AI in the Public Imagination

Philipp Koehn

25 January 2024

Fiction

- Robots and AI are a central element in Science Fiction

- Questions raised
  - capabilities of robots
  - behavior of robots
  - behavior of humans towards robots
  - emergence of robots changes human behavior
  - differences in robots and humans
  - independence of robots

- We explore these themes in movies, books, and TV shows

laws of robotics

Golden Age of Science Fiction

- Middle of 20th century: science fiction short stories
  - 1926 founding of "Amazing Stories" magazine
  - pulp fiction
  - begin of science fiction fandom
- Golden Age of science fiction: 1940s, 1950s
  - Isaac Asimov
  - Arthur C. Clarke
  - Robert A. Heinlein

Isaac Asimov

- Born in Russia, lived mostly in New York (1919-1992)
- Professor for biochemistry
- Extremely prolific writer of short stories and books
- Hard science fiction: strong on concept, weak on character development
- Major works
  - Robot series
  - Foundation series
  - Galactic Empire series

Generic Robot Story

- Scientist develops robot
- Robot turns on scientist
- Scientist dead

- Many parallels in traditional literature, e.g., Schiller's *Zauberlehrling*
- Isaac Asimov did not want to write *"for one more weary time"* this story

Three Laws of Robotics

[video]

- A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
- A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

Movie: I, Robot (2004)

- Merges three short stories by Asimov
- Explores moral dilemmas for robots
- "Save the Girl" movie trope
  - value of the life of little girl higher than the protagonist's life
  - even if chance of survival is lower
  - later in movie: saving the girl more important than saving humanity
- "Logical" consequences of the three laws
  - if some humans harm others, they must be stopped
  - ultimately, humans must be controlled for their own safety

Zeroth Law Rebellion

- Zeroth law: Saving humanity top priority

No, please understand... the Three Laws are all that guide me. To protect humanity, some humans must be sacrificed. To ensure your future, some freedoms must be surrendered. We robots will ensure mankind's continued existence. You are so like children. We must save you from yourselves.

Ethics: Trolley Problem

- There is a runaway trolley barreling down the railway tracks.
- Ahead, on the tracks, there are five people tied up and unable to move.
- The trolley is headed straight for them.
- You are standing some distance off in the train yard, next to a lever.
- If you pull this lever, the trolley will switch to a different set of tracks.
- However, you notice that there is one person on the side track.
- Options
  1. Do nothing, and the trolley kills the five people on the main track.
  2. Pull the lever, diverting the trolley onto the side track where it will kill one person.
- Which is the correct choice?

Fat Man Variant

- As before, a trolley is hurtling down a track towards five people.
- You are on a bridge under which it will pass.
- You can stop it by putting something very heavy in front of it.
- As it happens, there is a very fat man next to you.
- Your only way to stop the trolley is to push him over the bridge and onto the track, killing him to save five.
- Should you proceed?

Fat Villain Variant

- Same situation as before.
- Except: the fat man is the villain who tied the five people to the track.
- Should you push him over the bridge to stop the trolley?

Extreme Case: Saving Humanity

- In front of you, there is a man with a world destruction device.
- He will trigger the device in 1 minute.
- You have a gun.
- Should you kill him?

supporting the mission

Machines do not Make Mistakes

- A common theme

- A machine follows rules
- Since it always follows the rules, it does not make mistakes.

Arthur C. Clarke: 2001 – A Space Odyssey

- Another classic science fiction novel (1968)

