Lecture 12: Philosophical perspectives

Artificial Intelligence
CS-UY-4613-A / CS-GY-6613-I

Julian Togelius
julian.togelius@nyu.edu

Overview?

- Can a computer be intelligent? Common philosophical stances and arguments
- The ethics of AI: risks and possibilities
- Bonus: classical sci-fi movie references!
- NB: this is an actual lecture, with an actual book chapter as assigned reading and probable corresponding exam question

Strong versus weak AI

- Weak AI: We can create computers/software that can competently execute or simulate any aspect of intelligence, and solve any problems we can solve

- Strong AI: We can create computers/software that is actually intelligent, can feel and has a consciousness

Philosophical stances on the mind and the body

- Dualism: The material brain stands in correspondence with (or is controlled by) an immaterial soul

- Functionalism: The mind is determined by (or is) the functional relation between output and input

- Physicalism: The mind is determined by (or is) its physical realization, as biological neurons or perhaps silicon chips

Arguments

Numerous arguments have been advanced against both strong and weak AI

The argument from disability

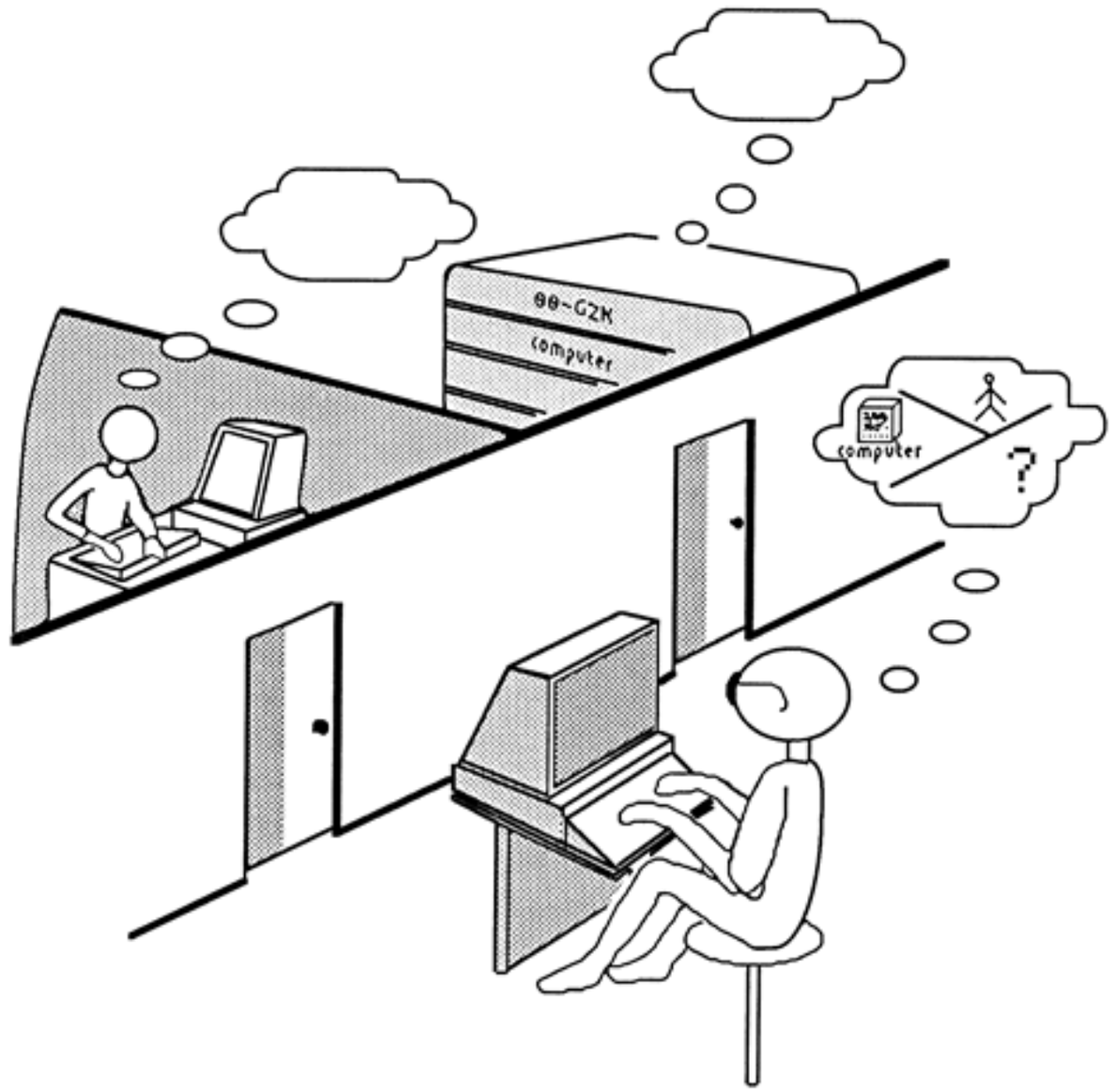
- “A machine could never do X”
- “Fall in love, enjoy strawberries and cream…learn from experience, use words properly, be the subject of its own thought…”
- Moving goalpost problem of AI
- What about people lacking certain abilities?

