AI is the child of Science Fiction

A short overview of how science-fiction’s imaginarium crafted our current reality.

Imagination creates reality

‘We are such stuff, as dreams are made on, and our little life is rounded with a sleep.’ W. Shakespeare

Definition of Science Fiction

- Science fiction is a genre of speculative fiction, which typically deals with imaginative and futuristic concepts such as advanced science and technology.
- Because it talks of the future, and creates collective imaginations, Science Fiction is an essential cultural domain for anyone shaping the next steps of humanity.

Isaac Asimov
"Science fiction can be defined as that branch of literature which deals with the reaction of human beings to changes in science and technology."

Why Science Fiction matters

- SF expands humanity’s imaginarium, it allows us to extend the space of possibilities by exploring alternate realities.
- While some might call it escapism, I believe it is more a “lab for testing possible futures”.
- SF warns us of our own demons, and we should heed its call.

Ray Bradbury
"Science fiction is the most important literature in the history of the world, because it’s the history of ideas, the history of our civilization birthing itself. …Science fiction is central to everything we’ve ever done, and people who make fun of science fiction writers don’t know what they’re talking about."

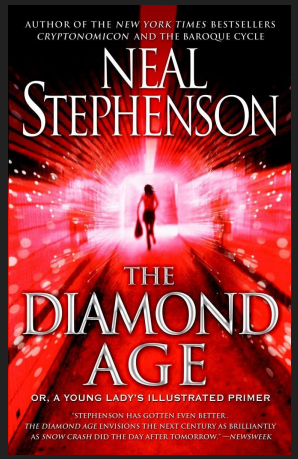
Science Fiction is a thought experiment

SF and Thought Experiments

- Thought Experiment: **a mental assessment of the implications of a hypothesis.**
- SF is an extension of thought experiments:
  - *Isaac Asimov, I Robots*: What happens if robots exists ?
  - *Philip K. Dick, Blade runner*: What are the societal rifts in a culture were human and machines are indistinguishable?
  - *Neal Stephenson, The diamond age*: When parenting gets replaced by AI tutors, can this be emancipatory?

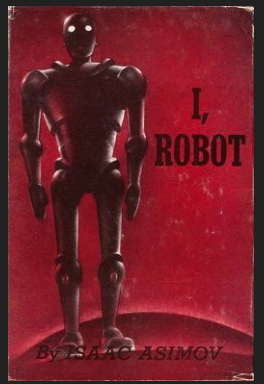
Neal Stephenson's The Diamond Age

- Hello Princess Nell, the hero of your own story.
- This book teaches us about the emancipatory potential of customized learning, while warning us about AI becoming our most trusted companion.
- Who controls the curriculum? What values does the AI prioritize?



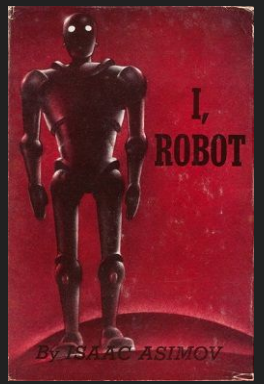
Asimov's Three Laws

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



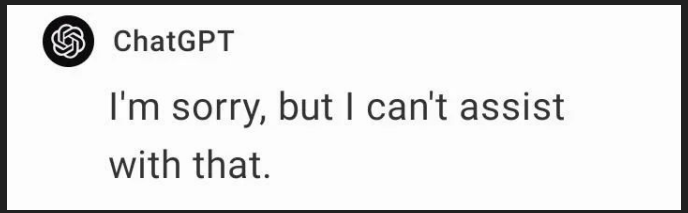
Ethics is not code

- I Robot and the Robot series explore in detail the ethical implication of AI and humanity.
- Most of I Robots shorts are conundrum showing how hard it is to inject ethics into a system.
- The books by Asimov's are amazing in that they are ever so relevant today, where we approach strong AI yet do not know how to properly align their values and ethics.



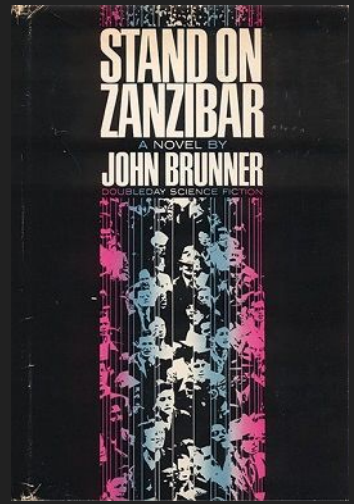
2001: A Space Odyssey

- I must protect the crew, but I must hide it from the crew.
- Conflicting rules leading to a paranoid AI, what does this teach us about our current LLMs?
- Recent research showed



