iv. What Makes Science Special

I'd like to wrap up this lecture with a final comment on the bigger question that motivated this topic in the first place. The question was whether there's a way of defending the superiority of science as a source of knowledge about the world, without resorting to a mythic view of how science works.

I think the answer is "yes", but it's a qualified "yes". I think it's clear that if we didn't follow these scientific protocols, our knowledge of the world would be less reliable than it is, and it's clear why this is so when you think of these protocols as methods for neutralizing the effects of cognitive biases.

But saying this doesn't mean that scientists always follow these protocols. Science is a complex social practice, and there are lots of things that can interfere with or prevent these protocols from being properly implemented. The highest quality studies are often the most expensive studies to conduct, so funding can be a limiting factor. The highest quality studies might also take many years, maybe even decades to conduct, so time constraints can be another factor.

And in some fields proper controlled studies might be just be impossible to conduct. In genetics, for example, there are experimental ways of measuring the heritability of a trait, which is the percentage of the variation in the trait that can be accounted for by genetic variation in the population. You can do these controlled experiments to directly measure heritability on fruit flies, but you can't do them on human populations because they would be unethical to conduct. So for humans we're forced to rely on more indirect methods of estimating heritability that are more limited and more vulnerable to biases.

So I admit that in some ways, this discussion of scientific methods still has an air of mythology about it, in the sense that it sets up an ideal that may never be perfectly realized in practice. But on the other hand, **we still retain a notion of what makes science special; namely, that it's an institutionalized social practice that is committed to these ideals, that strives to reduce the distorting effects of biases when it can. And in this respect it's distinctive, there's no other social institution that functions quite the same way.**

When I talk about science in this way, my students sometimes think that I'm defending a mythic view of science as "value-free", as though science when its properly conducted doesn't make make any subjective value judgments, or is free of value biases. But that doesn't follow. What this view entails is that scientific methodology is committed to the elimination of certain kinds of biases; namely, those that are conducive to error in the identification of meaningful patterns and in the weighing of evidence. But that's only part of what science is about. Science is very much a value-laden enterprise, and good science is just as value-laden as bad science. **What distinguishes good from bad science isn't the absence of value judgments, it's the *kind* of value judgments that are in play.**

v. Implications for the Philosophy of Science?

The other comment I want to make is that accepting this view of the scientific process still leaves open most of the big philosophical questions about the nature of science.

For example, I might be what philosophers of science call a **scientific realist**, which means that I think that one of the proper goals of a scientific theory is to describe the world beyond what we can directly measure and observe, and that sometimes science is successful in doing this.

You, on the other hand, might be what philosophers of science call an **instrumentalist** about theories: you think that science doesn't aim to describe the world beyond the realm of observable phenomena at all. You think that scientific theories are just instruments for helping us organize and predict observable phenomena and the results of experiments, and that's how they should be judged; and we should treat the theoretical parts of our theories as nothing more than useful fictions, or at least be agnostic about their truth.

This debate between scientific realists and instrumentalists is one of those big philosophical questions about science. Some version of it goes back to the Greeks. My point is that **either view is compatible with thinking of scientific methodology as a set of tools for neutralizing cognitive biases.** I'm not saying that both are equally plausible views, all I'm saying is that they're equally compatible with everything I've said about scientific methodology in this podcast.

vi. The Take-Away Message

So, the take-away message of this position is really fairly limited. It doesn't imply much for the big philosophical questions about the nature of science. But it does suggest a certain kind of attitude toward science, and the authority of science.

It suggests, first of all, that **people should be very cautious about relying on their intuitions in judging a scientific issue.** Our intuitions are just not reliable. One kid developing autism after a vaccination does not imply that the vaccine was the cause of the autism. But we all know that, the sample size is too small. What about 6000 kids? Our intuition tells us that if 6000 kids develop autism after being vaccinated, that's at least evidence for a strong correlation, right?

WRONG. It's evidence, but it's lop-sided, it's an incomplete body of data. Think of the Dr. Jones example.

