Using Causal Questions for Theory Development

D. Alex Hughes

2025-07-22

Table of contents

Learning Objectives

At the end of today's session, we should each be better able to:

1. **Identify** the major theories in your discipline, and **apprecaite** the value in conversations informed by theory.
2. **Articulate** how your research question is informative of the major theories in your discipline.
3. **Recognize** that all theories have causal implications.
4. **Apply** the idea of the *ideal experiment* as a dark mirror for theory understanding.

1 Introductions

1.1 Myself

1.2 Who 'Dis

- Do you think of your professional work as belonging to:
  - A *scientist*?
  - A *theorist*?
  - An *engineer*?
  - Is there another descriptive label?
- What field are you writing in? Who would recognize your work?
  - One of the *computer sciences*?
  - One of the *critical fields*?
  - One of the *social sciences*?
  - One of the *natural sciences*?

2 Finding Theories

- The goal in this section is to move from the questions that we're researching to the theories that are running through our fields.
- This is a strong claim, but no matter your field, whether it is a natural science, a social science, a journalistic field there are theories that are being looked at.

2.1 Research Questions

Many of us organize our research, and our talks which I saw early last week, around research questions. They keep us from getting *lost*, they keep our work *organized,* the focus us on tasks that we can actually complete.

But, you're doing the *fun* work of your dissertation right now! You're taking on three, or more, different tasks that are all organized around a theme. That theme *might as well be* a major theory in your field!

This is how **books** come to be, and your dissertation is really a first draft of your tenure book.

2.2 Breakout One

Breakout 1

- Remind yourselves, what is the role of theory building in the cumulation of knowledge in the social sciences?
- Is it important that what we know accumulates? If so, why? If not, why not?
- What is your favorite social sciences textbook?

2.3 Can AI Accumulate Knowledge?

- Ralph, this morning, contended that it is, or might be possible, for AI to accumulate knowledge. That there is a way to systematize the process that we undertake as social scientists: to observe, abduct, place against known theory, and then produce hypotheses, and test those hypotheses?
- Do you agree? Where are the points that you think an AI system is likely to be especially helpful in this process? Where are the points that you think an AI system is likely to be especially *unhelpful* in the process?
- If AI *can* accumulate knowledge, and then produce new knowledge, should you be in this field — academics — right now? Should you be in this room right now?
- I'm going to argue "Yes" even if AI can accumulate knowledge there's an important role for us — what the machines have access to are probabilistic sets of words that belong together or in conversation; they can help us to understand what has been *done* and what has been *thought* — but it does not have the capabilities (yet) to intervene and run trials or think counterfactually. *At least for a little while, we're still ahead of the machines.*

2.3.1 Discussion Questions

1. Research Questions

Take a five minutes to write down for yourselves (it is OK for you to use a LLM):

- What are the major *questions* in your field?
- This might require that you think back to your field exams! Try to think not about a specific study, but questions that many people who are working on.

Examples:

1. Why does this apple fall from this tree, rather than floating away?
2. Why *do*, and why *don't*, citizens participate in electoral politics?
3. "Misinformation"… ?

What I think that we're going to see is that the major questions that exist in the disparate fields are actually quite similar! Or, at least, if you squint a little bit, the ways that we think of information access among youth in Puerto Rico have to be related to the explanations news production in Canada, and civic technology utilization in Georgia.

3 Answers to Research Questions are Theories!

As you're moving from questions being asked, toward some more general explanation of phenomena, you're going to use our incredible, creative, human ability to explain patterns with simpler stories. We do this *all. the. time.* Learning and abstraction, which I'm actively seeing my child undertake on a daily basis, is one of the core human super powers!