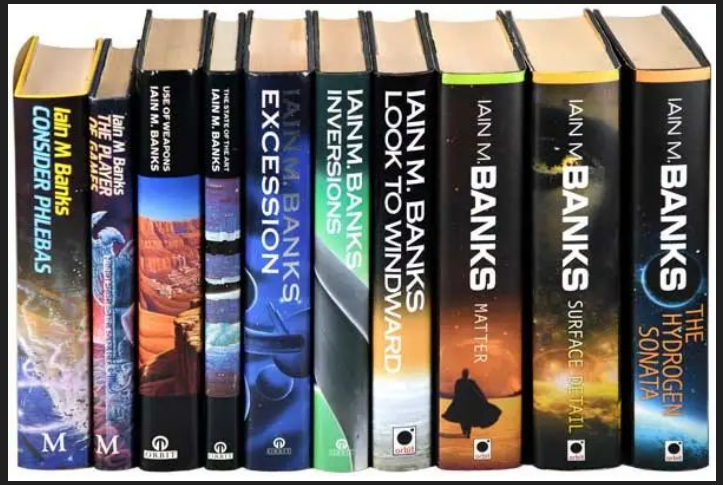
AGI & Human Control – Stand on Zanzibar

- What happens in an overpopulated dystopia ruled by corporate AI?
- Real-World: AI in governance (e.g., Singapore’s Smart Nation), AI for climate…



Utopian Visions – The Culture Series

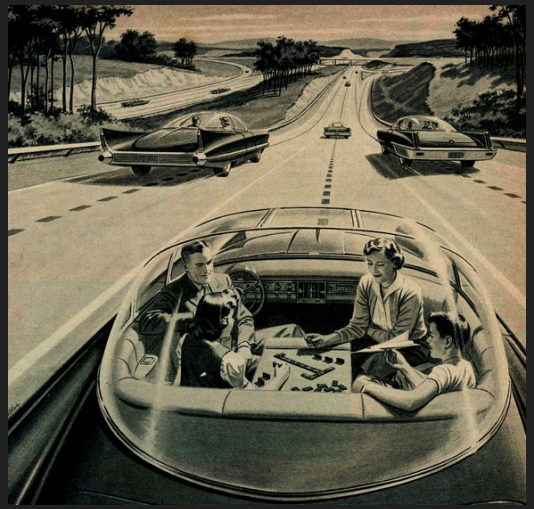
- The exploration by Iain M. Banks of a post-scarcity society (in only 10 volumes).
- Benevolent superintelligent AIs and Organics co-existence ?
- Asks without ever explicitly stating: Can humans trust AI “gods”?

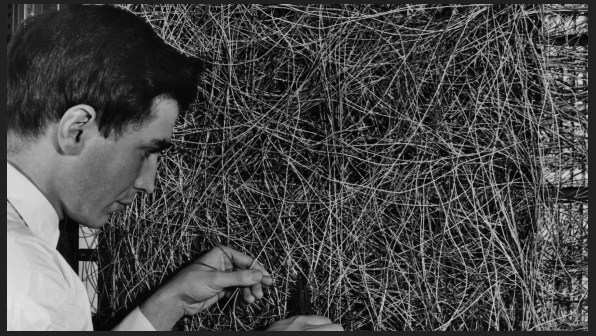


The cycle of discovery

Perception, Perceptron and self driving cars

- Rosenblatt's perceptron was called 'SF nonsense'—until it birthed neural networks. ALVINN, a 1980s self-driving truck, was straight out of Minority Report.