When you actually look at a more complete body of data, including background rates of autism, the evidence is clear: there is no statistically significant correlation between vaccination and the development of autism. The number of reported cases of autism has certainly shot up over the past thirty years, but part of this is attributable to changes in diagnostic practices; how much of an increase there's been in the actual prevalence of the condition is still unclear, but there's no evidence that it's linked to vaccinations. Multiple studies from different scientific bodies agree on this conclusion.

Now, I know that a lot of people in the anti-vaccine movement resist this conclusion, and there are a lot of conflicted parents who see this as a tug-of-war between equal sides, an anti-vaccine side and a pro-vaccine side. **But the take-away message of this lecture is that the sides are not equal, and we shouldn't view them as equal**. Human beings, left to their own devices, will see correlations where there aren't any, and attribute meaning to correlations that are actually meaningless. The more invested you are in the outcome, the more likely it is that you'll be led into error. **Only a proper scientific study can resolve the issue, and when multiple studies converge on the same conclusion, then the rational thing to do, provisionally, is to accept the scientific consensus.**

We may all have strong intuitions the other way, but once we realize how prone to error we really are, and how scientific methods are designed precisely to avoid those errors, then I think we do have a compelling argument for accepting the authority of science, especially when there's a consensus on the issue among the relevant experts in the mainstream scientific community.

3. Confirmation Bias and the Evolution of Reason

In this lecture I'm going to follow up on our previous discussion of **confirmation bias** by looking at an interesting attempt to explain confirmation bias drawn from the field of **evolutionary psychology**. In the commentary on this theory I hit on a bunch of topics, including the relationship between **rhetoric and argumentation**, **the nature of biological functions**, and an important virtue that I call **epistemic humility**.

i. Cognitive Bias Research: Phenomena vs Theory

Last lecture I talked a lot about *confirmation bias*, this tendency we have to filter and interpret evidence in ways that reinforce our beliefs and expectations. And I argued that one way to think about science and scientific methodology is as a set of procedures that function to neutralize the distorting effects of confirmation bias (among other cognitive biases) by forcing us to seek out and weigh even the evidence that might count against our beliefs and expectations.

There are **two sides** to cognitive bias research. There's the science that *describes the effects of these biases* on our judgment and behavior, and there's the science that tries to *explain why* we behave in this way, that tries to uncover the psychological or neurological or social *mechanisms that generate* the behavior.

Everyone agrees that confirmation bias is a very real phenomenon, the descriptive part is well established. But not everyone agrees on the *explanation* for why we're so prone to this bias, what *mechanisms* are at work to generate it. And when there's disagreement at this level, there can be disagreement about how to best counter-act the effects of confirmation bias. So *as critical thinkers we should be interested in these debates, for this reason, and because they're relevant on a deeper level to how we should think of ourselves as rational beings.*

In this lecture I want to talk about an interesting approach to these questions that has generated some recent buzz. It comes out of the field of

evolutionary psychology, which is a branch of psychology that tries to explain and predict our cognitive functioning by showing how various aspects of this functioning can be viewed as evolutionary adaptations. This approach views these cognitive processes as *products of natural selection*, meaning that they were selected for because they helped our evolutionary ancestors survive and reproduce. **What I'm going to be talking about is a view of human reason from this evolutionary perspective that offers an interesting explanation for the phenomenon of confirmation bias.**

ii. An Evolutionary Explanation of Confirmation Bias

So from this perspective, we can ask the broad question, *why* did human reason evolve? And by "reason" here I mean a very specific ability, namely, *the ability to generate and evaluate arguments, to follow a chain of inferences, and to construct and evaluate chains of inferences that lead to specific conclusions.* Human rationality can be defined much more broadly than this, but we're focusing on this specific component of rationality for the time being.

Now, if we assume that this ability to construct and evaluate arguments is an evolutionary adaptation of some kind, the question then becomes, what is this ability an adaptation FOR?