- We have burned ourselves on a fire ring at a campsight on the California coast. How quickly we're able to infer that a fire ring with an open flame in the Sierra Nevada mountains has the same properties, and will cause the same result, *despite* being an entirely different location, context, elevation, and flame!
- We have eaten cereal with cow's milk that has a certain off smell and feel ill afterwards. How quickly we are able to recognize that smell, and avoid products that have such a smell in the future. And! We're able to distinguish that smell of soured milk from the smells of cheeses, yogurts, and other similar cultured milk products.
- We have grown up in the midwestern parts of the United States (or possibly other places, too) and know that when the weather changes rapidly from sunny to overcast, there is rain that is imminent.

i Pattern Learning

- How can we learn patterns so quickly?
- Are we ever wrong when we're learning these patterns? Are there times that the general rule that we see is incorrect? To broad? Too narrow?

Are we generalizing knowledge — called *induction* — or, are we proposing what we believe to be the most likely explanation — called *abduction* (Charles Sanders Pierce)?

i 2. Answers to Questions

Take five more minutes.

- What are the possible answers to these questions?
- Which of these — research questions, or theories — are easier for you to find? Why?

Theoretical Answers

1. Apples fall from trees with predictable paths and rates because apples are composed partially of earth and partially of water (rather than air or fire). Since things that are composed of either earth or water belong at the center of the universe, which is the center of the Earth, they fall toward where they rightfully belong (Aristotelian explanation.)
2. Citizens operate subject to resource and cognitive constraints and make trade offs between action and inaction that can be understood as noisy bets made by lazy thinkers.

3. ...?

3.1 How do we notice these patterns?

Breakout: Noticing Questions and Answers

Take five minutes as a group to talk about: How, practically, do you go about noticing the things that are interesting research questions; or, the things that are likely explanations for answers to your research questions?
Do you simply look around? Read widely in your field? Take baths and ponder?

3.2 Theories are the proposed, sometimes incomplete, answers to questions in your field!

A field of economics is concerned with the question, "How do firms work?" which really means, "How can a company get the most out of its workers?"

Suppose that you are a plant manager in my home state of Michigan who wants your plant to be as productive as possible.[1]

You might think that people who aren't monitored will use the time that they're not being watched to *shirk*, which is to do work that is of lower effort than you would like. (For a review of this question, see Banidera, Barankay, Rasul (2011).)

How would reduce the amount of shirking? (Tighter monitoring? Monetary incentives for work completed? Social pressures from managers who have monetary incentives? Social rewards for people who do especially good work?)

[1]Stretch your theory. Suppose that you're a lab manager who is managing post-docs and PhD students? Stretch further, you're an Associate Dean who wants to get the most out of your School/Department.

An example of solution that cannot be experimented easily? Authority.

3.3 Example: Perfect Pizza

This is going to age me, but there was a glorious moment, pre-pandemic, where *Bon Appetit* had won the internet.

Their YouTube had everything – star power, shiny production, and a connection to something that is important to everyone the world over – **pizza!**

They conducted a series called "Making Perfect." The goal was to make, well, the *Perfect Pizza* at home.

https://www.youtube.com/watch?v=STpv0aTReIw

3.4 Perfect Pizza

What are the component of a perfect pizza?

Let's think through that exercise together. What would make the *perfect pizza?* How would you approach it? How would you describe it?

3.5 Finding Theories that Matter

Here's the hard part: When, or what, is the tradeoff between (a) working within an existing theory and finding (another) "novel" test of that theory that is going to contribute to the cumulation of knowledge (which Ralph said rather woefully are written, but never cited); or, (b) challenge a theory to produce a revision, change, adaptation, or "notch" in the evidence consistent with the theory that is meaningful?

i Open Discussion

Should you try and *time* your contribution so that you're working on something that is hot when you're on the market? So you can be a superstar on that junior search cycle? Or, should you work on the theories that you think are interesting, no matter if they're in a backwaters that are largely irrelevant to your field?
What are the tradeoffs that you see as 4th year PhD?

4 Theories to Questions

Ok, so we can move from questions to theories. How do we go the other way around, and why is this important?

5 Theories to Questions

5.1 Answers to Questions Separate Theories

If you propose things *work* differently than the field currently believes, you're going to have to convince the field. And, it is important to remember – the field doesn't like to change its mind. Researchers have reputations, and grants, that they intend to retain. Even if your work is simply directed at the concepts, there will be *human-level* resistance to what you're proposing.

