How First Principles Thinking Fails

commoncog.com/how-first-principles-thinking-fails

Cedric Chin December 1, 2020

Thinking Better

By Cedric Chin

Table of Contents

One of the threads I've been pulling on in this blog is the question "Where should the line lie between first principles thinking and pattern matching? When should you use one or the other?"

The question is interesting because many career decisions depend on good thinking, and thinking is split roughly into pattern matching against our experiences, or reasoning from first principles. My belief is that it isn't enough to have one or the other; you really need to have both.

The question I've chosen — "where should the line lie …?" — isn't a particularly good one, because the obvious answer is obvious: "it depends!" And of course it does: whether you do first principles thinking or perform some form of pattern matching

really depends on the problem you're trying to solve, the domain you're working in, and all sorts of context-dependent things that you can't generalise away.

As a direction for inquiry, however, the question has been fairly useful. You can sort of squint at the following posts and see the shadow of that question lurking in the background:

1. In Much Ado About The OODA Loop I wrote about John Boyd's ideas, and in particular Boyd's belief that good strategic thinking depends on accurate sensemaking, which in turn depends on repeatedly creating and then destroying mental models of the world. I followed that up with Good Synthesis is the Start of Good Sensemaking, which was an attempt to demonstrate the difficulty of good sensemaking under conditions of partial information, by putting you in the shoes of a battered CEO.
2. In Reality Without Frameworks I talked about the danger of using frameworks excessively. I argued that frameworks are compelling because they help give order to your world, but that they also colour and shape (and sometimes blind you to) what you see around you.
3. Then, in In Defence of Reading Goals, I went the other way and argued that for certain careers, it is a competitive advantage to build up a large set of patterns in one's head. The easiest way to do that is to read a large number of books; I argued that when seen in this light, it wasn't a terrible idea to set reading goals for yourself.

In fact, if you skim through the posts I've written over the past six months, you would find arguments for first principles thinking in some essays, and arguments for better pattern matching in others; I was essentially vacillating between the two positions. I suppose this is a way of arguing that you need both forms of thinking to do well — or that I think the art of good thinking lies in finding a balance between the two modes.

One useful way to look at this problem is to ask: *how can you fail?* That is: how can pattern matching fail you? And how can first principles thinking fail you?

In my experience, the second question is more interesting than the first.

First Principles Thinking, Evaluated

If I were to ask you how pattern matching might fail, I'm willing to bet that you can give me a dozen good answers in about as many seconds. We all know stories of those who pattern matched against the past, only to discover that their matches were wrong when applied to the future. There's a wonderful collection of bad predictions over at the Foresight Institute, though my favourite is perhaps the example of Airbnb (I remember hearing about the idea and going "What?! That would never work!" — but of course it did; the story has entered the realm of startup canon).

It's easy to throw shade at pattern matching — nevermind that expertise is essentially pattern matching, or that Charlie Munger appears to be very good at it. But I've always considered the second question a bit of a mystery. How can first principles thinking fail? It all seems so logical. How indeed?

One lazy answer is that pattern matching is often fast, whereas analysis isn't, and so pattern matching is better suited for situations where you have to make a decision quickly. But this is an edge case more than it is an interesting answer — we know that pattern matching may also be employed in slower, more considered forms of thinking: psychologists call this analogical reasoning.

(I'm mixing terminology a little; but bear with me — this is a blog post, not an academic paper).

Another possible answer is that you make a mistake when executing your reasoning:

- One or more of your 'principles' or 'axioms' turns out to be mistaken.
- You make a mistake in one of your inference steps.

In practice, mistakes like this aren't as huge a problem as you might think. If you run your reasoning past a sufficiently diverse group of intelligent, analytical people, it's likely that they'd be able to spot your errors.

What is interesting to me are the instances when you've built logically coherent propositions from true and right axioms, *and you still get things wrong anyway*.

Back when I was still running an engineering office in Vietnam, my boss and I would meet up every couple of months to take stock of the company, and to articulate what we thought was going on in our business. Our arguments were reasonably water-tight. Sometimes we would start from observations; other times, we would argue from first principles. My boss was rigorous and analytical. I respected his thinking. Between the two of us, I was pretty sure we'd be able to detect flawed assumptions, or unreasonable leaps of logic. We thought our conclusions were pretty good. And yet reality would punch us, repeatedly, in the face.

The *way* we got things wrong was always strikingly similar. When we first started selling Point of Sales Systems, most external observers told us that there was no money in it: that there were too many competitors, that the margins would be too low, that we wouldn't be able to bootstrap the business. *We* thought the same things, but my boss decided to give it a go anyway. I remember spending months afterwards, talking with him, trying to answer the question "why are we making so much money?" We obsessed over this for *months*. In retrospect, we were blessed to be able to ask this question. But the fact that we didn't understand why there wasn't more competition made us antsy as hell. We had any number of theories. They were all wrong.