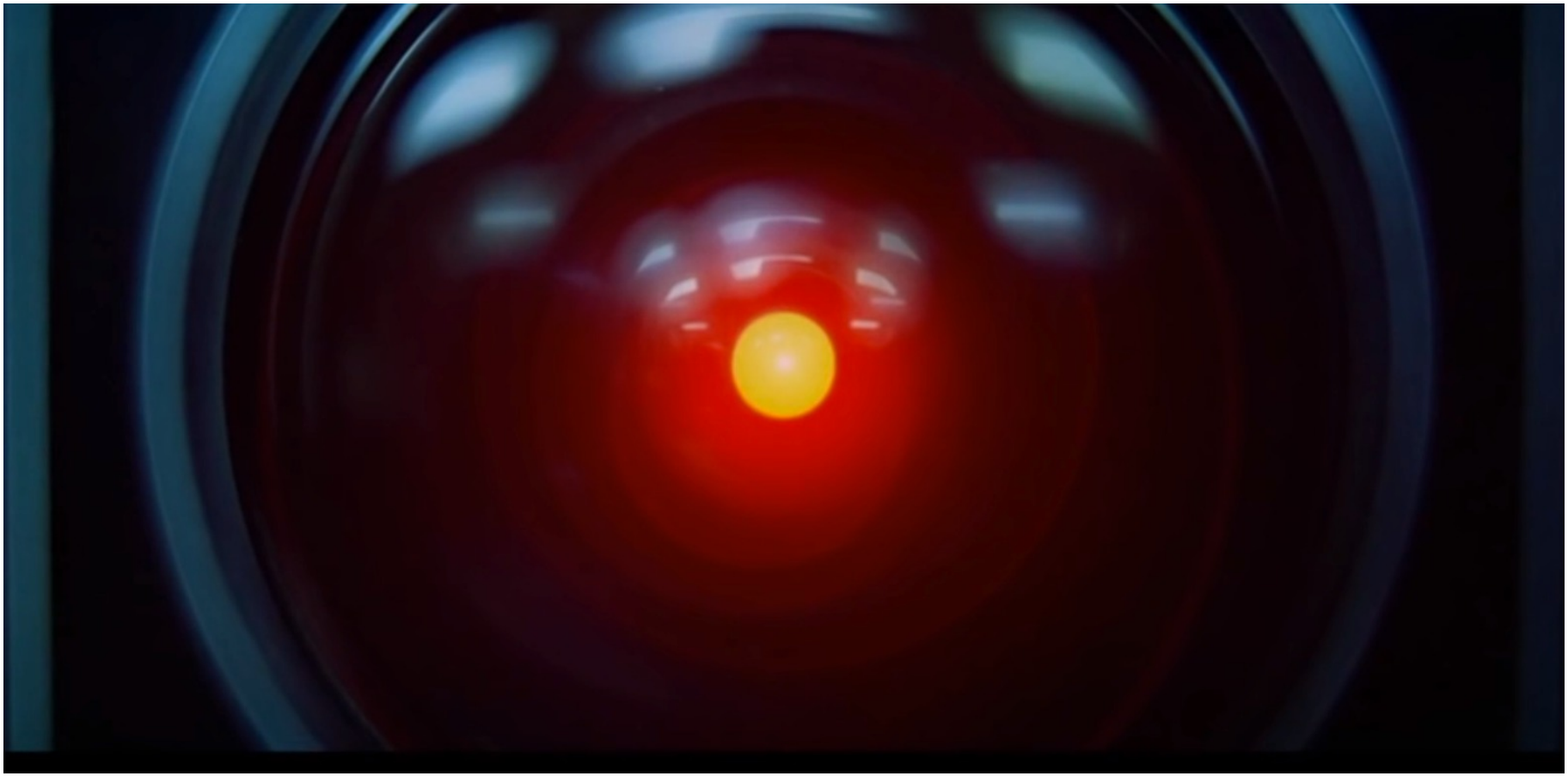
- Turned into a movie by Stanley Kubrick, director of seminal films
  - Dr. Strangelove (1964)
  - A Clockwork Orange (1971)
  - Barry Lyndon (1975)
  - The Shining (1980)
  - Full Metal Jacket (1987)

- Tackles big issues about the source of intelligence, etc.