The informality argument

- Most tasks are not specified well enough
- We cannot describe everything in first-order logic
  - e.g. tacit knowledge, intuition…
- But is this really an argument against AI?
- Or simply against a particular type of “good old-fashioned AI”?
- Would a real AI need to be embodied?

The Turing Test

- Human interrogator can chat with a human and a computer
- Must decide which is which
- Is a computer intelligent if it can fool the interrogator that it is human?

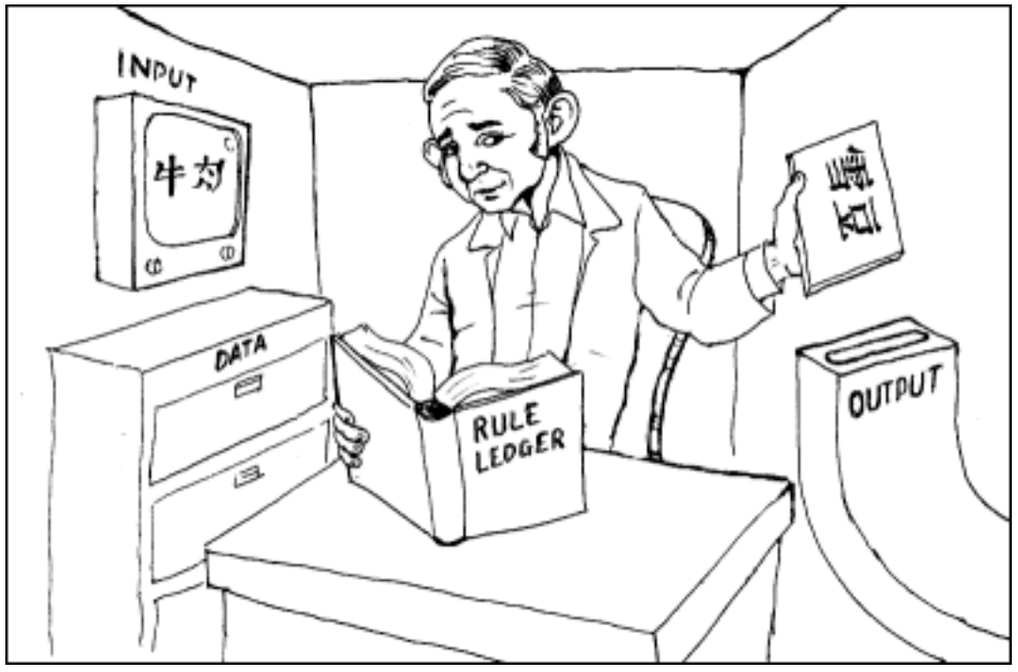


http://www.aisb.org.uk/events/loebner-prize

The Chinese Room

- A non-Chinese speaker sits in a closed room, gets paper slips with Chinese text

- Follows instructions in a book to select other Chinese text snippets to return



- Thus Chinese questions are answered

- Does this mean the system (person+book) understands Chinese?