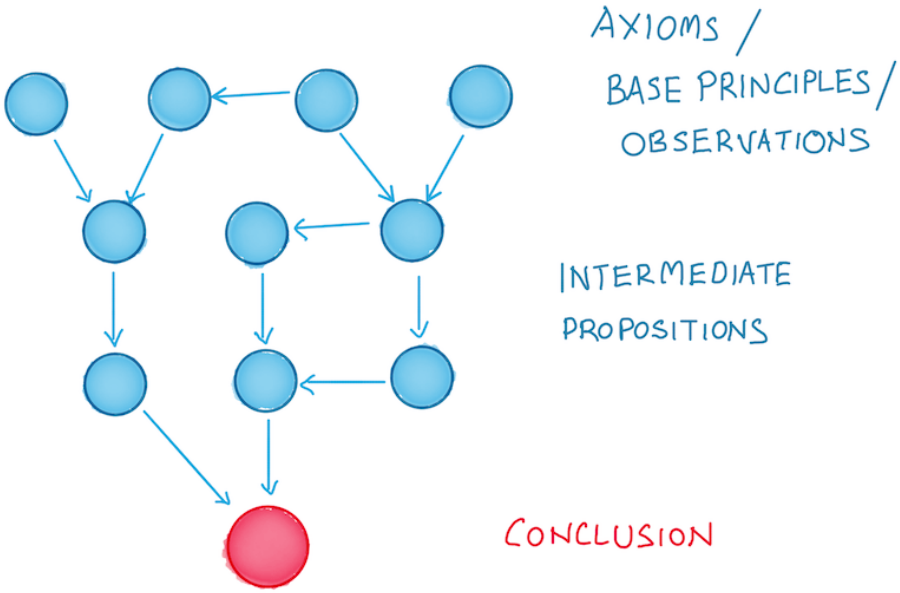
We discovered the answer much later. It turned out that the market was distorted in a very particular way. The government wanted to encourage the adoption of self-service checkout machines, and Point of Sales systems were included under that definition. This meant that they were willing to give out thousands of dollars in grants in order to subsidise the cost of adoption. If you were a vendor during that period, and you got onto an approved vendor list, you had access to those grants; the vendor list served as a chokepoint to the rest of the market. This explained the margins we were seeing, along with a large number of market characteristics we had observed but not understood. (Note: I'm leaving a number of details out of this account, because my old company is still a player in the market. The shape of the market has changed, naturally, and the grant situation is different.)

In other words, there was a fact that we didn't adequately understand. Without knowledge of that fact, we couldn't bootstrap a good explanation.

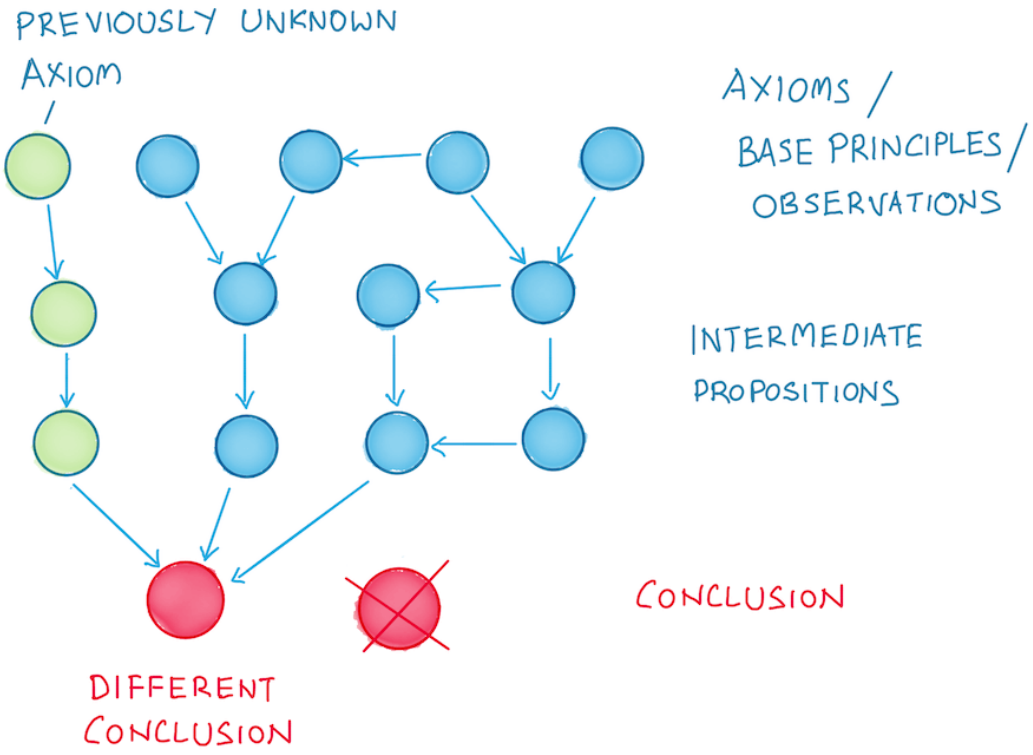
It eventually got to a point where we — or more accurately, *I* — no longer said things like "Ok, this argument makes sense. I think it's right." Reality had punched me in the face one too many times to be that confident. I began to say things like: "Ok, this analysis checks out. It seems plausible. *Let's wait and see.*"

Part of the reason I had to write Seek Ideas at the Right Level of Abstraction before writing this post was because I had to work out the general idea from my experiences. Today, I know that you can build analyses up from base principles … only to end up at the wrong level of abstraction. This is one way to fail at first principles thinking.