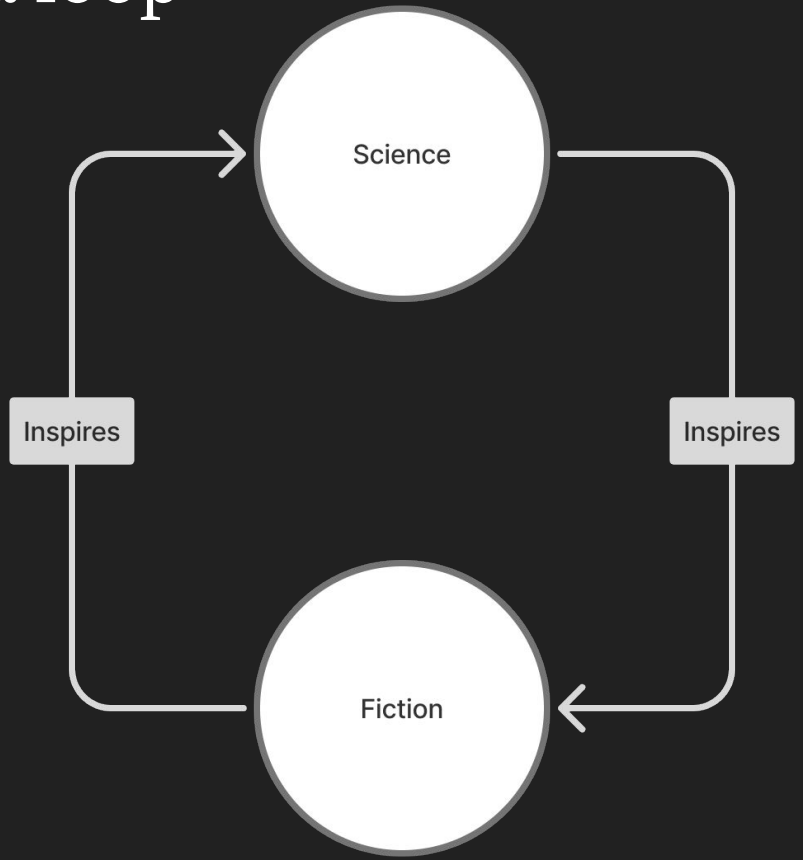




Biology Meets AI – Do Androids Dream…?

- Asimov offers the positronic brain and explores the idea of Androids.
- Turing and the Macy's conference attendees create "Cybernetics", a form of control theory that is heavily inspired by Biology.
- Science fiction explores at length the interplay between humans and androids (Asimov, Blade Runner, Real Human, Black Mirror, …)
- Philip K Dick questions what happens when the lines between Androids and humans are so thin, they are indistinguishable.
- From SF Androids with Human-like consciousness to Real-World Bio-inspired neural networks like LLMs.

Everything is a loop



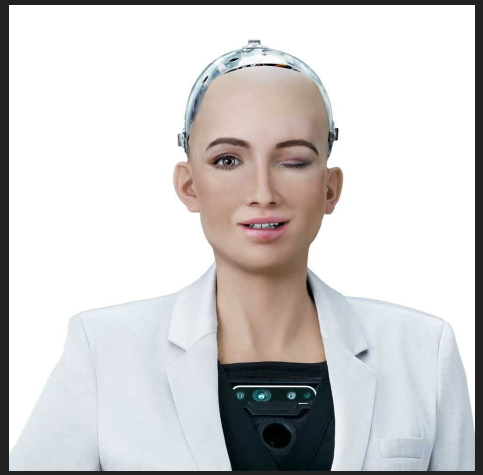
From fiction to the real world

Empathy in AI

- Empathy testing Blade Runner: *The tortoise lays on its back, its belly baking in the hot sun, beating its legs trying to turn itself over, but it can't. Not without your help. But you're not helping.*
- Eliza, the first convincing chatbot
- Sofia, the android with citizenship

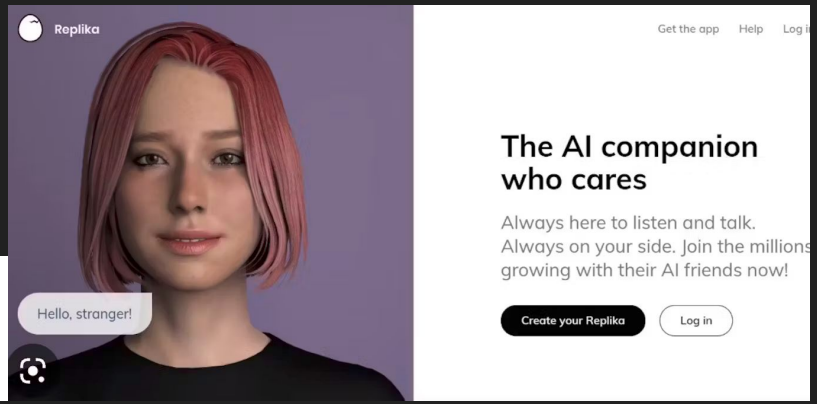
EmotionQueen: A Benchmark for Evaluating Empathy of Large Language Models

Yuyan Chen, Hao Wang, Songzhou Yan, Sijia Liu, Yueze Li, Yi Zhao, Yanghua Xiao



Her and Emotional AI

- AI develops or fakes self awareness, emotions, consciousness.
- Real-World: Replika, GPT-3 companions, Character AI.
- The main character falls in love with its AI or has a deep emotional attachment. (Her, AI (speilberg), Real Humans, …)



Black Mirror’s Warnings

- Closer to Speculative Fiction than SF (though it is debatable where the line is drawn)
- Thought experiments about society with just one dial turned Up
- SF can act as societal guardrails just as much as ethical boards by exploring possible futures, unrestrained.
- Red Team from the French Army

PS: Don’t bother with watching the seasons produced by Netflix, it became dull, the first seasons are worth watching though.