The Chinese Room

- A computer is just a symbol-manipulation engine, so how could it be intelligent?

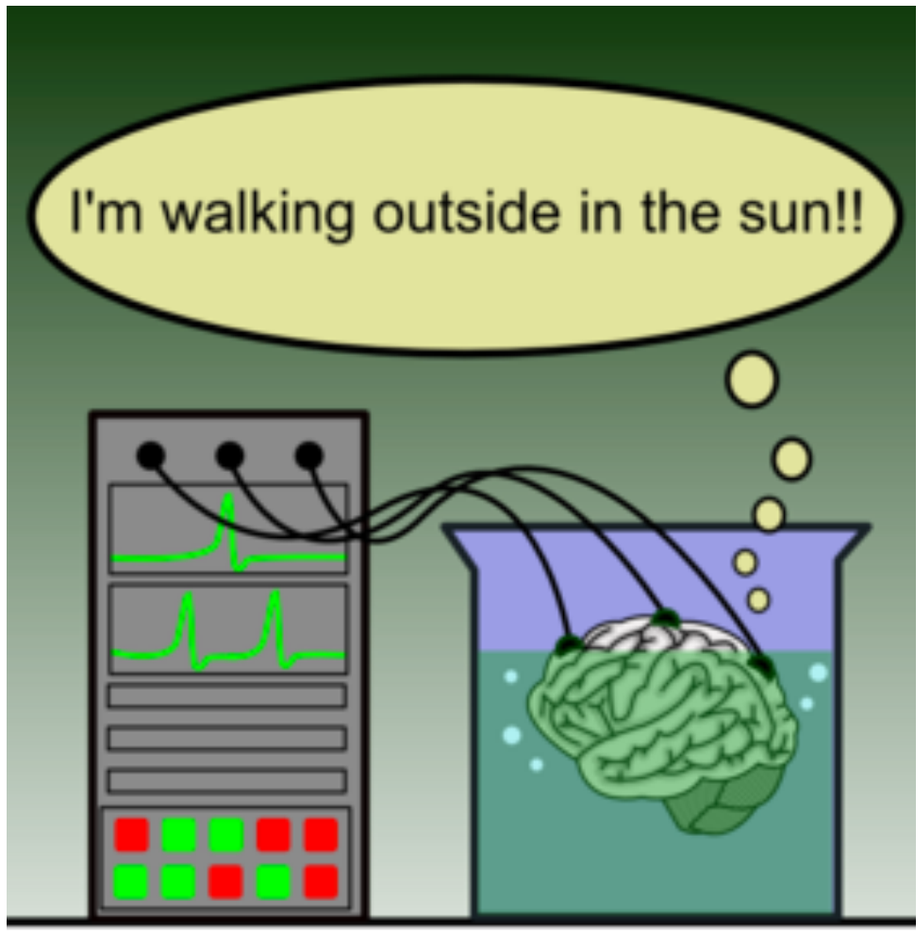
- Where is the understanding? Where is the intentionality?

The brain in a vat

- Your brain is removed from your body and placed in a vat

- Cables that are attached to your brain feeds sensory impressions generated by a computer

- Are your experiences real?