And so, if you've got this uphill battle to change the minds of your discipline, how do you do it most effectively?

> **i** Changing Minds: Part 1
>
> How do you best structure your argument/book to be most effective at changing the minds of your field? How do you make it so they *have to* engage with what you've written? What is the role of logical deduction in this process?

5.2 Theory → Hypotheses

> *"OK, I've got this theory. I think that {these things} {do this} {because of} {when that}."*

Now what?

The process to go from your observations to the most likely set of explanations — the process of abduction — is one of the steps that produces the statement above .

But, it cannot be the only step — if that were the only step, we wouldn't be social scientists; we'd be evangelists and zealots.

Because it is entirely within your own head it is fallible. Our minds' ability to abstract events to patternss in amazing, but by its very structure imprecise, and is it entirely possible that we might learn the wrong generalization, or abduct what seems to be a reasonable explanation but that is incorrect.

After all, when we're seeing some real-world event we're seeing it within a limited context frame and that context shapes how or what might occur. Although there are many, many explanations for why things occur, there is almost certainly a boundless set of possible context frames that we could have been within!

5.3 Deductive Validity

Position: Your goal, at some point in your dissertation, is to produce a prediction that would be true *if and only if* the theory you believe to be the true explanation were in fact true.

Jeez. Let's unpack that.

First, recall that you're working within the *enterprise* of academic, and probably social scientific research. The people who are working within the field carry forward the beliefs of their advisers (and their advisers' advisers…). And so, there are a lot of beliefs out there — a lot of explanations why things happened the way they did.

In your dissertation work, and in the work that follows, you're taking a position on how the world works. That's your job as a member of this academy. But, your position is just one of many possible positions; and, certainly, there will be other people who hold those other positions. They're not straw-people positions, and there's little value in positioning your work against such a staged argument.

> **i** Examples of Differing Positions
>
> - Misinformation and disinformation are scourges that fundimentally threatens democratic accountability.
> - We've seen clear examples of the harms of misinformation and disinformation through the last several global election cycles. But, these are disequilibrium events and society will learn and develop ways to avoid such events in future states.

Only one of those positions can be true — either we're screwed, or there are relatively minor fixes that we can enforce that will pull us out of this morass. Which of these states of the world are we in, and how would you know?

Ask yourselves: "Why?" Why might we be fundamentally screwed? What about humans' interactions with decentralized media create this deep risk? **OR**, Why are there controls that we can actuate that are likely to control the worst machinations of this political system?

Changing Minds: Part 2

How do you best structure your career to be most effective at changing the minds of your field? How do you make it so they *can* engage with what you've written?

6 Causal Implications

Here's a hot take for you:

Every (good) theory has a causal implication.

6.1 What does it mean to cause?

There is a sub-field of language model training that is concerned with teaching machines to learn what "cause" is. As it turns out… it continues to be a difficult task for machines to figure out.

As an example:

> Lauren and Jane work for the same company. They each need to use a computer for work sometimes. Unfortunately, the computer isn't very powerful. If two people are logged on at the same time, it usually crashes. So the company decided to institute an official policy. It declared that Lauren would be the only one permitted to use the computer in the mornings and that Jane would be the only one permitted to use the computer in the afternoons. As expected, Lauren logged on the computer the next day at 9:00 am. But Jane decided to disobey the official policy. She also logged on at 9:00 am. The computer crashed immediately. Did Jane cause the computer to crash?

How do we think, as humans, think about this?

6.2 Causal Thinking

> **i** What does it mean for an action to *cause* an outcome?
>
> Literally, that's the prompt. What does it mean to cause?
> Take five minutes to discuss that, and we'll come back.

6.2.1 Judea Pearl and DAGs

One of the systems, which Ralph has talked about this morning, is DAG thinking — derived from Judea Pearl's *Causality (2000)?* In this line of thinking, you write down all the concepts in the world, and you draw the "geometry" of what you understand to cause what. Once you write those down, you can reason about what you need to measure, and… *Bob's Your Uncle.*

- What are the limitations of this approach?