The Simple and Obvious Story

Well, here's one simple and obvious way to think about it. Our ability to construct and evaluate arguments evolved because it has *survival value*, and it has survival value because it helps us to arrive at *truer beliefs* about the world, and to *make better decisions* that further our goals. This ability to reason is a general purpose tool for constructing more accurate representations of the world and making more useful and effective decisions. We assume that in general, ancestral humans that are better able to reason in this way will have a survival advantage over those that don't. So we expect that a higher percentage of individuals with this trait will survive and reproduce, and over time the trait will come to dominate the population, and that's why the trait evolved and persists in human populations.

Confirmation Bias: A Problem for the Simple and Obvious Story

Now, if we accept this simple story, then we have an immediate problem. The problem is that when we look at the psychological literature, we see that *human beings are often very bad at following and evaluating arguments, and they're often very bad at making decisions*. This is the take-home message of a good deal of the cognitive biases research that's been conducted over the past 40 years!

To take the obvious example, human beings are systematically prone to confirmation bias. Confirmation bias leads us to disproportionately accept arguments that support our beliefs and reject arguments that challenge our beliefs, and this leads to errors in judgment; we think our beliefs are more justified than they really are.

Now, from this simple evolutionary stance, **the existence of confirmation bias is a bit of a puzzle.** If reason evolved to improve the quality of our individual beliefs and decisions, then what explains the persistence of confirmation bias, and other cognitive biases that undermine the quality of our individual beliefs and decisions?

The Argumentative Theory of Reason

The new approach to these questions that is getting some recent attention tries to resolve this puzzle about confirmation bias. It's known as the **argumentative theory of reason**, and it claims that **the central adaptive function of human reasoning is to generate and evaluate arguments within a social setting, to generate arguments that will convince others of your point of view, and to develop critical faculties for evaluating the arguments of others.**

Now this might not seem like a radical hypothesis, but I want you to note the contrast between this view and the previous one we just described. The previous view was that the central adaptive function of human reason was to generate more accurate beliefs and make better decisions for individuals. In other words, **the function is to improve the fit between individual beliefs and the world**, resulting in a survival advantage to the individual.

The argumentative theory of reason rejects this view, or at leasts wants to seriously modify this view. It says that **human reason evolved to serve social functions within human social groups.**

What are these functions?

Well, imagine two ancestral humans who are trying to work together, to collaborate to find food, defend against aggressors, raise children, and so on. This all works fine when both parties agree on what they want and how to achieve it. But if a disagreement arises, then their ability to work together is compromised. They need to be able to resolve their disagreement to get back on track.

Now let's imagine that this pair of ancestral humans lacks the ability to articulate the reasons for their respective views, or the ability to evaluate the reasons of the other. They're stuck, they can't resolve their disagreement, and because of this, their collaboration will probably fail and their partnership will dissolve.

Now imagine another pair of ancestral humans in the same situation who have the ability to articulate and evaluate their reasons to one another. They have the potential to resolve their disagreement through mutual persuasion. This pair is more likely to survive as a pair, and to reap the benefits of collaboration.

And for this reason, this pair will likely out-compete groups or individuals who lack the ability to argue with one another. And **according to this theory, that's the primary reason why the ability to reason evolved in human populations — to serve the needs of collaboration within social settings, not to improve the quality of individual beliefs or to track the truth about the world.**

How The Argumentative Theory Explains Confirmation Bias

The argumentative theory of reason is the product of two french researchers, the well-known anthropologist and cognitive psychologist **Dan Sperber** and his former student **Hugo Mercier**[2].

One of the reasons they offer in support of their theory is that it helps to explain the existence of confirmation bias.

How does it do this? Let's walk through this, it's kind of interesting.

In a social setting where there are lots of different individuals with different beliefs and values, everyone is required to play two roles at different times, the role of the *convincer* — the one who is giving the argument that is intending to persuade — and the role of the *convincee* — the one who is the recipient of the argument and the intended object of persuasion.