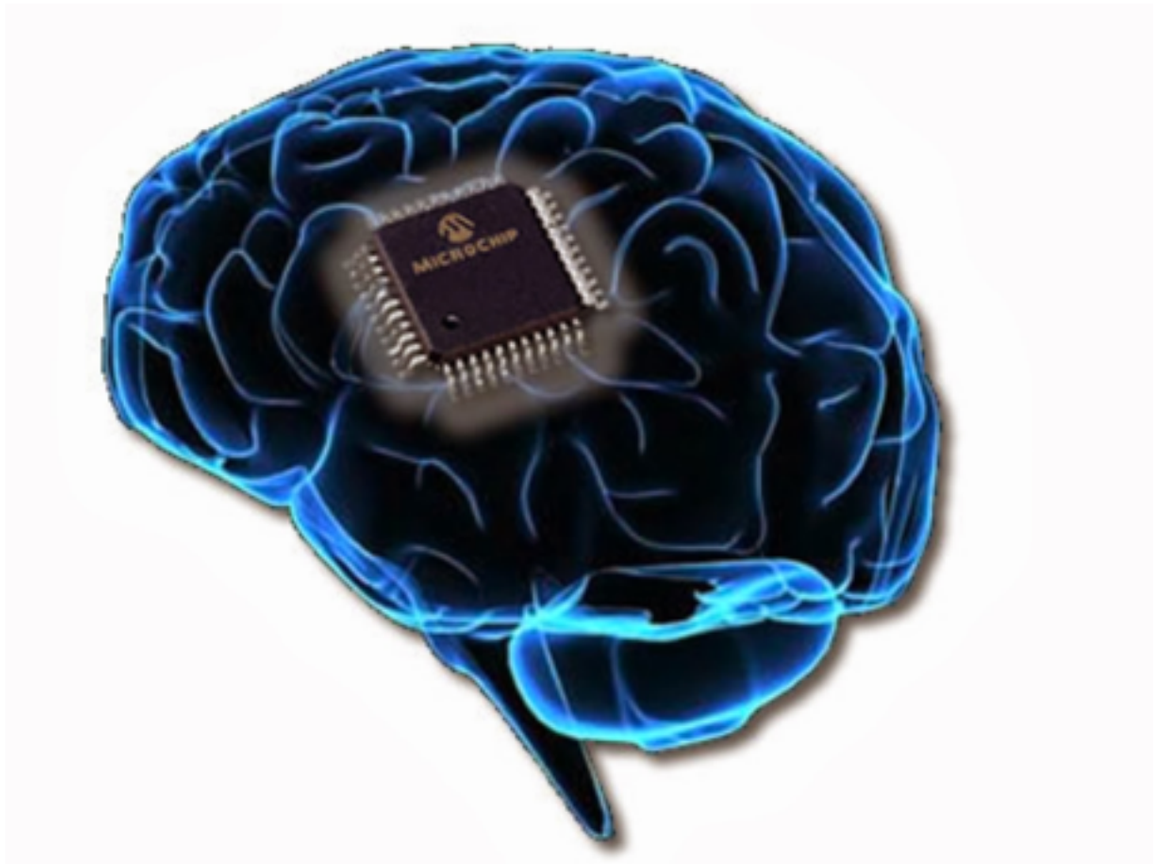


Brain replacement

- Part by part, your brain is being replaced with digital circuitry

- Every replacement of neurons with a chip retains the original functionality

- Are you still the same person? Are you still conscious? If not, when did you stop being you?



Ethics of AI

- Surveillance
- Employment
- Loss of control
- Superintelligence

Surveillance

- Ubiquitous computing and networking makes massive data collection possible (inevitable?)

- AI / machine learning methods make it possible (inevitable?) to analyze all this data and draw conclusions

Gattaca (1997)

Employment

- First and second industrial revolution: mechanization replaced physical labor, workers moved to more intellectually demanding jobs

- Early 21st century: a “service economy” (in the west)

- Will AI replace our need for service workers?

“It was a tough decision, but I’ve decided on which one of you I’m going to hire.”

Employment

Truck and taxi drivers, nurses, nannies, accountants, lawyers, psychologists, game designers, researchers…?