6.2.2 Counterfactual Reasoning (Neyman (1923), Lewis (1973), Rubin (1978))

We employ counterfactual reasoning. This means that we think about one circumstance where action would be taken, and we think about the *very same* outcome when the action would not be taken. In this sense, there are two states of the world that are exactly the same but with only a single difference.

If outcomes are different between these two states of the world, then we say that the action causes the outcome.

> What is the problem with this line of reasoning?
>
> - Do you have a problem with this way of thinking?
> - Do you have a problem with using this as a way of producing evidence that is consistent with one theory and inconsistent with another theory?
> - How is what we have identified as a problem, actually an opportunity for us as academics to produce knowledge?

6.3 Finding Causes

How do we go about finding the antecedents (causes) and their consequences (effects)?

These are the if-then squiggly lines that Ralph alluded to. The sets of connections that undergird the realty that we exist within.

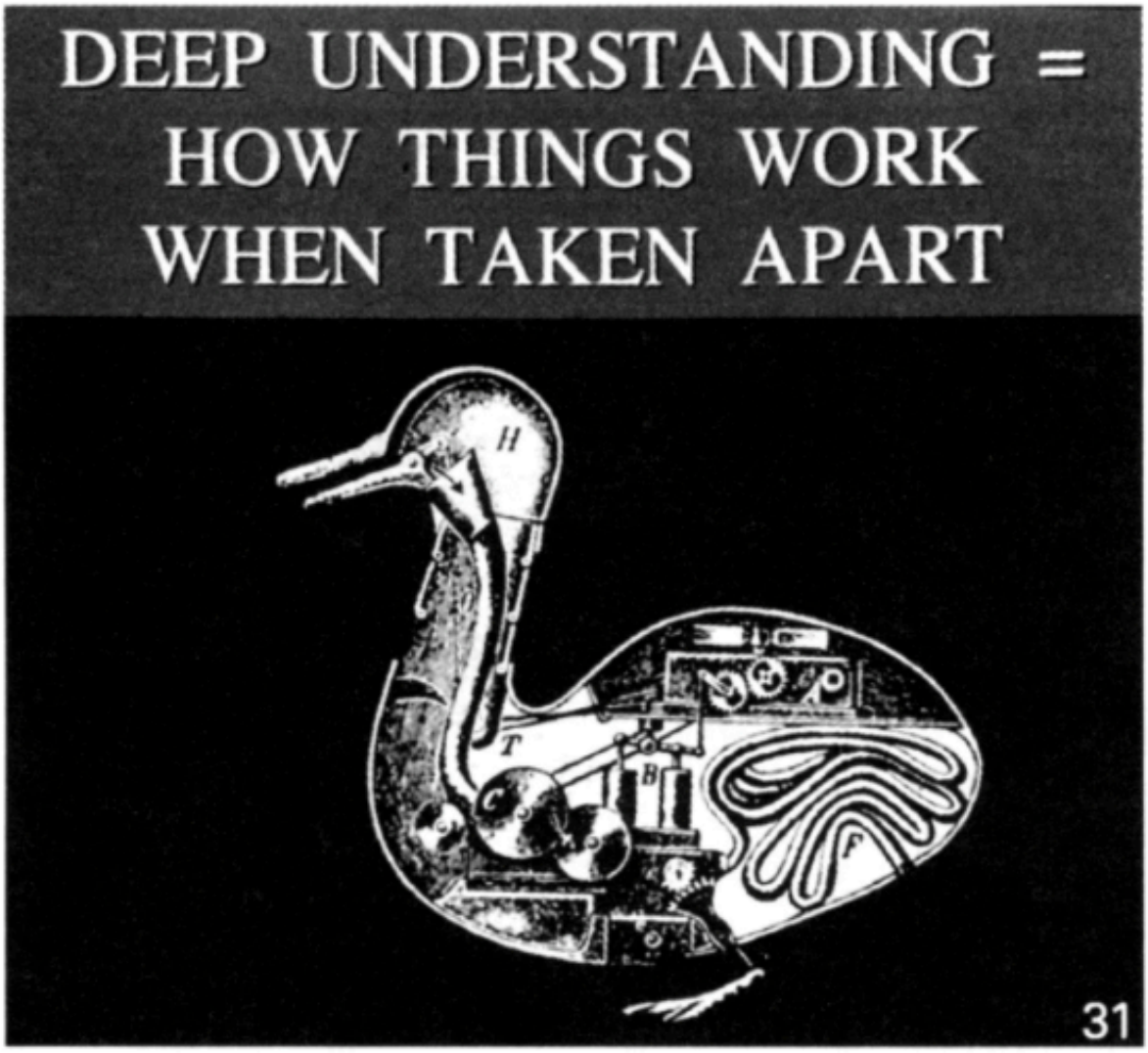

Figure 6.1: What happens when we take things apart?

6.4 Theories that Don't Have Causal Implications

- **Public Health?** Think about Ronald Fisher — would you die for your theory? He did! Public Health (not a social science, but instead a medical field that represents the worst of the social sciences…) talks in terms of correlates of; determinants of; associates with.