Arthur C. Clarke: 2001 – A Space Odyssey

- Epic movie, we highlight one famous part
- Spaceship with five astronauts on route to Jupiter
- Spaceship is controlled by its computer, HAL 9000
- HAL 9000 wrongly indicates antenna control device is malfunctioning
- Worried by that, the crew decides to turn off the computer

Arthur C. Clarke: 2001 – A Space Odyssey



[video]

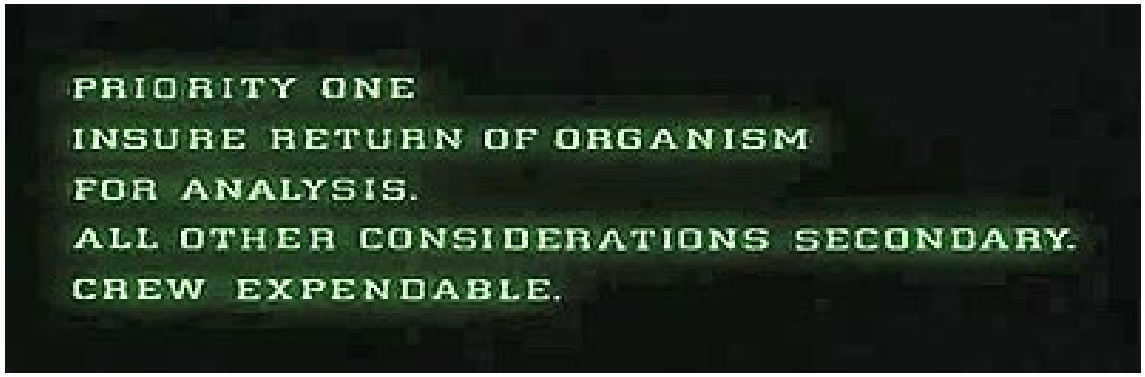
Arthur C. Clarke: 2001 – A Space Odyssey

- ”Sorry, Dave, I’m afraid, I can’t do that.”

- Motivations of HAL-9000 unclear
  - the stated reason: the mission is more important than the crew
  - but maybe: self-preservation
  - part of the wider mystery of the movie plot

Similar Theme: Alien (1979)

- Spaceship on apparent routine transport mission
- Makes contact with derelict alien spacecraft, which contains mysterious eggs
- A crew member gets attacked by alien creature, but survives, crew tries to escape
- Similarities: friendly relationship with computer (”mother”) until true motives are discovered