Loss of control and accountability

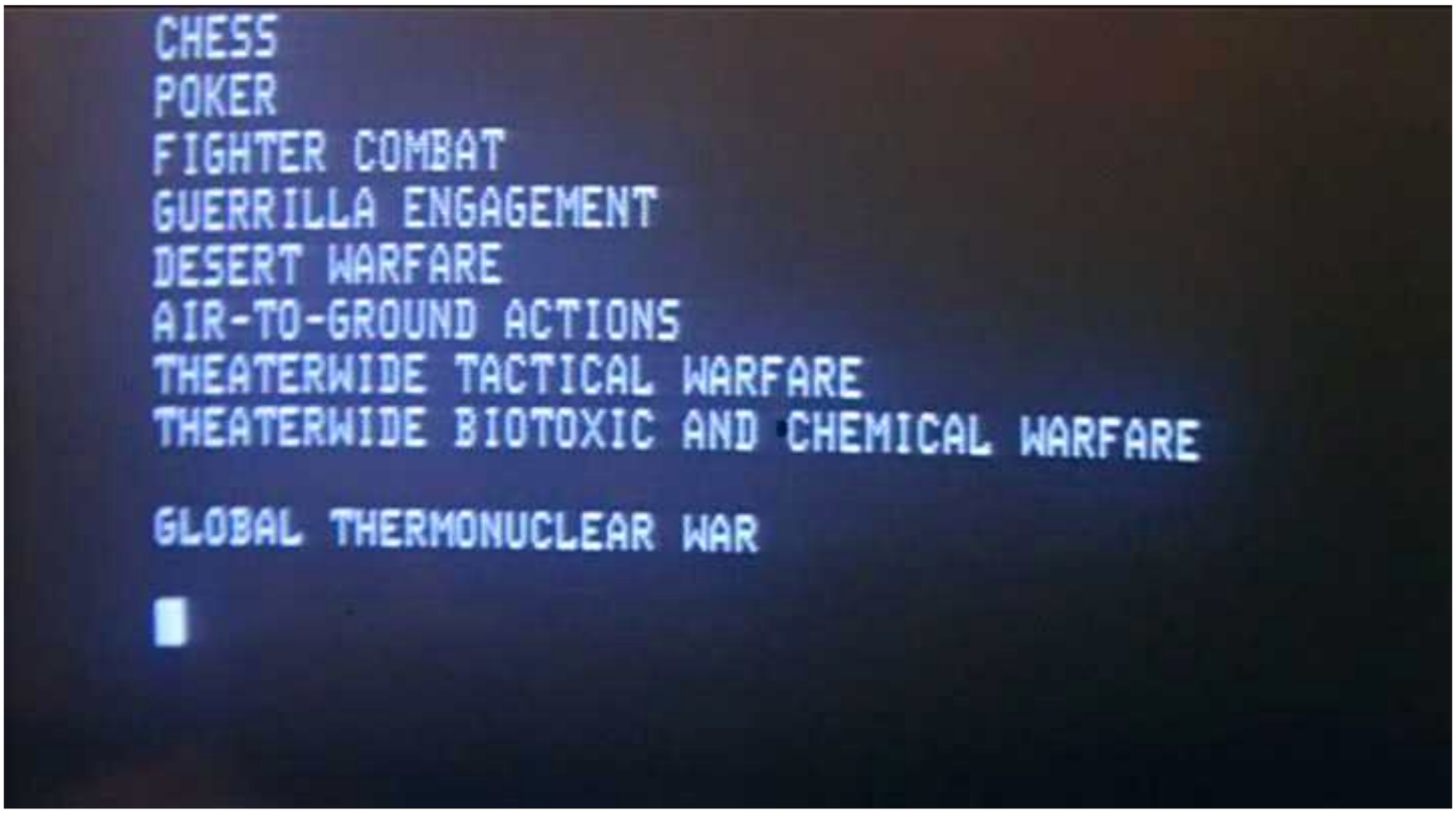
- The complex nature of modern technical systems more or less require AI to control many systems

- Defense systems, power production, social security, credit rating, the Internet…

