The Turing Trap Handout (Gens-176-V, Feb 17 2026)

Erik Brynjolfsson is the author, and is the director of the Stanford Digital Economy Lab.

The following summarizes the first part of the video we'll be watching (from 2022)

`https://www.youtube.com/watch?v=r3-mudok27s`

- From Allen Turing, we get the "Turing test" (or *Imitation Game* (1949). In the test, human beings judge the transcript of a natural language conversation between an human being and a machine. The machine "passes" the test if the evaluator cannot reliably tell them apart.

- Interesting side note: Things like "CAPTCHA" boxes online are a type of *reverse* Turing test, where the objective is to identify "bots" and stop them from accessing data on a website.

- In the movie "Blade Runner", there is a fictional test called the "Voight-Kampff machine", which determines whether an individual is a replicant, not by what they say, but what they *feel*: it measures respiration, blush response, heart rate, and eye movement in response to questions dealing with empathy.

  Some suggest that this test is more like what we would need to determine consciousness- not by what the machine is saying, but by what phenomena are occuring.

The 10 Theses of the Turing Trap

1. The benefits of human-like AI are enormous.

2. Not all types of AI are human-like.

3. The more human-like a machine is, the better substitute it is for human labor.

4. Labor substitutes (automation) tend to drive down wages.

5. Substitution can reduce the economic and political power of those replaced.

6. Taking away power and agency creates a trap. (This is the Turing trap)

   What happens in this scenario is that the economy is stuck in equilibrium. However there is a different path:

7. Alternatively, AI can complement (or augment) labor.

8. In our history, technology has been used to amplify human power, which increases its value.

9. Augmentation creates not just new capabilities, but also new goods and services.

10. Today, there are excess incentives for automation instead of augmentation.

Note that the speaker says that there were predictions when an AI would win gold in a Mathematics Olympiad. In 2021, they said rather than it happening in 2041, it would happen in 2031 (it actually happened this year):

> *Google Gemini's Deep Think and something experimental from OpenAI both achieved gold medal-level performance at the 66th International Math Olympiad in July 2025. They both generated natural language proofs that were praised as clear, precise and easy to follow.*

Going back to the talk:

- Thinks of AI as General-Purpose-Technology (like the steam engine).
- For economists, as labor hours gets small, productivity goes up.
- But, if labor hours go to zero, what happens to labor income?

  This is not hypothetical- it's happening now (chart using education vs wages)

  This is not restricted to wages, but also things like life expectancy (suicide rates among white, no college education have been growing very large).

He brings in three principles of Knowledge, Wealth and Power:

P1 Most "useful knowledge" is inherently decentralized across millions of human brains.

P2 When knowledge is codified(*) and digitized, becoming alienable and possibly centralized.

Codified meaning that experience-based understanding is turned into explicit, shareable forms. Alienable meaning able to be removed or given to others.

P3 Concentration of economic power begets concentration of political power.

(We'll stop after about 20 minutes to discuss)

For Discussion

1. Discuss the measurement problem: How should we measure "success" in AI if not by human imitation?
2. A question for education: What skills should schools and universities emphasize in an AI-rich world?
3. Policy: Should governments encourage augmentation over automation? How?
4. And going with policy, comes equity: Who gains and who loses when AI **replaces** human labor?
5. Ethics: Is replacing humans always a bad outcome—or only in certain contexts?

Takeaway Message

The most important question about AI is not “Can machines replace humans?” but rather “How can machines help humans do more?” Avoiding the Turing Trap means designing technology—and institutions—that place human capability, dignity, and shared prosperity at the center.