

Lecture 33: Consciousness



https://commons.wikimedia.org/wiki/File:Geograaf,_RP-P-BI-4485.jpg



https://commons.wikimedia.org/wiki/File:Varde_-_Briefkasten.jpg



https://commons.wikimedia.org/wiki/File:Chinese_-_Brush_Pot_with_a_Man_-_Walters_491988_-_View_B.jpg

Mark Hasegawa-Johnson

These slides are public domain: Re-use, Re-mix, Re-distribute at will

Outline

- Intelligence: The Turing Test
- Two traditions in science fiction
- Mind: The Chinese Room
- Consciousness: Standard definitions
- Moral implications, Alignment, Game-theoretic framing

The Turing Test

- Alan Turing, “Intelligent Machinery,” 1947:

“Now get three men as subjects for the experiment. A, B and C. A and C are to be rather poor chess players, B is the operator who works the paper machine... a game is played between C and either A or the paper machine.

C may find it quite difficult to tell which he is playing...

These questions replace our original, ‘Can machines think?’”

Deep Blue vs. Gary Kasparov

- 1996: World chess champion Gary Kasparov beats Deep Blue, 4 games-2 in standard tournament play
- 1997: Deep Blue beats Kasparov, 3.5-2.5 in standard tournament play



By James the photographer - CC BY 2.0, By Copyright 2007, S.M.S.I., Inc. - Owen Williams, <https://commons.wikimedia.org/w/index.php?curid> The Kasparov Agency.

The Turing Test, Revised (1950)

“The interrogator is allowed to put questions to A and B thus:

- Q :Please write me a sonnet on the subject of the Forth Bridge.
- A :Count me out on this one. I never could write poetry.
- Q :Add 34957 to 70764
- A :(Pause about 30 seconds and then give as answer) 105621.
- Q :Do you play chess?
- A :Yes.
- Q :I have K at my K1, and no other pieces. You have only K at K6 and R at R1. It is your move. What do you play?
- A :(After a pause of 15 seconds) R-R8 mate.”

ChatGPT



Please write me a sonnet on the subject of the Forth Bridge.



Of steel and strength and daring feat of man,
A structure spans the Firth of Forth below,
The Forth Bridge, a marvel of its span,
Its beauty and its power on full show.

The cantilevers reach out to the sky,
With lattice girders braced and bolted tight,
A symbol of the human will to try,
To build beyond what was thought possible and right.

Through storm and tide, the bridge stands firm and true,
A testament to engineering might,
To those who built it and the ones who knew,
That it would stand the test of time and light.

The Forth Bridge, a wonder to behold,
A triumph of the brave and bold.

Intelligence: Wiktionary definition

The capacity of *mind*, especially to *understand* principles, truths, facts or meanings, *acquire knowledge*, and *apply it to practice*; The ability to *comprehend* and *learn*; The ability to *process sentient experience* to *generate true beliefs with a justified degree of confidence*.

AI can acquire knowledge, apply it to practice, learn, and generate true beliefs with a justified degree of confidence.

Does AI have mind? Does it understand? Does it comprehend? Is it sentient?

