Doing Research in the AI era

Jon Kleinberg

Cornell University



Part 1: Context

What you think is happening, is happening.

- Our sensible aversion to hype
- The temptation to reason from what we wish were true
- The uncertainty level is very high
- A way of thinking about the uncertainty: discuss scenarios without trying to assign them probabilities.

Disruption, anxiety, and human costs

- What's happening feels unfair, for good reasons.
- The effect lands differently at different career stages.

Part 2: Goals

The goal has remained the same: do research that has an impact, in a responsible way

- In other words: use AI tools to do the best research you can
- Incentive systems, conferences, journals, awards are all in service of this goal (not vice versa).
- This "goal" doesn't answer anything; it's just a new question: What does high-impact research look like when you have access to powerful AI tools?

Two reasons we do science/engineering research:

- (A) achieve concretely specified targets
- (B) advance human understanding of the world
- Our two main modes of work arguably look like

$$.8 * (A) + .2 * (B)$$

$$.2 * (A) + .8 * (B)$$

An agenda: Pick the most ambitious problem you can, in either of modes (A) or (B)

- At the moment, AI can't do either (A) or (B) on its own.
- But it can take you much further than you'd get on your own.
- Example:
  (one of many, this a holy grail of the math. social sciences):
  Can we derive macro-level social phenomena from models of individual-level behavior?

What's a sign that your work is having impact in academia?

- People tell each other about your work even when you're not around.
- People build on your work even when you didn't ask them to.

What does that suggest about how to allocate your effort with AI?

- Writing as many papers as possible will probably be ineffective.
- Maximize the chance that something you do gets people's attention.

Part 3: Skill Sets

Pushing skill sets down to earlier career stages.

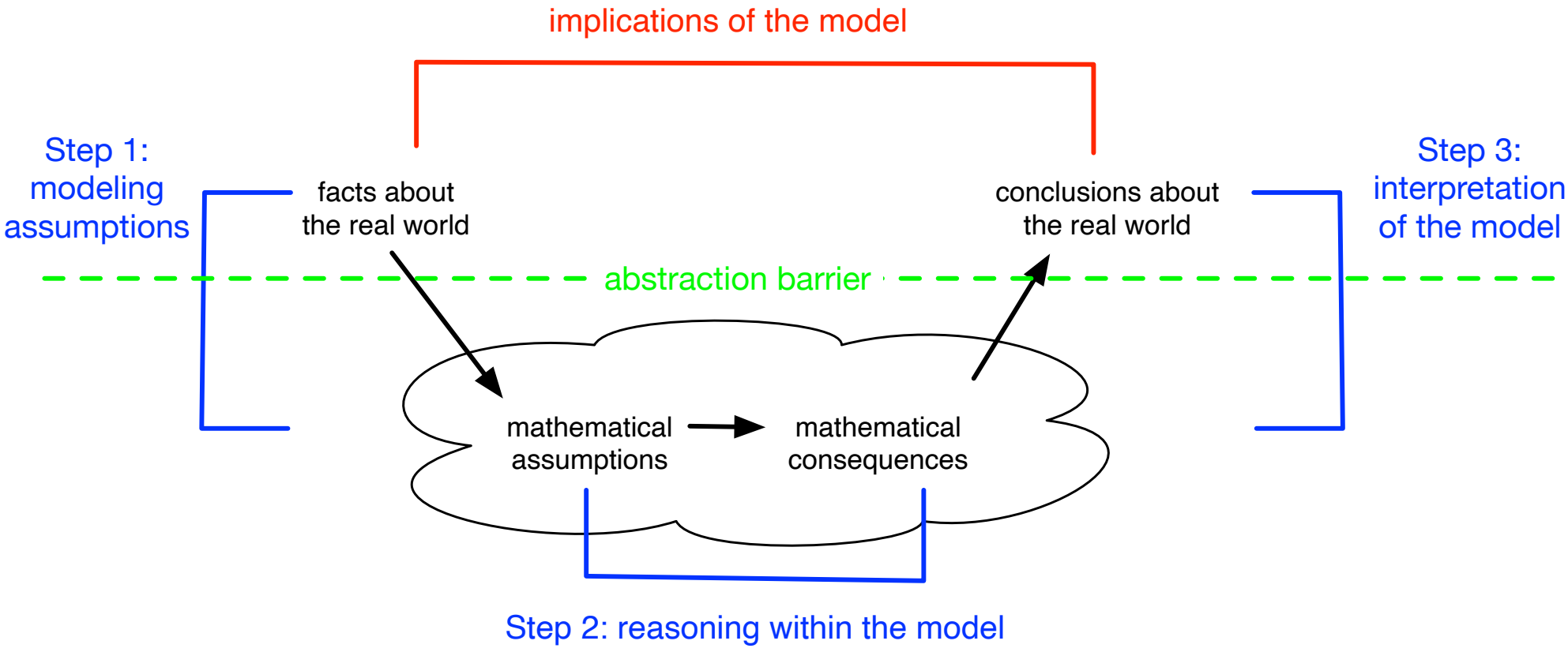
- 1970: Graduating PhD might have had one paper that was just published.
- 2015: College juniors might have a first-authored NeurIPS papers.