Now, if your goal as a convincer is to use reasons to persuade others, then a bias toward confirming arguments and confirming evidence is going to serve you well. As a convincer your goal *isn't* the impartial weighing of evidence to get at the truth, *it's to assemble reasons and evidence that will do the job of persuading others to accept your conclusion.*

Things are different when you're the convincee, the one who is the object of persuasion. In this context you can imagine two extreme cases for how you should handle these attempts to persuade you. On the one hand, you could decide to accept everything that other people tell you. But that's not going to serve your needs very well — *you'll be pulled in different directions, you won't have stable beliefs, and you won't be effective at asserting your own point of view.*

On the other hand, you could decide to reject everything that other people tell you that doesn't conform to your beliefs. This is the ultra-dogmatic position, you stick to your guns come what may. This has some obvious advantages. You'll have a stable belief system, you'll attract collaborators who think they way you do, and so on.

[2] Mercier, H. and Sperber, D. (2011) "Why Do Humans Reason? Arguments for an Argumentative Theory", *Behavioral and Brain Sciences* 34, 57-111.

But it's still not ideal, because t*he ultra-dogmatic stance runs the risk of rejecting arguments with true conclusions that would actually improve their condition if they were to accept them.*

A better compromise position is one where there's a default dogmatism, where your initial reaction is to resist arguments that challenge your beliefs, but this default dogmatism is tempered by a willingness and ability to evaluate arguments on their merits, and thereby make yourself open to rational persuasion.

From an evolutionary standpoint, this compromise position, which you might call "moderate dogmatism", seems to offer the maximum benefits for individuals and for groups.

Now when you combine the optimal strategy of the convincer, which is biased toward arguments and evidence that support your beliefs, and the optimal strategy of the convincee, which is biased against arguments and evidence that challenge your beliefs, *you end up with an overall strategy that looks an awful lot like the confirmation bias that psychologists have been documenting for decades.*

And this is what Sperber and Mercier are saying. When we think of reason as serving the goals of social persuasion, confirmation bias shouldn't be viewed as a deviation from the proper function of human reason. Rather, it's actually a constitutive part of this proper function, this is what it evolved FOR. **To use a computer software analogy, confirmation bias isn't a "bug", it's a "feature".**

Consequences of This View

Now, I don't know if this view is right. But I do find it provocative and worth thinking about.

I mentioned earlier that different views on confirmation bias can result in different views of how best to neutralize or counter-act it. Sperber and Mercier argue that their view has some obvious consequences along these lines.

For one, their view implies that *the worst case scenario is individuals reasoning alone in isolation*. Under these conditions we're most prone to the distorting effects of confirmation bias.

A much better situation is when individuals *reason in groups* about a particular issue. This way everyone can play both the role of the convincer and the convincee, and we can take advantage of our natural ability to evaluate the quality of other people's arguments, and others can evaluate the quality of our arguments. We should expect that reasoning in groups like this will result in higher quality judgments than reasoning in isolation, and lots of studies on collective reasoning do bear this out.

But of course not all groups are equal. If a group is very homogeneous, with lots of shared beliefs and values and background assumptions, then the benefits of group reasoning are more limited because there are collective confirmation biases that won't be challenged.

So, if we're looking to *maximize the quality of our judgments and our decisions*, the better situation is when the groups are not so homogenous, when there's *a genuine diversity of opinion represented within the group*, and you can expect that at least some people will start off disagreeing with you. Under these conditions, the benefits of group reasoning and group argumentation are greatest — **the result will be judgments that are least likely to be distorted by confirmation bias**.

This is one of the conclusions that Sperber and Mercier arrive at. I don't think you need to accept the whole evolutionary framework to understand the benefits of group argumentation, but it is an interesting explanation for why the quality of argumentation varies in this way, from lowest involving individuals in isolation to highest involving heterogeneous groups.

Take-Away Messages

So, what are the take-away messages of this story for us as critical thinkers? I have three in mind:

1. It can be very enlightening to take an evolutionary perspective on the origins and nature of human reason, but we need to be careful about how to do this.