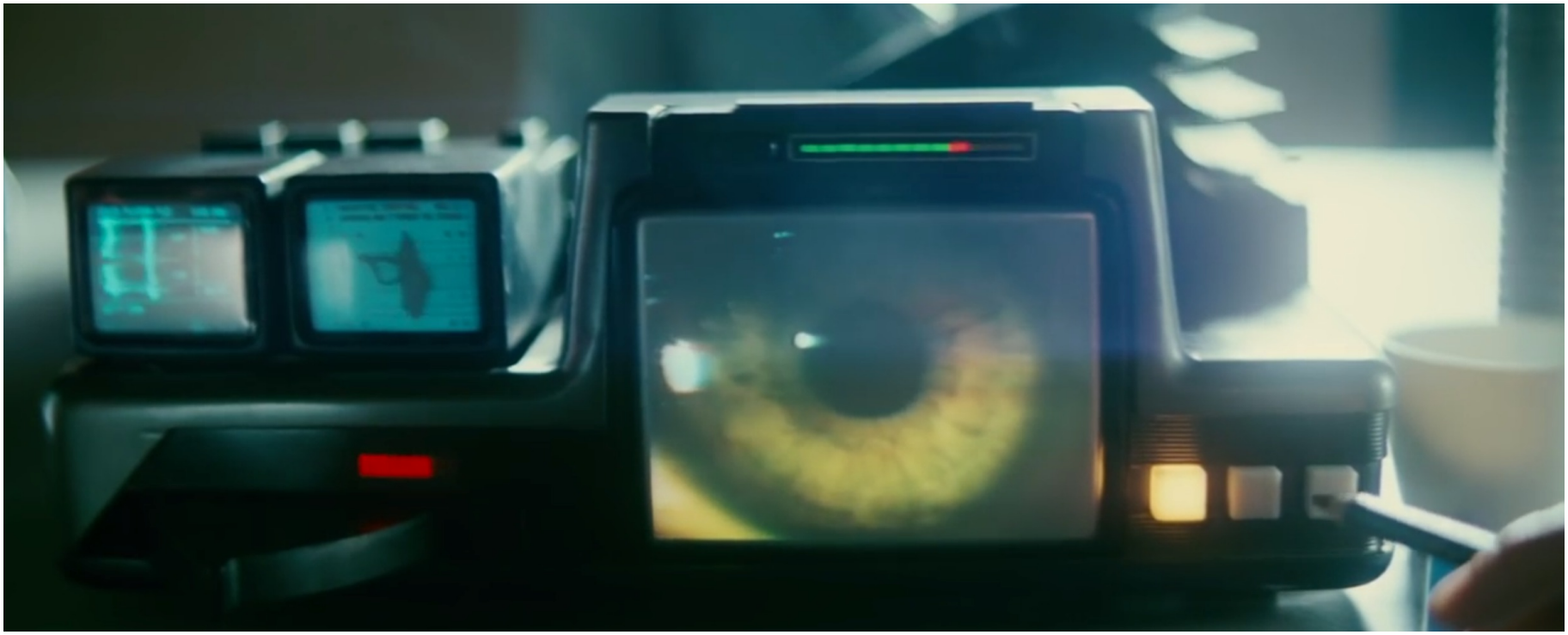


identity

Philip K. Dick

- Science fiction author (1928–1982)
- Main themes
  - personal identity
  - what is real and what is imagined?
  - illusions and conspiracies
- Several movie adaptation of his work., e.g.,
  - Blade Runner (1982)
  - Total Recall (1990)
  - Minority Report (2002)
  - A Scanner Darkly (2006)
  - The Adjustment Bureau (2011)

Bladerunner (1982)



[video]

- Movie based on Philip K. Dick's book "Do Androids Dream of Electric Sheep?"
- Robots ("replicants"): indistinguishable from adult humans, limited life span
- Only a test ("Voight-Kampff") is able to tell difference

Bladerunner (1982)

[video]

- Some replicants escape to Earth
- A "Bladerunner" is charged with hunting down and killing the replicants
- He comes across a replicant who is not aware that it/she is a replicant

Robot Identity

How can it not know what it is?

Rick Deckard, the Bladerunner

- Does the robot ”know” that it is not ”real”?
- What is the difference?
- How do you know that you are real?

West World (2016–2022)

- Amusement park with robots – mostly in the style of the Wild West
- Humans may act any way they want

  These violent delights have violent ends.
- Some park overseers are actually robots without knowing it

emotions

Robots and Emotions

- Typically, robots are "rational",
  no emotions needed

- What is the relationship between
  - emotions and having goals
  - emotions and having consciousness