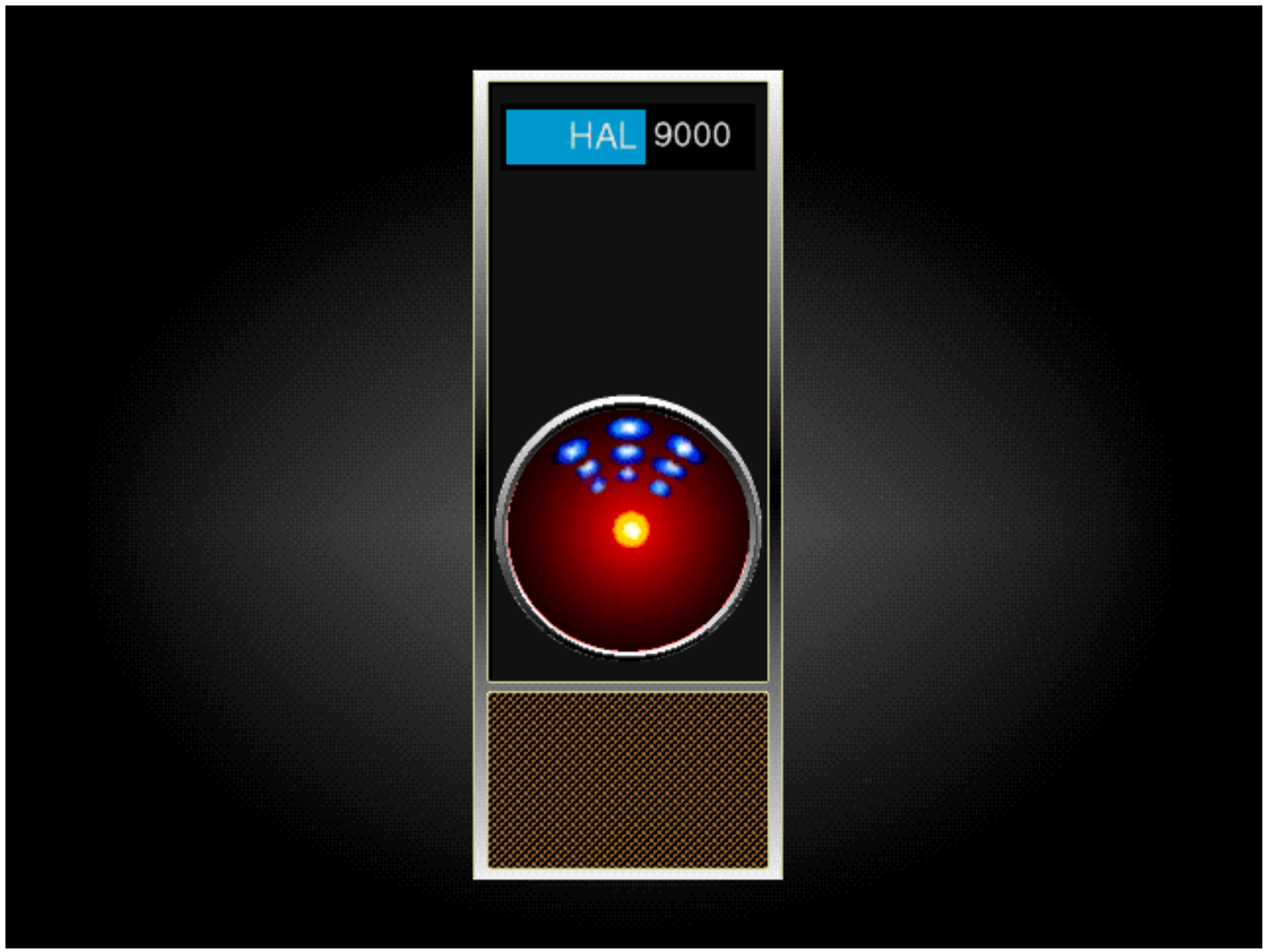
- There are so many systems that we don’t completely understand anymore and therefore don’t really control

- And this is likely to get worse!

