AI Snake Oil
Exercises and discussion prompts

October 10, 2024

We plan to update this document periodically. If you are an instructor who uses the AI Snake Oil book in your course, we'd love to hear from you. Please feel free to suggest questions or discussion topics.

Chapter 1: Introduction

Fact checking Moravec's paradox.

Moravec's paradox is the observation that tasks that are hard for people, such as playing chess, are easy for computers, and tasks that are easy for people, such as vision or walking, are hard for computers.

This statement tends to have a strong and immediate intuitive appeal, and is taken as a truism by AI researchers. Yet, to our knowledge, it has never been formulated precisely enough to be testable or falsifiable. So let's try to do that.

What do "easy" and "hard" mean in the context of AI? Does it refer to computational requirements, or the human effort needed to build AI to perform a task, or something else? And what does easy/hard for people mean?

Based on your definitions of these terms, pick a variety of tasks and try to place them on a 2-dimensional spectrum where the axes represent people's and computers' ease of performing the task. What sort of relationship do you see?

Some points to keep in mind:

- What counts as a task? How do you select tasks to analyze? In discussions of Moravec's paradox, people often forget about tasks that are "easy" for both people and computers, such as detecting how bright an image is. Similarly, there are an endless number of tasks that are "hard" for both people and computers. Of course, if you eliminate two opposing quadrants in a 2x2 figure, the relationship between the remaining points will appear to be negative!
- When Moravec's paradox was formulated back in the '80s, people thought reasoning was "easy" because it could be solved by symbolic systems of the time. Unfortunately, while those systems work well in closed, toy domains, they lack common sense and struggle in the real world. Today's reasoning systems rely on generative AI, so it's unclear in what sense reasoning is easier than, say, image classification.

Additional remarks about this question:
We don’t know if Moravec’s paradox has any merit to it. But we strongly suspect that it’s a just-so story that helped AI researchers make sense of the perceived failures of the early decades of AI. Moravec offered a theoretical explanation for the paradox based on the difficulty of reverse engineering human skills, but this explanation did not anticipate the role of machine learning in AI. Today, reverse engineering plays only a minor role in AI and researchers often eschew it in favor of learning from examples even in domains where it is possible.

Our book is about what AI is and isn’t good for. But we are *not* trying to perform the kind of analysis that Moravec’s paradox hints at, and we aren’t optimistic about finding any criterion that can help us anticipate which as-yet-unsolved tasks will be easy or hard for AI. (We’d love to hear from you if you think there is such a criterion that’s supported by evidence — whether Moravec’s paradox or something else.)

When we talk about limits to prediction (primarily in Chapter 3), we mean intrinsic limits to the predictability of certain phenomena, regardless of whether it’s a person or AI doing the prediction. And we talk plenty about areas where AI shouldn’t be used for moral reasons *even if* it can perform accurately.

Desirable types of AI

The text gives many examples of AI that quietly work well, like spellcheck. Can you think of other examples? What do you think are examples of tasks that AI can’t yet perform reliably but one day will, without raising ethical concerns or leading to societal disruption?

Controversial applications

Figure 1.2 presents a framework to understand AI snake oil, hype, and harms. Pick a few AI-related controversies and discuss where in the figure you think they should be placed. For example: Levi’s used AI-generated clothing models to show more diverse body types and skin tones; DoNotPay was alleged to be engaged in the unauthorized practice of law by advertising a “robot lawyer”; an AI model was claimed to predict crimes with 90% accuracy.

Chapter 2: How predictive AI goes wrong

Case studies of failures

Two case studies from the UK are relevant to the themes of this chapter: The Post Office scandal and the 2020 A-levels grading controversy. Look up and read analyses of what went wrong in each case.

What, if any, lessons from the chapter are relevant? Note that AI wasn't involved in the Post Office IT system and it's debatable if the A-levels grading algorithm can be called AI.

Common sense

Predictive models make "common sense" mistakes that people would catch, like predicting that patients with asthma have a *lower* risk of developing complications from pneumonia, as discussed in the chapter. What, if anything, can be done to integrate common-sense error checking into predictive AI?

Gaming

Think about a few ways people "game" decision-making systems in their day-to-day life. What are ways in which it is possible to game predictive AI systems but not human-led decision making systems? Would the types of gaming you identify work with automated decision-making systems that do not use AI?

Hiring automation

In which kinds of jobs are automated hiring tools predominantly used ? How does adoption vary by sector, income level, and seniority? What explains these differences?

Chapter 3: Can AI predict the future?

Forecasting methods

For forecasting each of the following phenomena, which methods are used — machine learning, simulation, or something else? How has the mix changed over time?