Pushing skill sets down to earlier career stages.

- 2015: Newly tenured professors might be just starting to run a large research group.
- 2026: First-year PhD students have the ability to act as leaders of research groups.

The proficiency-is-assumed principle

- 1970: PhD theses in CS contained pages and pages of source code.
- 2015: No one wanted to see your source code.
- The question is what you can do with your newfound proficiency.
- Similar principles in music, in visual art, in accounting, ...


implications of the model
Step 1:
modeling
assumptions
facts about
the real world
conclusions about
the real world
Step 3:
interpretation
of the model
abstraction barrier
mathematical
assumptions
mathematical
consequences
Step 2: reasoning within the model

Part 4: The Situation in Math and Theoretical CS

What you think is happening, is happening.

- Math research has been irrevocably transformed.
- Entirely reasonable objections to how AI companies rolled out this capability, but the change itself was arguably inevitable.

We've known since 1900 that math was mechanizable

- Math is a tree with the axioms at the root.
- From each internal node, apply each possible inference to get the node's children.
- This is all the theorems that can ever be proved.
- A mechnical wind-up toy can solve all of mathematics.

Math research is therefore about the curation of theorems (and definitions)

- Theorem: $27 \times 43 = 1161$.
- Theorem: If I color the points of the plane in two colors, there will be two points an inch apart that get the same color.
- The mechanical wind-up toy does not one of these two theorems as more interesting than the other.
- So far, curating the definitions and theorems is still a human activity
- Maybe: this is something that humans will continue to do well at even as AI keeps improving (more on this later).

Maybe: human math research might also use AI to try doing things that would otherwise take a century

- Impossibility of trisecting an angle with compass and straight-edge
- The $3x + 1$ problem: "Math is not yet ready for such questions" (Erdos)
- But maybe this will be too hard for us.
  And maybe the AI won't need us for this activity.
  (More on this later.)

Part 5: Chess as Model System for Superhuman AI

What you think is happening, is happening (1990s version):

- 1988: First win by chess engine over human grandmaster in tournament.
- 1997: Deep Blue defeats world champion Garry Kasparov.
- 2005: Last time a human player defeated a full-strength chess engine.
- 2014: End of the centaur era in chess.
- 2018: AlphaZero achieves superhuman performance purely via self-play.

Stylized facts we learned from superhuman AI for chess:

- Caveat: chess $\neq$ math $\neq$ science $\neq$ healthcare $\neq$ ...
- Chess is more popular then ever
- Human chess-playing style has been influenced by engines
- The spectators have been dramatically empowered
- The chess engine gives you the illusion that you know what's going on

Part 6: What we can learn from the chess analogy

Some analogies between chess and math/theory/type-(B) research:

- Human-AI collaboration on concretely defined math problems might have an expiration date in the medium-term future.
- Our notions of "elegance" in math arguments might change.
- This change might be generational, not psychological.
- Math might acquire spectators who feel empowered.

One more analogy: The aesthetic ceiling

- Conundrum: what if AI becomes superhuman at formulating elegant theories of the world?
- Precedent from chess: chess engines can play beautiful games, but people don't prefer them to beautiful human games.
- Maybe for research aimed at human understanding,
  there is a ceiling that humans are already near.

Part 7: Our everyday workflows

Performance versus practice

- Thus far: doing research and practicing skills have been almost the same activity.
- This is changing; analogy is now closer to music or sports
- Involves believing that strengthing basic skills is still important.
- Once again: at what level of abstraction do we assume proficiency? (E.g. analogy to machine-language programming with birth of compilers)

Working with co-authors

- It's easy to miss how much we develop a theory of mind of our co-authors.
- If we're not transparent about our AI use in collaborations, it messes with this theory of mind.
- When we're transparent, we have AI as one (or more) additional collaborator, and we need a theory of mind for it too

How we communicate our research

- No one seems to believe that packaging research into discrete papers represents the future
- Work on larger things
- Disseminate the research as more of "living object"
- Still not clear what this will look like, or how rapidly to innovate and experiment with the form