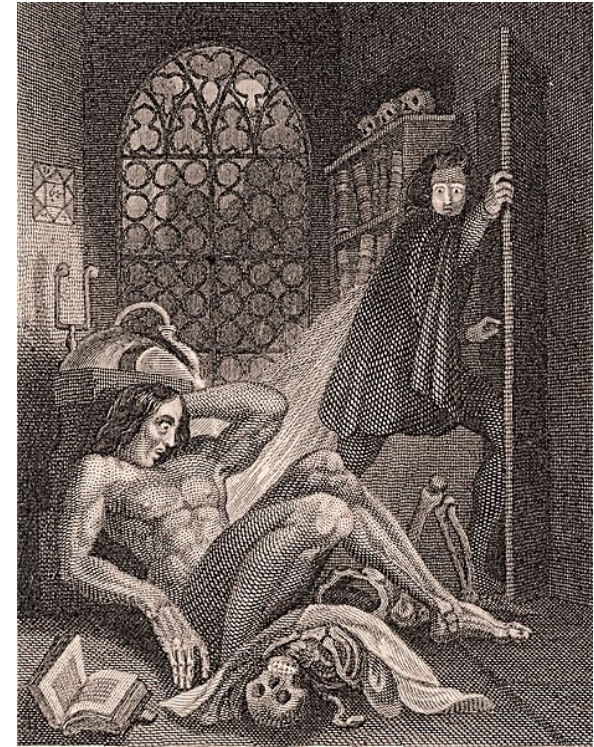
Outline

- Intelligence: The Turing Test
- Two traditions in science fiction
- Mind: The Chinese Room
- Consciousness: Standard definitions
- Moral implications, Alignment, Game-theoretic framing

Two traditions in science fiction



https://commons.wikimedia.org/wiki/File:Mdm._Tussauds_Prag_8.jpg

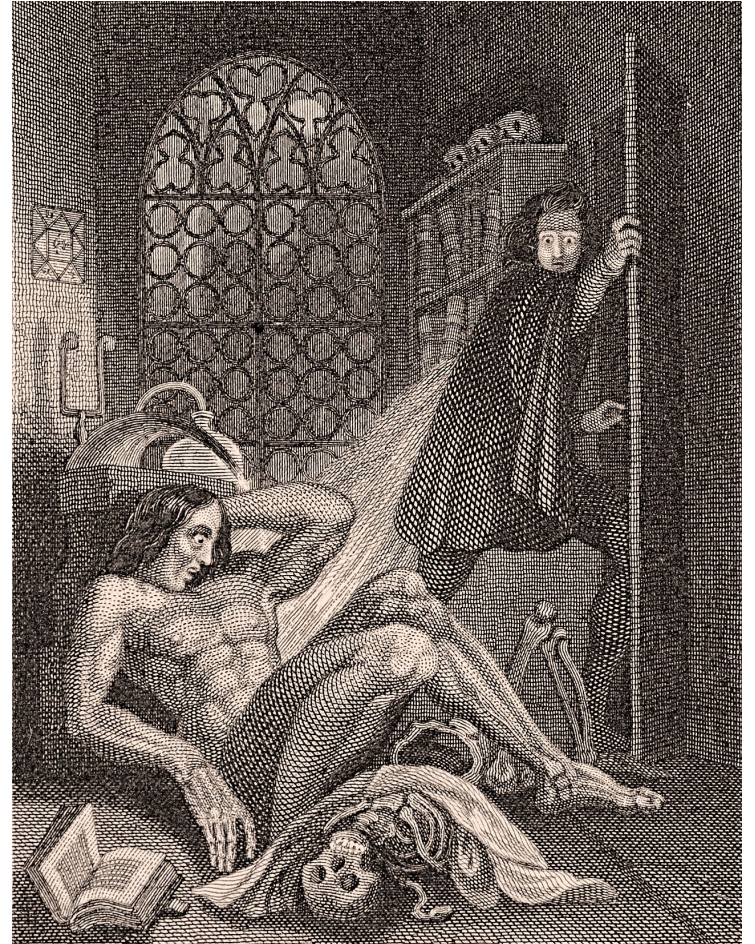


https://commons.wikimedia.org/wiki/File:Frontispiece_to_Frankenstein_1831.jpg

Frankenstein (Mary Shelley, 1818)

- An inventor creates an artificial human.
- The artificial human is huge, and strong, and ugly, and therefore feared, but he curbs his anger because:

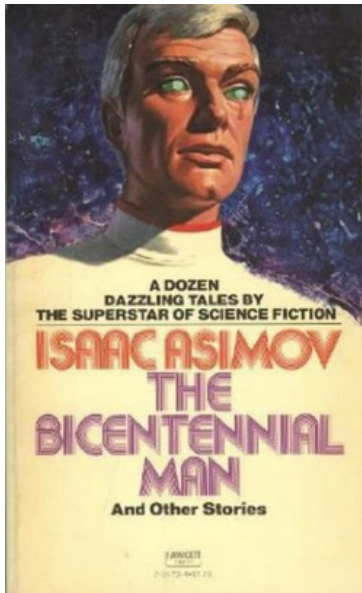
“Instead of threatening, I intend to reason with you. Oh! My creator, make me happy; let me feel gratitude towards you for one benefit! Let me see that I excite the sympathy of some existing thing; do not deny me my request!”