- The weather
- The progression of infectious diseases
- The demographics of populations
- Elections

Meteorologists

Weather forecasts have long been automated. Why, then, does the job of operational meteorologist exist (as opposed to research meteorologists, or weather presenters who pretend to be meteorologists on television)? What do operational meteorologists do?

Ethics

Suppose a research group at a big tech company finds that it can build a model to predict which of its users will be arrested in the next year, based on all the private user data that it stores, such as their emails and financial documents. While far from perfectly accurate, it is more accurate than any model that uses public data alone.

- Does it seem plausible that a model like this might work in any meaningful sense? If so, what signals might it be picking up on?
- What laws, rules, or norms should govern companies' ability to undertake research projects of this sort?
- Is there any ethical and responsible way in which technology like this can be put to use, or should we as a society reject such uses of prediction?
- What, if anything, prevents a company from partnering with police departments in your country to use such a predictive model for surveillance of individuals deemed high risk?

Chapter 4: The long road to generative AI

Deepfakes

Clone your own voice and video using a commercial product. How good is it? Show a friend a real video of you talking, and a cloned video, and see if they can tell which is which.

Learning

Spend at least an hour using a chatbot for learning. Here are the guidelines:

- You can go in depth into one topic or pick a few different topics. But don't scatter yourself too thin.
- Use a state-of-the-art chatbot.
- Pick topics that you have actually been planning to learn, so that you have a stake in the outcome.
- The chatbot should probably not be your sole resource and you might want to have a process in place for verifying its outputs.
- You might find it helpful to peruse strategies for how to use chatbots for learning.

Reflect on your experience and discuss it with your peers. What worked well, and what didn't? Do you plan to continue to use chatbots for learning?

Annotation work

AI companies are notoriously secretive about the labor that they employ for data annotation. How would you estimate the number of workers in the world currently engaged full time or part time in AI annotation work?

The use of creative work for training

Generative AI is built using the creative output of journalists, writers, photographers, artists, and others — generally without consent, credit, or compensation. Discuss the ethics of this practice.

How can those who want to change the system go about doing so? Can the market solve the problem, such as through licensing agreements between publishers and AI companies? What about copyright law — either interpreting existing law or by updating it? What other policy interventions might be helpful?

Next-word prediction

Discuss why large language models trained to accurately predict the next word in a sequence of words end up exhibiting a range of other capabilities. Read the research on this topic and try to summarize it in an intuitively accessible way.

Environmental impact

Discuss the environmental impact of generative AI.

- Start with the current impact. Keep in mind that in addition to energy, data centers require water, land, metals and minerals. Consider both the global impact (e.g. the impact of energy consumption on climate) and local impacts (water use, environmental degradation of mining sites).
- How might this change in the future? There are many unknowns — rate of AI adoption, scaling trends, energy efficiency, resource depletion, regulation, and more. Come up with a few scenarios.
- What do you think about the argument that AI use will substitute for activities that use even more energy, just as videoconferencing technology sometimes substitutes for air travel? What about the argument that advancing AI will help solve environmental problems, such as through better disaster prevention or making traffic routes more efficient?
- What, if anything, is distinct about AI's environmental impact compared to computing in general or other specific digital technologies with a large energy use such as cryptocurrency?
- How should policymakers respond to the environmental impacts of AI?

Chapter 5: Is advanced AI an existential threat?

Policy making

In AI safety policy, entrenched camps have developed, with vastly divergent views on the urgency and seriousness of catastrophic risks from AI. While research and debate are important, policymakers must make decisions in the absence of expert consensus. How should they go about this, taking into account differences in beliefs as well as values and stakeholders' interests?

Forecasting

Make predictions on the forecasting website Metaculus on a few AI- and AGI-related questions. Be sure to read the "resolution criteria" carefully. What data or information did you consider? What do you think of the community predictions? Discuss with your peers.

Recall the difference between foxes and hedgehogs in Philip Tetlock's view of prediction experts. Why do you think forecasters that integrate information from many domains ("foxes"), are better at predictions than forecasters who are experts in one domain ("hedgehogs")?

Automating AI research

As of 2024, there have been a few attempts to automate AI research. Read some of this work. What set of activities are researchers trying to automate? Assess how close they are to their goal. What are the implications of being able to automate AI research?

Chapter 6: Why can't AI fix social media?

General-purpose AI

Recently there has been a lot of excitement about using general-purpose AI models, notably large language models and vision-language models for content moderation. As of 2024, however, the use of these technologies remains relatively niche, and purpose-built models for violations such as hate speech or incitement to violence are still the norm.

- The chapter discusses many limitations of AI for content moderation. Which, if any, of them can be mitigated by shifting to general-purpose AI models?
- What new risks or limitations would the use of general-purpose models introduce?
- The use of general-purpose models for content moderation tends to be more attractive to smaller platforms than larger ones. Why do you think this is?