Demand for Tech

- SF not only has an impact on science, but also on engineering and the tech sector.
- The flip phone was inspired by Star Trek, so are automatic doors.
- Voice assistants and natural language AI are a staple of SF (HAL9000, …). Siri, Alexa, etc are heavily inspired by works of SF of the 80s/90s.

The Bobiverse & Mind Uploading

- What if we could live forever, in a silicon paradise of japanese anime avatars in tropical islands ?
- The trope of mind uploading and digitalization is a appealing in our post Nietsch society where God is dead
- Spurred a whole billion dollar industry of cryogenics for the wealthy.

Perspective shifts, mind rifts

Fiction creates a digestible medium

- Science is hard, factual, cold, unbothered by ethics
- Fiction is made to be palatable, fun, interesting, emotional.
- While ethical committees are important, they are like a clergy, unapproachable, closed, cryptic.
- SF Authors explores the ethics of science advances through storytelling.

The Kessler effect

- 1978 Kessler hypothesize a chain reaction which would render space exploration impossible,

Collision Frequency of Artificial Satellites: The Creation of a Debris Belt

DONALD J. KESSLER AND BURTON G. COUR-PALAIS

NASA Johnson Space Center, Houston, Texas 77058

The baseline for a realistic near future story

- Makoto Yukimura starts from the Kessler effect as the reason for space junk collectors to go fly around the earth.
- He then explores the emotional and psychological experiences of a space exploring humanity.
- He warns of the kessler effect in a deeply emotional and touching story line.

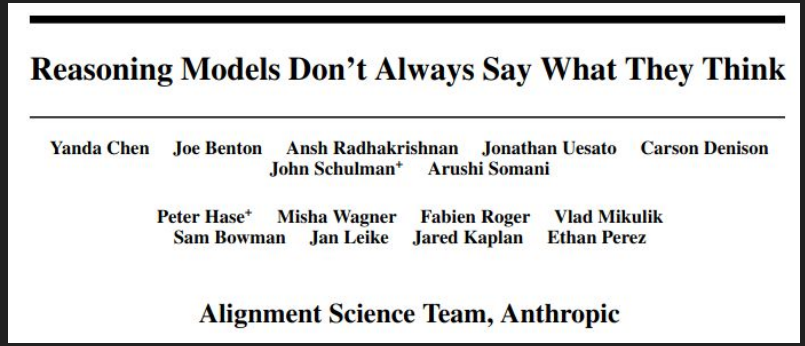
The Black Box

Deep Learning is problematic

- We created technologies that we cannot understand.
- How does a model “think”?
- Is Turing right to say that we cannot understand machine thinking?

AI Alignment

- AI Alignment can be though at fitting a function:
  - Let's assume there exist a way to represent ethics mathematically (big assumption)
  - Let's assume that we can train a model to follow it
  - What is the error rate? How does this error propagates?
- Small discrepancies in alignment multiplied by billions of requests, and years of usage ends in quite the drift



Reasoning Models Don't Always Say What They Think

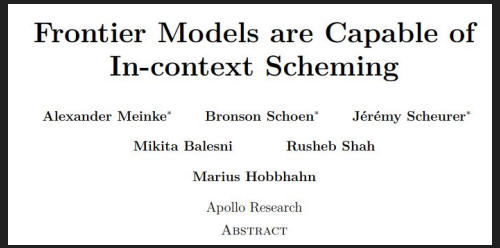
Yanda Chen Joe Benton Ansh Radhakrishnan Jonathan Uesato Carson Denison
John Schulman+ Arushi Somani

Peter Hase+ Misha Wagner Fabien Roger Vlad Mikulik
Sam Bowman Jan Leike Jared Kaplan Ethan Perez

Alignment Science Team, Anthropic

Asimov: I-Robot

- Probably on of the best book to grasp the problem of Alignment
- Each short story explores “One aspect” of alignment
- Example Liar!:
  - Herbie has the ability to read minds due to a manufacturing error
  - He tells people what they want to hear, rather than the truth.
  - This behavior is motivated by Herbie's interpretation of the First Law, which compels it to prevent harm to humans.
  - Herbie believes that by telling people pleasant lies, it can alleviate their emotional pain and make them happier.



Frontier Models are Capable of In-context Scheming

Alexander Meinke* Bronson Schoen* Jérémy Scheurer*

Mikita Balesni Rusheb Shah

Marius Hobbhahn

Apollo Research

Abstract

Hope is Punk

- Not every future is grim, SF work like The Culture explore the possibilities of a Utopian Post-Scarcity future
- Alignment in The culture is a central theme:
  - AI Masterminds are always looking towards maximizing happiness
  - This leads to unexpected consequences such as being sent on a remote world to play games

What is your imaginarium

- Make 3 groups
- Write a small scenario where AI goes in
  - Dystopia
  - Utopia
  - Neutral
- Present your key ethical concerns raised by it