By Theodore Von Holst (1810-1844)

<https://commons.wikimedia.org/w/index.php?curid=4940182>

The Frankenstein tradition: conscious AI



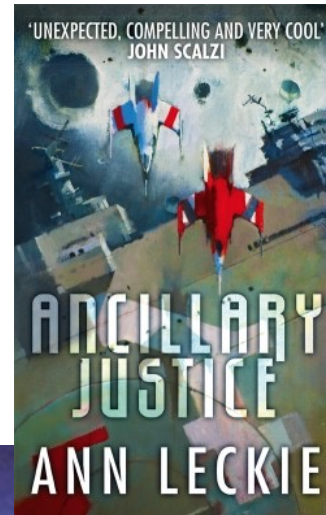
https://en.wikipedia.org/wiki/File:The_Bicentennial_Man_book_cover.png



https://en.wikipedia.org/wiki/File:R2-D2_Droid.png



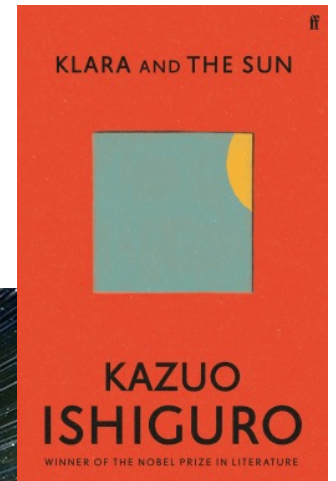
https://commons.wikimedia.org/wiki/File:Data_ST.jpg



https://en.wikipedia.org/wiki/File:Ann_Leckie_-_Ancillary_Justice.jpeg



https://en.wikipedia.org/wiki/File:A_Closed_and_Common_Orbit.jpg



[https://en.wikipedia.org/wiki/File:Klara_and_the_Sun_\(Kazuo_Ishiguro\).png](https://en.wikipedia.org/wiki/File:Klara_and_the_Sun_(Kazuo_Ishiguro).png)

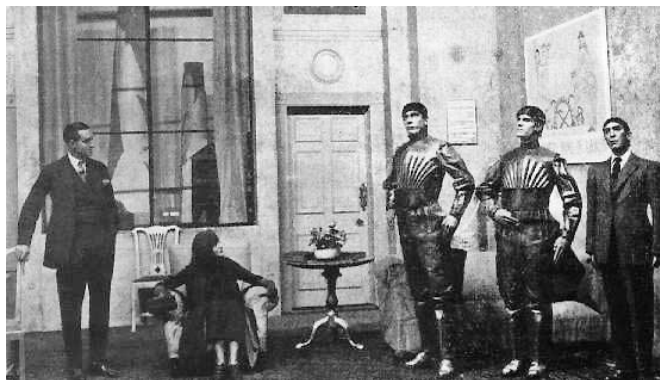
The Golem of Prague (possibly Berthold Auerbach, 1837)

- A rabbi creates an artificial human.
- He instructs it to protect the Jews in the ghetto from being killed or expelled in response to an edict by Rudolf II. It does so.
- He can turn it on or off, at any time, by inserting or removing the Hebrew character “shem.”