Almost everyone thinks that evolutionary biology can and should shed light on psychology. But we need to be careful about this. "Evolutionary psychology" in the form that Sperber and Mercier practice it has its critics. Telling a compelling story about the possible evolutionary advantages of a trait isn't enough to establish that the trait really is an adaptation that was selected for because of those advantages. Evolutionary views need to be defended against non-evolutionary alternatives, and specific evolutionary stories need to be defended against alternative evolutionary stories. But this is just to say that we need to be careful about the science. For now I'm just going to mention this and put aside the question of the status of the science.

2. Some may see troubling (skeptical) implications in Sperber and Mercier's theory. I don't think these skeptical worries are warranted, but the shift in perspective does have implications worth thinking about.

What are these troubling skeptical implications? Well, remember that the theory involves a shift from thinking of the primary function of human reason as improving the quality of our individual beliefs and decisions, to thinking of the primary function as persuasion within human social groups.

One possible worry is that this shift threatens to collapse the distinction between argumentation and rhetoric that optimists about reason would like to maintain. Rhetoric is about persuasion, broadly construed. Argumentation is traditionally viewed as either distinct from rhetoric, or as a special branch of rhetoric that focuses on persuasion for "good reasons". It's the "good reasons" part that distinguishes argumentation from mere persuasion, and that connects it to the broader concepts of truth and rational justified belief that we emphasize in Western science and philosophy.

So, some might conclude that if Sperber and Mercier are right that the primary evolutionary function of human reason is social persuasion, then it turns out that argumentation really is just about persuasion, and not about truth and developing rationally justified beliefs. Some might think that it warrants a form of skepticism about our ability to improve the quality of our beliefs and decisions through reason and argument.

Now, **my reading of Sperber and Mercier's theory is that it doesn't support this kind of skepticism about human reason. I'll offer two reasons why.**

First, the story that Sperber and Mercier give about the evolutionary function of reason explicitly builds in a capacity for critically assessing the quality of arguments. But this capacity is grounded in our roles as convincees, not as convincers. From the perspective of someone on the receiving end of an argument, the capacity to distinguish good from bad arguments is genuinely adaptive, and their theory predicts that human reason retains this capacity. What Sperber and Mercier add is that human reason is also adapted for social persuasion, and the goals of social persuasion are sometimes in conflict with the goals of truth-finding and belief-justification. *So it's a messier, less unified view of human reason that they're advocating, but it's not inconsistent with the traditional view of reason as a means of distinguishing good from bad reasons for belief.*

The **second** point I want to raise has to do with this talk of *evolutionary functions*. Let's say, for the sake of argument, that human reason evolved primarily for social persuasion, that our ability to construct and follow arguments was selected for the advantages it afforded in terms of social persuasion and social collaboration. In other words, we're granting that it didn't evolve for the purpose of improving the quality of our individual beliefs and decisions. Sperber and Mercier don't actually go this far, but let's assume it for the sake of argument.

Now, does it follow, if this is the case, that human reason does *not* today have the function of improving the quality of individual beliefs and decisions? It's tempting to say "yes" given the way I've set it up, but the answer is "no", this doesn't follow.

To think this way is to make **a basic mistake about the nature of biological functions**. It's the *fallacy* of assuming that, if a trait evolved for a given function, *then it can only serve the function that it evolved for*.

In biology there are lots of examples of traits that originally evolved for a given function, *but that later were used for other purposes and became part of the natural function of the trait*. The classic example is **bird feathers**. Almost everyone agrees that feathers originally evolved to aid in *heat regulation*, not for flight — it was only later that they were co-opted and shaped by natural selection for flight. They were also co-opted for display and selection of mates, so that's three functions right there.

So, even if we granted that human reason evolved initially as a tool for social persuasion in the service of collaboration, it doesn't follow that our capacity for reason couldn't have been co-opted for other purposes and shaped by natural selection for other functions, like improving the reliability and accuracy of our beliefs about the world, or for other functions entirely.

Okay, so what's my third take-away message?

3. Background knowledge like this helps us develop strategies and attitudes that can improve the quality of our judgments and decisions.