Content moderator autonomy

What are the advantages and disadvantages of having a content moderation system with highly specified rules? What do you think about a content moderation system that gives more autonomy to content moderators?

Algorithmic choice

Algorithmic choice is the idea of allowing users to choose or personalize their own recommendation algorithms, perhaps even through a marketplace of such algorithms through which anyone can distribute their own algorithms.

- Of the many ills that have been blamed on recommendation algorithms, which ones could algorithmic choice conceivably combat? Which ones are structural and can't be solved through the lens of individual empowerment?
- As of 2024, none of the major platforms offer meaningful algorithmic choice. Why do you think that is? What do you think are the technical impediments to doing so? In what ways do you think it's against their commercial interests?
- If platforms gave you algorithmic choice, what changes would you make to the algorithms behind the feeds that you consume?

Your experiences

Discuss your own encounters with content moderation as a social media user or creator.

- Have you ever received any warnings from platforms? Had content removed?
- Were you convinced by the reasons, did you think the platform made a mistake, or did you think the policy was too onerous?
- Have you ever suspected that you were being shadowbanned? What do you think about shadowbanning as a policy?

The Haugen documents

Facebook/Meta has been accused of many things, including: allowing problematic content to proliferate in order to profit from the engagement that it generates; a bias against conservatives; and suppressing internal evidence of the harms of its platforms, such as Instagram's impact on teen girls' mental health.

The Haugen documents, commonly known as the Facebook papers, are a trove of internal documents leaked by whistleblower Frances Haugen. You can read them here.

Spend an hour or two browsing the documents, generally. No need to look for any specific information. Try to form a high-level impression of how developers at Facebook think about the impacts of their platform and their responsibility. What motivates them? Which of the concerns listed above seem to be supported by these documents? What do you think about the quality of the ethical reasoning in the documents?

The future

Do you think it's possible that recommendation algorithms can be far more accurate in the future (in the narrow sense of predicting how likely a user is to engage with a piece of content), or are there fundamental limits that will prevent this? For this thought experiment, imagine that platforms will have access to more invasive data about users than they do today, and that computational limits aren't an issue.

Chapter 7: Why do myths about AI persist?

AI versus other fields

One difference between AI research and other kinds of research is that most AI research is purely computational, and doesn't involve (for instance) experiments involving people or arduous measurements of physical systems. In what ways does this make it easier to have confidence in the claims of AI research? In what ways does it make it harder?

Impact on professionals

The chapter argues that we shouldn't read much into the performance of AI on professional licensing exams. A better way to understand the impact of AI on occupations is to the productivity and satisfaction of professionals with and without AI assistance. Many such studies exist. Pick an occupation, find the relevant studies, and discuss their implications. If the studies seem to contradict each other, what might explain the differences in findings?

News reporting

Analyze a few news articles on AI using our checklist of eighteen pitfalls in AI journalism.

What techniques do you personally use to stay grounded when you hear of seemingly amazing AI advances in the news? Discuss with your peers.

Accountability

What are some ways to improve accountability for companies making unsubstantiated claims? These could include legal remedies as well as non-legal approaches.

Read about the U.S. Federal Trade Commission's crackdown on exaggerated AI claims. What are the types of exaggerations the FTC can go after? What are the limitations of the FTC's authority?

Chapter 8: Where do we go from here?

Broken institutions

The chapter makes the point that broken AI appeals to broken institutions. What are some examples of broken institutions enamored by other dubious technologies? Is there something about AI, as opposed to other technologies, that makes it liable to be misused this way?

Partial lotteries

How would you feel if a college application, or a grant proposal, or another valuable opportunity to which you applied were decided by a partial lottery? What if you were in a world where this was the norm rather than the exception?

Regulation

Look up some examples of AI-related legislation or regulation recently enacted or being debated in your country. Discuss the pros and cons of specific actions and proposals, as well as the overall approach to AI policymaking.

Impact on occupations

What impact do you think AI will have on your chosen or intended profession in the next five to ten years? What levers do we have to steer this impact in a way that is positive for society?

Personal decisions

Having read the book, are there any changes you plan to make in your life or career?

Bonus

Once you’ve finished answering / discussing the above questions, use a chatbot to answer some of them. You might have to modify the prompts to give the bot instructions on what constitutes a good answer; upload papers or documents for it to draw from; and ask follow-up questions. Compare and reflect on the two processes of learning (with and without chatbots). Did the chatbot provide useful insights that you had not thought of?

We are grateful to Zachary S. Siegel and Nitya Nadgir for suggesting questions for discussion.