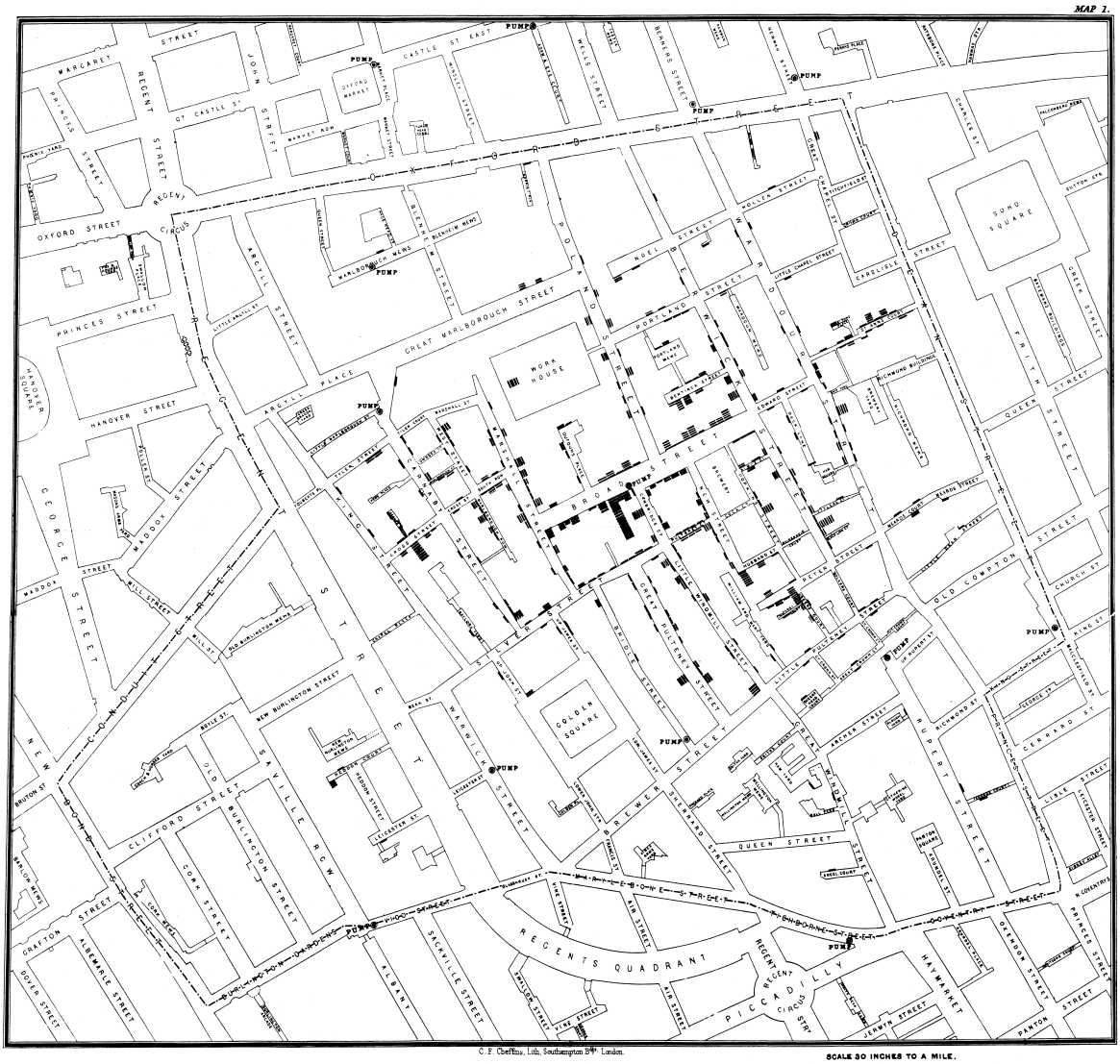

Figure 6.2: John Snow's Map of London, Showing Outbreaks of Choleara

Discussion Question

Look back to the theories that you wrote down earlier — are there any of these, for which, you cannot imagine a causal implication?

7 Ideal Experiments

Here's the thing — you're probably a 4th year PhD student. Which means…

… and running field experiments in the real world is *mind-bendingly expensive*. For example, experiments that Alex has run in Guatemala ($25M total program; $1M experiment) and Pennsylvania ($2M total cost)

What's more — you've probably already proposed your work to your advisers and the worst outcome that could happen from this Summer Doctoral Program is that you return to your institution, throw out your work, and blame Eric, Ralph, David and Kimmy.

But. Can you imagine the **ideal experiments** that you could conduct that would be consistent with your theory and inconsistent with other theories?

7.1 Ideal Experiment

The **ideal experiment** is the experiment that you would conduct with:

- Unlimited Time
- Perfect Measurement
- Infinite Budget

There's a part of me that says, "Well, that's easy if you've got unlimited resources and perfect measurement." But, trust me, *it isn't...*

The ideal experiment has to produce evidence that is consistent with one theory but not consistent with any other theory; and it has to do so *cleanly*, that is, without complex causes (where more than one feature moves at at time).

The ideal experiment is *not* concerned with:

- **Practical considerations**: it would be hard to…
- **Normative considerations**: it would create compensatory harm if one unit got {this} while the other unit got {that}