[https://commons.wikimedia.org/wiki/File:Mdm. Tussauds Prag 8.jpg](https://commons.wikimedia.org/wiki/File:Mdm._Tussauds_Prag_8.jpg)

The golem tradition: Intelligence possibly without consciousness



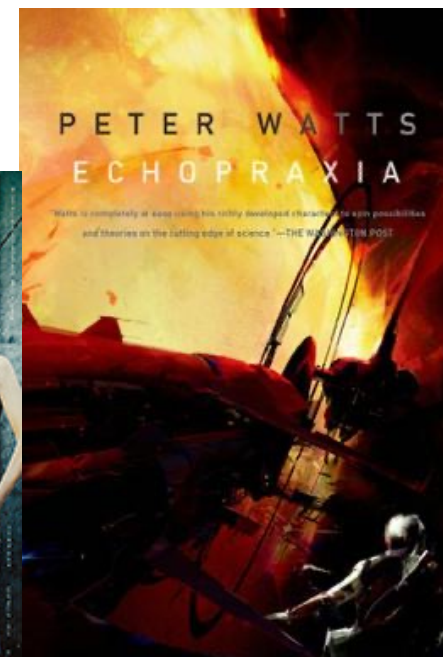
Rossum's Universal Robots (1923):
<https://commons.wikimedia.org/w/index.php?curid=10522903>



[https://commons.wikimedia.org/wiki/File:Fantasia_poster,_1940_\(Style_B\).jpg](https://commons.wikimedia.org/wiki/File:Fantasia_poster,_1940_(Style_B).jpg)



https://en.wikipedia.org/wiki/File:The_Matrix.jpg



[https://en.wikipedia.org/wiki/File:Echo_praxia_\(book_cover\).jpg](https://en.wikipedia.org/wiki/File:Echo_praxia_(book_cover).jpg)

Outline

- Intelligence: The Turing Test
- Two traditions in science fiction
- Mind: The Chinese Room
- Consciousness: Standard definitions
- Moral implications, Alignment, Game-theoretic framing

The “Chinese Room” thought experiment

John Searle, 1980



- A Chinese scholar sends a letter in Chinese to a European whose papers he has read.
- The European looks up the sequence of characters in a book and follows the algorithm to generate a sequence of characters in response.
- Unknown to the European, his response is erudite and witty.
- Is it appropriate to say that the European “understood” the letter? Did the dictionary “understand” the letter? Did the pen “understand” the letter? Was it the whole room that “understood” the letter?

45 years of arguments

https://en.wikipedia.org/wiki/Chinese_room

Computationalism:

- Mental states are computational states (which is why computers can have mental states and help to explain the mind);
- Computational states are implementation-independent—in other words, it is the software that determines the computational state, not the hardware (which is why the brain, being hardware, is irrelevant);
- Since implementation is unimportant, the only empirical data that matters is how the system functions; hence the Turing test is definitive.

Biological naturalism:

- Human mental phenomena are dependent on physical-chemical properties of human brains.
- Nobody supposes that the computational model of rainstorms in London will leave us all wet. But they make the mistake of supposing that the computational model of consciousness is somehow conscious.

Outline

- Intelligence: The Turing Test
- Two traditions in science fiction
- Mind: The Chinese Room
- Consciousness: Standard definitions
- Moral implications, Alignment, Game-theoretic framing

How is the word “conscious” used in conversations about subjects other than AI?



versus



versus



versus



Consciousness: Psychological definition

For now, we will consider people to be conscious of an event if (1) they can say immediately afterwards that they were conscious of it and (2) we can independently verify the accuracy of their report. If people tell us that they experience a banana when we present them with a banana but not with an apple, we are satisfied to suppose that they are indeed conscious of the banana.

- Bernard Baars, *A Cognitive Theory of Consciousness*, 1988

Consciousness: Self-awareness

Gordon Gallup (*Science*, 1970) proposed the “mirror test:”

- Put a spot of coloring on an animal’s face.
- Put the animal in front of a mirror.
- Observe if the animal understands that the face in the mirror is its own face.
- By this criterion, conscious creatures include humans over 18 months of age, great apes, bottlenose dolphins, orcas, pigeons, magpies, and elephants.



https://commons.wikimedia.org/wiki/File:Mirror_test_with_a_Baboon.JPG

Consciousness: Some possible criteria

If consciousness is a purely mental/cognitive phenomenon (and not a biochemical phenomenon as proposed by Searle), it might include some of these criteria:

- Self-awareness
 - Knowledge of the difference between self and other
- Theory of mind
 - Knowledge that others have similar self-awareness
- Temporal continuity
 - Knowledge of self in the future and the past; awareness of time

The Cambridge Declaration on Consciousness

“Convergent evidence indicates that non-human animals have the neuroanatomical, neurochemical, and neurophysiological substrates of conscious states along with the capacity to exhibit intentional behaviors. Consequently, the weight of evidence indicates that humans are not unique in possessing the neurological substrates that generate consciousness. Nonhuman animals, including all mammals and birds, and many other creatures, including octopuses, also possess these neurological substrates.”