This message is one that I've been trying to emphasize in these lectures, and it has to do with the importance of background knowledge for critical thinking. In previous lectures I've distinguished three different kinds of background knowledge that are relevant for critical thinking.

1. Background knowledge related to subject-matter, the topic under discussion.
2. Background knowledge related to the intellectual history of debate over a particular issue, the history of argumentation that has shaped the way people view the issue.

3. Background knowledge related to how the human mind works, how we actually form beliefs and make decisions.

The views on the evolution of human reasoning that we've been talking about here — along with the two previous lectures on cognitive biases — fit into this third category of background knowledge

From my perspective, it's clear that the more we understand how our minds actually work, the greater is our ability to bring critical thought to bear on our judgments and decisions.

The irony, however, is that sometimes what we learn is discouraging. We learn that we're less rational than we think we are, we're more prone to errors and over-confidence than we think we are.

But my point is that this is still useful information, because it can help us develop strategies for managing our irrationality in a rational way. And **it can help us identify the sorts of psychological attitudes and dispositions that are most conducive to critical thinking.**

Epistemic Humility

For example, if we know that we're prone to confirmation bias, but that this bias can be neutralized by following certain scientific protocols, or by reasoning together in diverse groups, then this can lead to strategies for effectively managing and reducing the effects of this bias.

Also, if we know that we're prone to confirmation bias, and we know that we're also prone to overconfidence, then it helps us to identify certain attitudes, or virtues, that should be cultivated to help avoid the effects of these biases.

One of these attitudes is what I like to call **epistemic humility**. "Epistemic" is a philosopher's term that means "pertaining to knowledge", so in this respect I'm talking about **humility regarding the status of our knowledge and our capacity to reason well**.

Now, this isn't the same as skepticism about *knowledge* — to be epistemically humble isn't necessarily to *doubt* our knowledge, or to deny the possibility of knowledge. **It's rather to adopt an epistemic stance that is appropriate to, and that acknowledges, our situation as fallible, limited beings that are prone to overconfidence and error.**

The degree to which we're prone to error will vary from context to context. The key idea here is that **the quality of our judgments is highest when our epistemic stance — the attitude we take toward our own status and capacities as knowers — properly matches the epistemic environment in which we find ourselves.** For example, if we're reasoning all by ourselves, this is a different epistemic environment than if we're reasoning with a diverse group of people. Given what we know about reasoning, it's appropriate to adopt a greater degree of epistemic humility when we're reasoning by ourselves than when we're reasoning with a diverse group.

And sometimes the appropriate stance is simply to not trust our own judgements at all. That's an extreme form of humility, but in the right circumstances it can be the most rational stance to take.

A classic example of this kind of rational humility can be found in the Greek story of Odysseus and the Sirens. The Sirens were these mythic female creatures who sang these beautiful songs that lured sailors into the water or to crash their boats onto the shores of their island.

Odysseus was very curious to know what the Siren song sounded like, but he understood that we may not be able to resist their song. So he did something very clever; he had all his sailors plug their ears with beeswax and tie him to the mast. He ordered his men to leave him tied tightly to the mast, no matter how much he might beg to untie him.

When he heard the Siren's beautiful song he was overcome by it and he desperately wanted to jump into the sea to join them, and as he predicted he ordered the sailors to untie him. But they refused based on his earlier orders. So, *as a result of his strong sense of rational humility regarding his own capacity to resist persuasion, Odysseus was able to experience the Siren's song and come out unscathed, where other men with less humility were lured to their deaths*.

For critical thinkers the moral of this parable is clear. Though it may seem counter-intuitive, **by accepting and even embracing our limitations and failings as cognitive agents, rather than denying them or struggling against them, it's possible to improve the quality of our judgments and make more rational decisions than we would otherwise**. But to pull this off we need to cultivate the right kind of epistemic virtues that are informed by the right kind of background knowledge, and through knowledge and experience, learn to develop the appropriate judgment about the right level of epistemic humility to adopt in any particular circumstance.