But I think there's a more pernicious form of failure, which occurs when you reason from the *wrong set of true principles*. It is pernicious because you can't easily detect the flaws in your reasoning. It is pernicious because all of your base axioms are true.



If you pick the wrong set of base principles — even if they're all true — you are likely to end up with the wrong conclusion at the end of your thinking. In other words, the only real test you have is against reality. Your conclusion should be useful. It should produce effective action.



Conversely, it’s possible to ‘figure everything out’, only to learn a new piece of information that changes everything, that reconfigures all the reasoning chains in your argument.

As the old saying goes: in theory, theory and practice are the same. In practice, theory and practice are different.

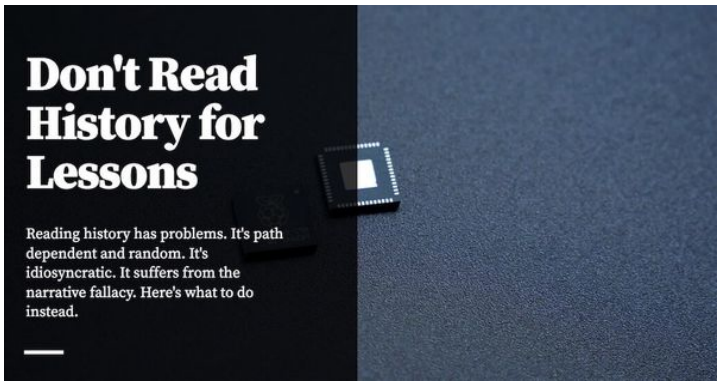
To paraphrase that aphorism: in theory, first principles thinking always leads you to the right answer. In practice, it doesn't.

Update: I wrote a follow-up post to this, which goes deep on one real-world example of a first principles argument gone wrong: The Games People Play With Cash Flow.

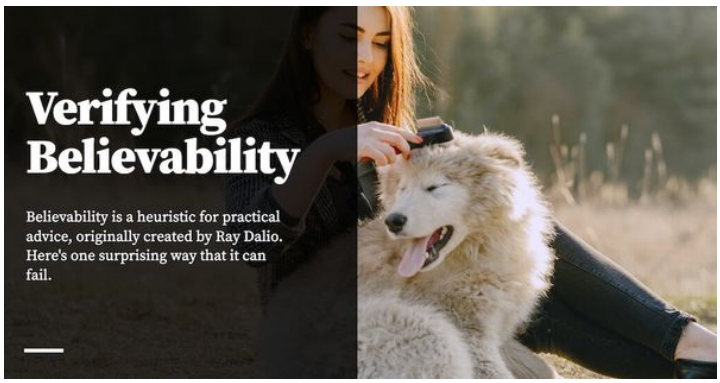
Originally published 02 December 2020, last updated 09 December 2020.

Member Comments

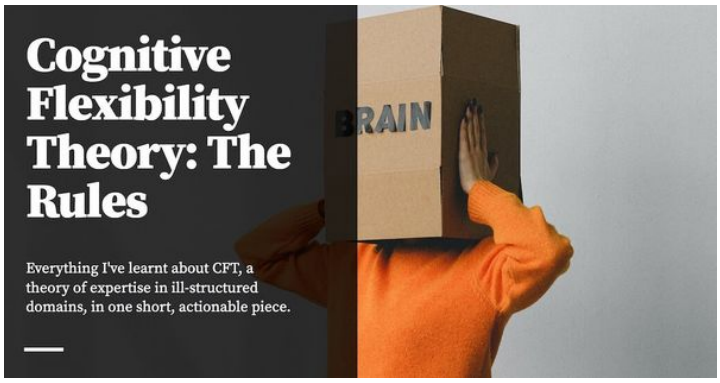
More in Thinking Better



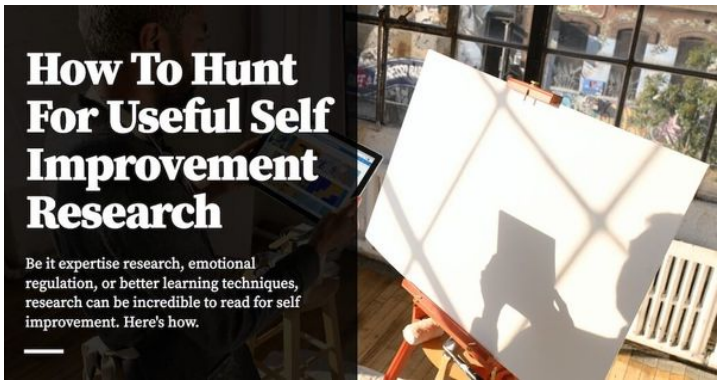
Don't Read History for Lessons



Verifying Believability



Cognitive Flexibility Theory: The Rules



How To Hunt For Useful Self Improvement Research