

Defining Ethics + Project 2 intro

15-113 – Monday 9/28

We'll (anonymously) share these results

- Show off and discuss HW4 with your neighbor and then fill this out! <https://tinyurl.com/55edce6n>



Onward: Project 2!

This is very open-ended (though it overlaps the last few projects)

- Purpose: You should build something you're proud of *and* can *comfortably* talk about in interviews!
- This is a 1.5-week project. Be detail-oriented and make it nice! (And don't forget the midpoint check-in)
- Due before Fall Break (probably 10/8 to leave time for 1-on-1 presentations)

Get creative!

Just note that we'll be looking for about 9 hours of focused *new* work for this project, ideally building something entirely fresh. (If you really want to improve something you already have, it must be *transformative* enough to feel new.) We'll be looking for indications that you've learned something and are using AI more effectively to achieve greater goals than you could at the beginning of the course.

HW5

Just two things to read.
Take notes, and be ready to discuss!
(No explicit deliverable to hand in)

- “Cognitive Helmets for the AI Bicycle” by Cat Hicks
- “The Evitable Conflict” by Isaac Asimov

Some preliminaries before discussing ethics

- Mike is not a philosopher! (But I spend a lot of time worrying about things)
- The goal here isn't to teach you right and wrong; it's to help you navigate the borderlands for yourself
- Much of this is based on what I learned from Prof. Illah Nourbakhsh, who will be joining us for a future lecture
- I highly recommend his book "Robot Futures" (I'll order some extra copies for my office soon because I keep lending them out / giving them away)

Please participate!

Voice your thoughts, ask questions! This is only going to be valuable if you're willing to add to it.

- It's ok (and probably even good) to be uncertain about what you know or feel or think. Voice that too!
- I know you'll all respect each other. On top of that, let's add a house rule for anonymity: If you tell anyone about our discussions outside of class, please don't attach names to anyone's comments.
- On my end, I promise to represent your interests to the university, other faculty, etc.

What is ethics?

- Academic definition: The study of moral phenomena
- Casual definition: How do we make important decisions? How do we decide what's right and wrong, worth it or not worth it? How do our decisions (or indecisions) affect others? What should we prioritize, and what are we willing to sacrifice for a particular end?

What is ethics?

- Ethos: “the distinguishing character, sentiment, moral nature, or guiding beliefs of a person, group, or institution” (Merriam Webster)
 - What is our nature?
- Mores: “the fixed morally binding customs of a particular group” (MW)
 - How does that nature manifest in our decisions?
- Taken together: What does responsible conduct look like?

Is “AI Ethics” the right term here?

- AI doesn't advance without technologists
- AI doesn't act on the world without users
 - (individuals, corporations, governments, etc)
- Deliberately or not, an AI's perceived “ethics” flow from us
- So, instead of framing tough questions as “What should an AI do or not do in this situation?” I think we should take a more anthropocentric approach:

Ethics of humanity in the context of AI

Questions we should be asking

- What should we demand from an AI system?
- How do we use AI while staying consistent with our values?
- How do we maintain accountability for indirect consequences?
- What are our obligations to humanity and the environment?
- What are our obligations to ourselves as individuals?
- What should we prioritize when we have the opportunity to change the world?

What can we actually do, though?

- None of us have the ability to simply reach out and turn all the knobs of the world (bummer). You *can* reach some of them though, and you'll reach more in the future.
- "Give me a lever long enough and a fulcrum on which to place it, and I shall move the world." (Archimedes)
- KEEP BUILDING YOUR LEVER
- In the meantime, find your ethical algorithm

My (personal) ethical algorithm

1. First, start at the individual.
 - a. What personal agency do I have in this moment?
 - b. Am I being true to myself and my values?
 - c. **Am I being deliberate in my decisions?**
2. Move to the community.
 - a. What spheres do I move within? Collectively, do these spheres act responsibly?
 - b. To what extent can I steer these groups (toward happiness, responsibility, safety, etc)
 - c. **Am I steering deliberately?**
3. Move to the present and future.
 - a. What ripples am I sending out into the world?
 - b. To the extent that I can predict, are my actions likely to steer the world to a happier and more equitable future?
 - c. **Am I deliberate in considering the broader impacts of my actions?**
4. **Be exemplary**
 - a. Be seen, allow people to learn from your successes *and* your mistakes

Next (depending on time)

- We'll focus more on how all of this applies specifically to AI
- We'll discuss what agency we have, and what agency we should work to gain
- We'll discuss several groups' attempts to define priorities for AI development and use
- My goal is for you to define your own priorities, and a deeper understanding of your own ethical algorithm

One video of Illah:

<https://www.youtube.com/watch?v=giKT8PkCCv4>

First Topic:

One of the major promises of generative AI is that it will make us more productive. For example, we'll be able to generate code faster than we could write it ourselves.