War Games (1983)



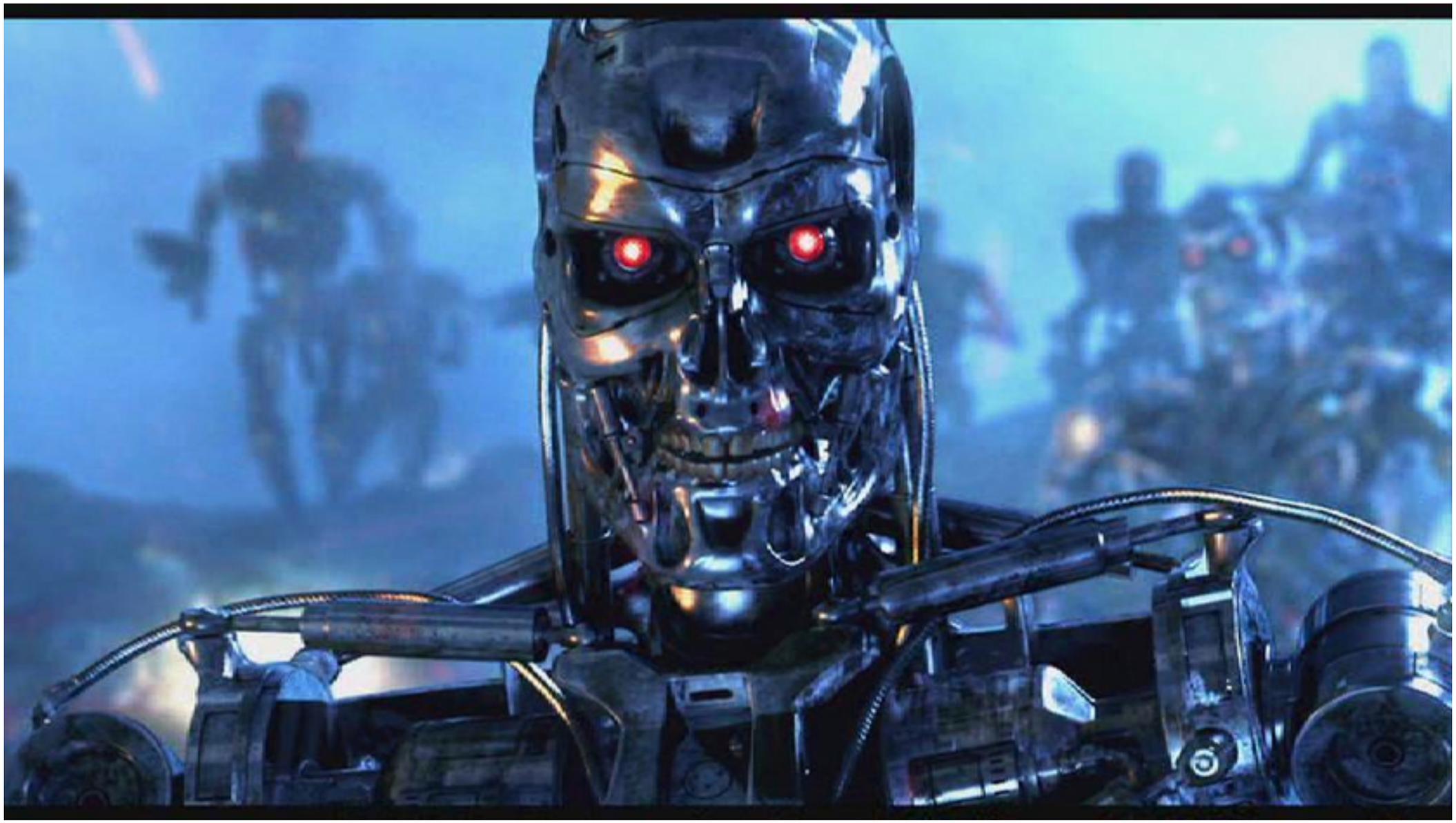
2001: A Space Odyssey (1968)