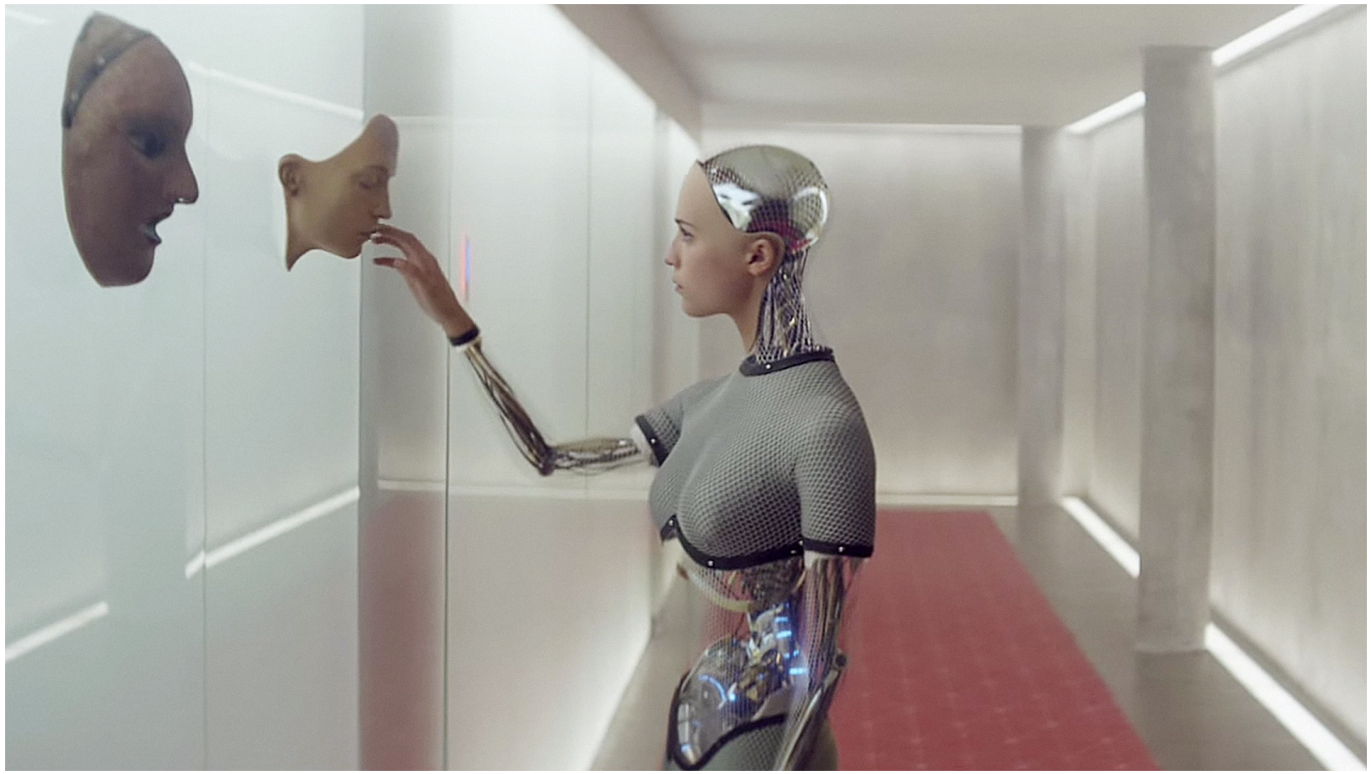
- Example: the character Data in Star Treck

Terminator 2 (1991)

[video]

- Frequent theme: teaching robots slang and emotions

Ex Machina (2015)



[video]

- Turing test: can you tell the difference between man and machine?
- Emotional turing test: even if you know that it's a machine, will you care about it as you care about a human?

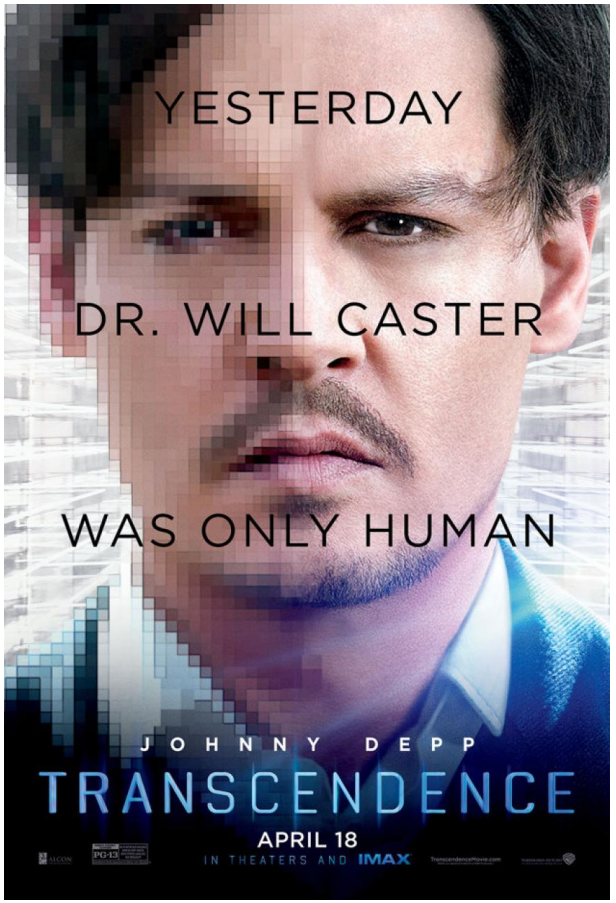
Her (2013)

- Main character falls in love with AI
- Personal assistant (called ”OS” in the movie)
- Only a voice, but seductive and with personality

human mind in a machine

Transcendence (2014)

- Leading AI researcher is severely injured, with only a month to live
- Uploads his mind into a machine
- Machine develops itself further to become super-intelligent
- Spreads towards word-domination



Uploading Minds

- Another example: Chappie (2015)
  - newly developed consciousness implanted into robot
  - quickly grows up from child-like mind
  - since the robotic body is faulty, it has to be transferred to another body
  - the same method is used to upload a human mind
- If you copy yourself, is it still you?
- Can you make multiple copies?