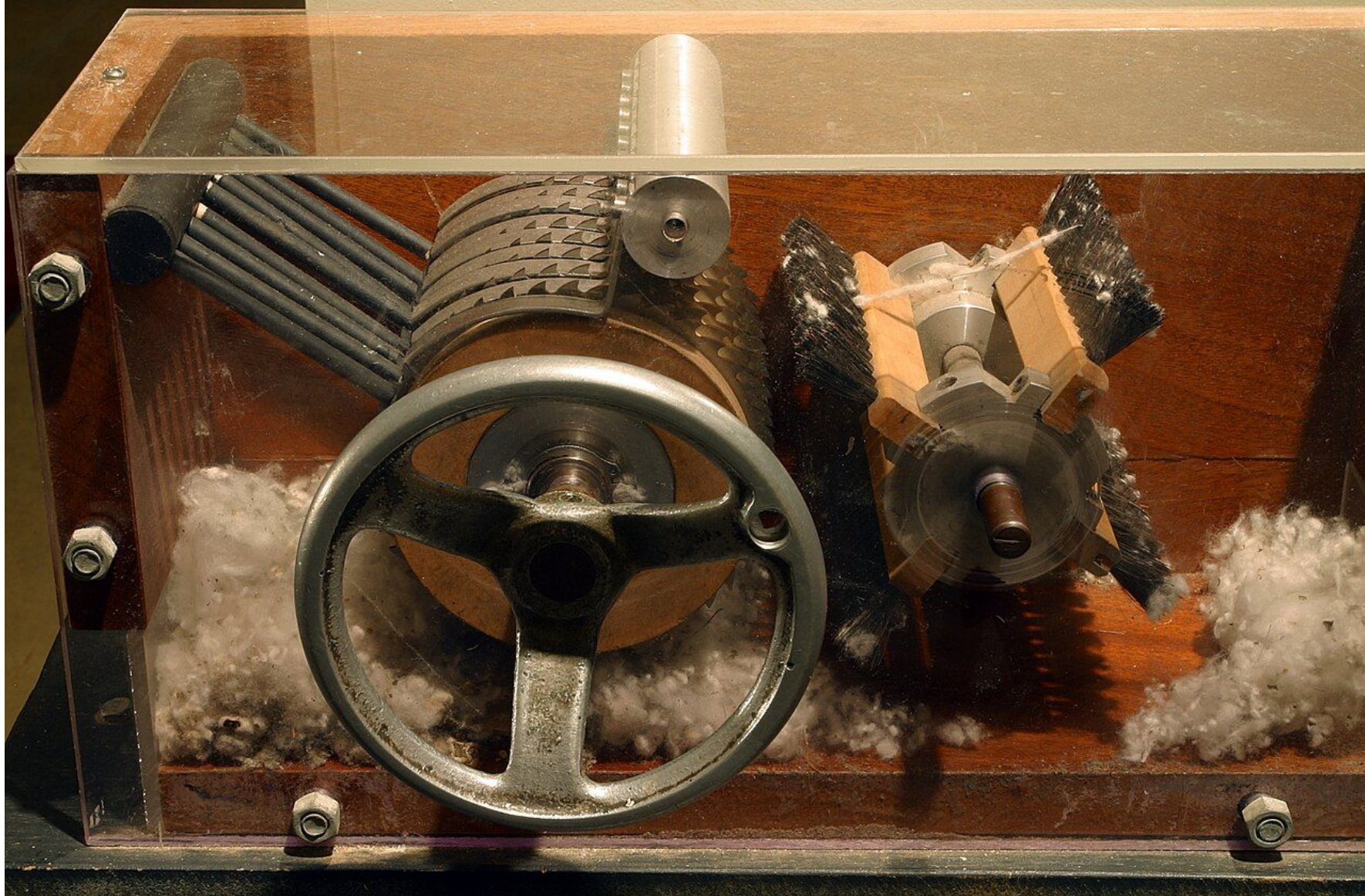
- Do you feel this is already true?
- Do you feel that it is likely to be true in the near future?

For the moment, assume the standard of productivity is that the output “gets the job done,” and that we’re talking about low-creativity low-risk tasks. We’ll get to questions of safety, artistic merit, education etc later.

If modern *AI can* make us more productive...

- Who benefits from this?
- How might this change how much money you make, or how much free time you have?
- How might this change how you feel about the work that you do?

An aside: What's this?



Eli Whitney's cotton gin, 1793

- Function: Automates and *massively* accelerates cotton seed/fiber separation by 4900%
- Along with other industrialization, this created a bottleneck: harvesting the cotton
- Rather than easing the workload, a 'productivity tool' skyrocketed the demand for slave labor in the field, resulting in longer hours, worse conditions, and higher expectations
- Back to AI: Does this seem relevant? Or do you think this is too far in the past?

If time (or watch on your own)

- Excerpt from 60 Minutes on AI supervision and training (Nov 2024):
<https://www.youtube.com/watch?v=qZS50KXjAX0>
- CW: Corporate exploitation, poverty, descriptions of physical/sexual violence in the context of AI-generated imagery
 - General note: You are always welcome to step out of class if needed