Memo: Volitional AI as Abduction
Dec 6, 2023

Goal:

- Explain inferential logic in a broad way so that I can…
- Explain why I don't think autoethnography is the correct framing
- Advocate for embracing abduction first (and deduction second)

Inferential Logic

There are broadly three forms of logical inference

- Induction - we work from a set of examples and observations to generalize and explain some broad or more abstract concept. This is - in the most generic sense - the process of theory building. **Inductive inference is from specific to general.**
- Deduction - starting from a theory we state a hypothesis and design a test of the hypothesis. Our hypothesis is typically related to how a theory works in a particular environment, setting, population, etc. **Deductive inference is from general to specific.**
- Abduction - Incomplete observations can yield likely explanations (these explanations can then be tested deductively)

(see caveats in this comment)

"Cool Nic. WTAF does this have to do with our FaCCT paper?"

Ethnography and even auto-ethnography suggest the lab is engaged in a scientific endeavor that is inductive. I don't think that is true.

I think this lab is engaged in an exciting and important form of **computational abduction** (as well as deductive hypothesis testing)

Computational abduction allows us to short-circuit some of the inductive protocols that require building up, confirming, and testing alternative explanations that a responsible inductive process would require.

Here are some of my observations of the lab:
- Robert spins up a new infrastructure to design experiments.
- Using an existing dataset on hiring - Robert and Isaac show some promising early results of differences between OS and proprietary models
- Bingbing suggests the architecture is important - Bill proposes reconsidering experimental designs to test this explanation.
- Etc etc

> **This is not induction, even if we reflect on it individually inspired by auto-ethnographic methods.**

This lab is making (abductive) explanatory arguments about models - and it is doing so by holding some external variables constant (time, money, etc) and testing hypotheses about model variations under those constraints. Learning of failures and successes begets new questions that can improve a deductive research design.

How do we bring this into a lab process and explain it in FaCCT paper?

The part of auto-ethnography that I think is more salient is the concept of reflexivity.
Reflexivity is one of the hallmarks of all good inquiry.
Reflexivity is what separates scientific practice from Taylorism: We don't just specify a hypothesis and rotely execute experimental protocols. We actively interpret, refine, and adjust the design of an experimental protocol based on values, judgements, commitments to objectivity, etc.

Reflexivity is critical to the research and evaluation this lab is engaged in for a number of reasons:
1. Models can be black boxes
2. Temporal validity (understanding of emergent model behavior can change rapidly)
3. Pace of field (related to temporal validity, but new models are being released constantly)
4. Harms and biases are diffuse - they are not isolated in one model or one type of interaction

My proposal

- Let's find a way to describe how we can inject reflexivity into research practices that are inherently NOT inductive (whether you buy my description about abduction or not)
- I'm inspired by this description of computational grounded theory (which does something similar for the inductivism)
  https://journals.sagepub.com/doi/full/10.1177/0049124117729703