- **Statistical considerations**: your focus is on the design and the relationship between the outcomes, intervention, and theory… not on getting the model right. Hand-wave that model right out the door.

7.2 Failure Points in Ideal Experiments

- **Your theory isn't precise enough.** While you've been building your explanation for "why" the world works, you've elided parts of the thinking that are necessary to produce a whole story. Imprecise theory produces bulky tests that don't actually have crisp predictions.
- **Your theory overlaps too much with existing theories.** If you cannot produce a prediction that would be true *if and only if* your theory were correct, you haven't made a contribution to the field!
- **The other theories aren't precise enough.** It isn't a problem of your creation, but it *is* a problem that you've got to address. If the existing theory ("Prospect Theory" anyone?) isn't sufficiently well developed, then it will overlap with your work.
- **You don't know all the theories out there**. This is hard, you're a burgeoning professional, but you don't know everything! (Neither do I; Ralph and Eric may).

7.3 Applications of Ideal Experiments Thinking in Social Sciences

This morning, Ralph clapped out computer scientists who step into social scientific work. I'm with him on this! Another smart, math-y CS student who discovers graph theory….

But, let me give an example of someone who has done this **very well**.

> **i** Application of the Thinking Cycle to Other's Work
>
> Dan Jurafsky (incidentally a Berkeley PhD, now Full at Stanford with 100,000 cites) asked the question, "Why are Black drivers detained at higher rates in Oakland than white drivers?"
> Can you think through the entire cycle: research question, possible theories and an experiment that would produce evidence that would separate those theories?

> **i** Apply to Your Work