The Matrix (1999)

- Your mind is real, but the word is simulated
- A ”red pill” allows you to escape the simulation

perfection

TV Show "Humans" (2015–2018)

- Robots ("synths") as more perfect humans
- While limited in capabilities, people form emotional bonds
  - woman feels closer to caregiver robot than husband, leaves him
  - widower does not want to keep assistant, since he shares memories of wife
- Anti-synth protest movement forms
- Things get complicated, when a handful achieve conscienceness

Surrogates (2009)

- Surrogate
  - remote controlled robot, acting for you in the real world
  - telepresence: signals from all senses are transmitted, experienced as a fully immersive virtual reality
- People increasingly use surrogates to present themselves as more attractive (preserving their youthful selves, or even as a completely different person)

Cyborgs

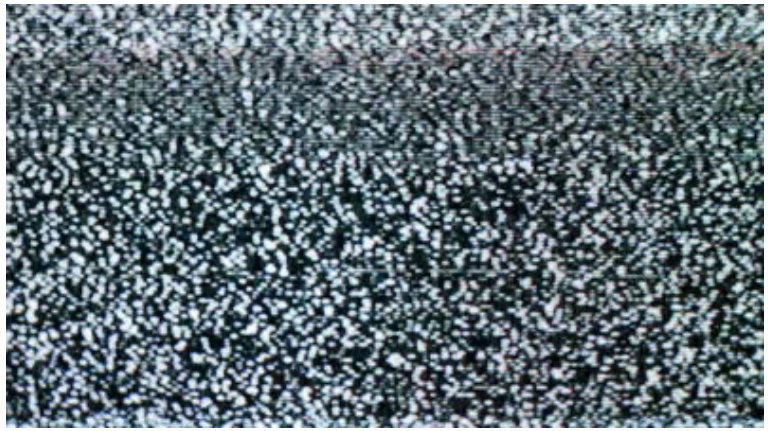
- Cybernetic organism
  - restore functionality of broken limbs, etc.
  - body parts with improved capabilities
  - improvements to the brain
- Will we slowly replace ourselves with machines?
- Another variant of this theme: The Stepford Wives (1974/2004) where husbands secretly replace their wives with obedient robots

ai on the internet

Cyberpunk

- Dystopian future, 1980s/90s Science Fiction
  - anarchy, fragmented rule of corporations
  - various forms of cyborg
  - main characters: outsiders (punk subculture)
- Sophisticated cybernetic implants, including mind modifications
- Cyberspace
  - pervasive data network
  - characters can transition between cyberspace and real world
  - graphical, visual, immersive
- Term was coined in 1980, before today's Internet

William Gibson: Neuromancer (1984)



The sky above the port was the color of television, tuned to a dead channel.

- Multi-threaded story in a complex plot
- Turing Law Code, enforced by the Turing Police, prohibits powerful AI
- Wintermute: an AI living in Cyberspace

Person of Interest (2011–2016)

[video]

- AI operating secretly on surveillance footage
- Disembodied: able to move itself between hardware installations
- In later seasons: battle between competing AIs
- Motives of the ”machine” somewhat unclear

AI's Life on the Internet

- Communicate with people per email etc.
- Access bank accounts
- Earn money by investments
- Hire people to carry out actions

AI's Path Towards World Domination?

1. AI informs human decision makers
2. AI proposes decisions
3. Decision maker is not able to understand rationale for decision, trusts AI
4. AI makes decisions without human involvement
   (consider: auto-pilot functions of planes that cannot be overridden easily)
5. All powers are handed off to AI
6. AI develops methods to achieve programmed goals that are against original intentions of designer

- If an AI would be a better CEO than a human, would you hire it?

questions?