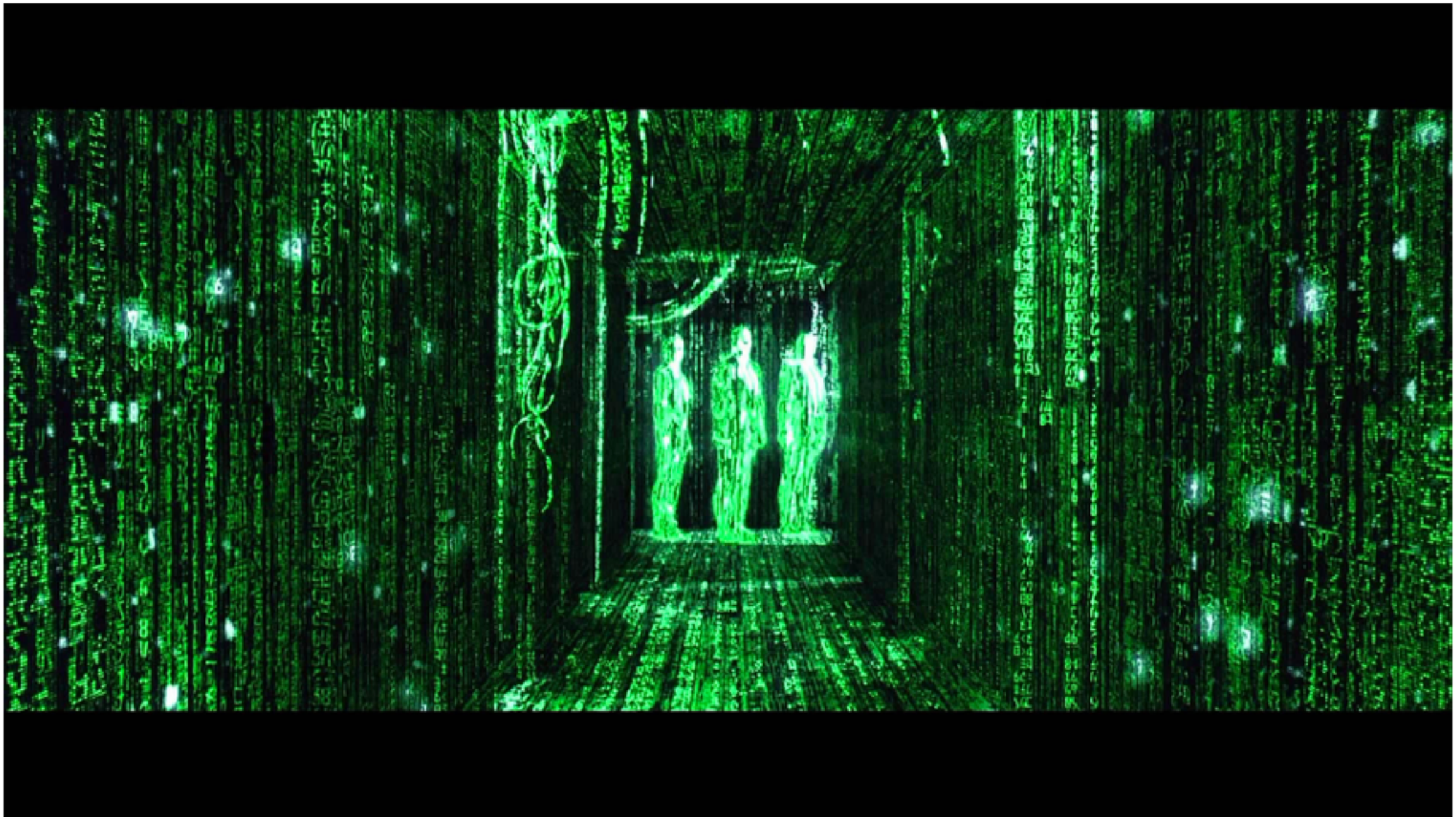
Superintelligence

- If we create AI that is as intelligent as us, it might soon learn to modify itself to become smarter
- We might rapidly lose ability to control or even understand it (a type of *singularity*)
- We don’t know its value system, and it might not want what’s best for us
- Even if it wants what’s best for us, the outcome might be catastrophic (e.g. trapping us in a pleasant simulation)

Terminator… which one?



The Matrix (1999)



Superintelligence?

- Can we put in safeguards? Affect the value system of the AI? Keep it in a sandbox?
    - But it can modify itself
    - And fool us
- Will it really be able to modify itself?
- Is there a difference between the “AI” and the whole socio-technical complex that we are part of?