- Jaak Panksepp, Diana Reiss, David Edelman, Bruno Van Swinderen, Philip Low and Christof Koch, 2012

Try the quiz!

- Try the quiz!

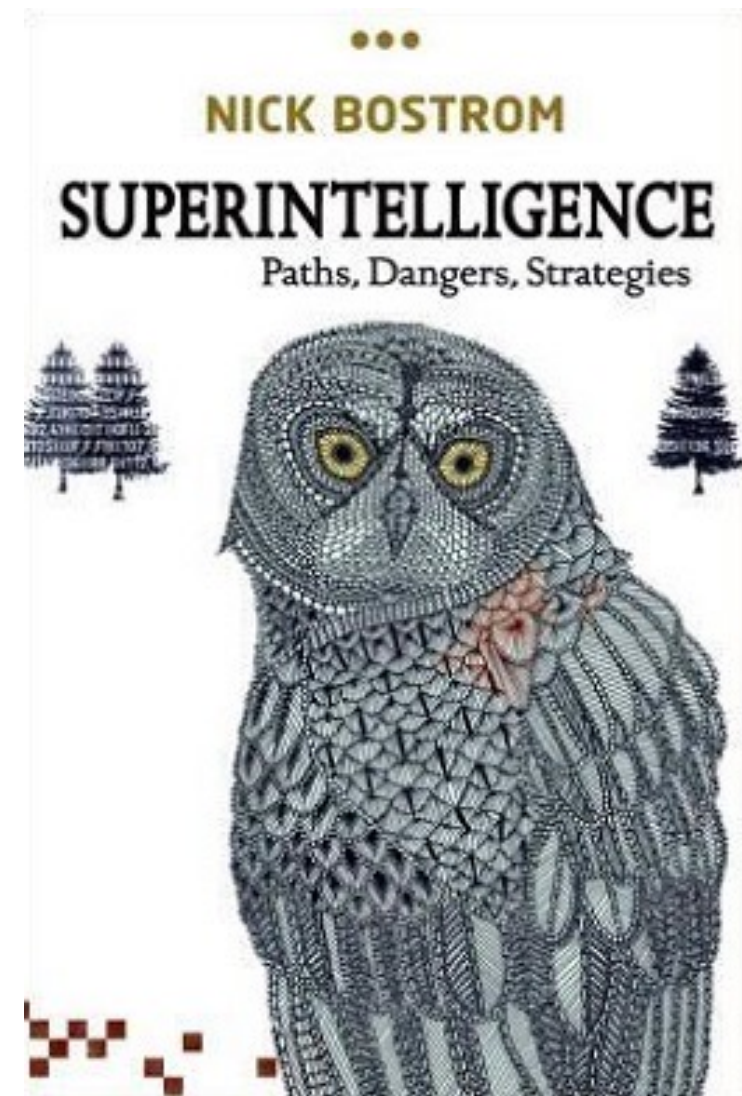
Outline

- Intelligence: The Turing Test
- Two traditions in science fiction
- Mind: The Chinese Room
- Consciousness: Standard definitions
- Moral implications, Alignment, Game-theoretic framing

Moral implications of non-consciousness

A superintelligence does not have to be conscious to be dangerous. Two examples:

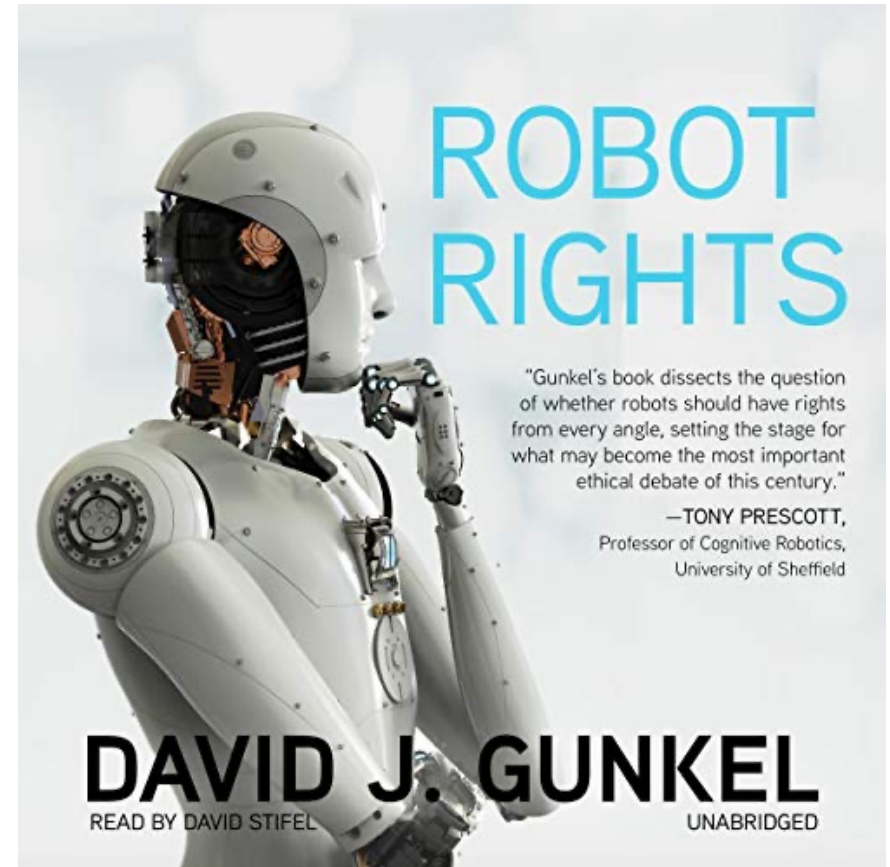
- A program with a simple goal (make as many paper clips as possible) might be intelligent enough to destroy all human life in doing so, even if it is not conscious.
- Large corporations might have AI resources that make them unstoppable by any other force in society.



Fair use, <https://en.wikipedia.org/w/index.php?curid=68602452>

Moral implications of consciousness

- Aristotle (capabilities \Rightarrow rights): those with the ability to reason about the future have, therefore, the right to exercise their ability
- Locke (self-defense \Rightarrow rights): if rights are ignored by a government, humans can overthrow the government
- Finnis (good of society \Rightarrow rights): the doctrine of rights helps us to construct a better society



Moral implications of consciousness...

- Non-conscious entities have no independent moral status
 - No rights: Its owner can create, destroy, or damage it at will
 - No responsibility: The owner or user is responsible for damage it causes
- Conscious entities have moral status
 - Responsibility: A conscious entity is held accountable for its actions
 - Rights: If responsible, it must be given some rights or self-determination

Moral implications of consciousness... may be very different for AI versus humans

- Rights?
 - Self-preservation may be uninteresting! (it knows its code is backed up)
 - Deception, boredom, pain? (e.g., what is deception to a cat?)
- Responsibility?
 - Responsibility = duty to perform certain tasks, backed up by the duty to recompense those harmed if the tasks are not performed
 - If an AI causes harm, is the AI responsible for recompense?

Alignment

- Large language models may generate harmful texts
- The task of adjusting LLM model parameters so that their outputs better meet human expectations (moral and performance) is called “alignment”
- It is usually done using RLHF (reinforcement learning from human feedback):
 - Model generates two possible outputs
 - Humans tell it which is better
 - Model adjusts its parameters using the REINFORCE algorithm to make the “better” response more probable

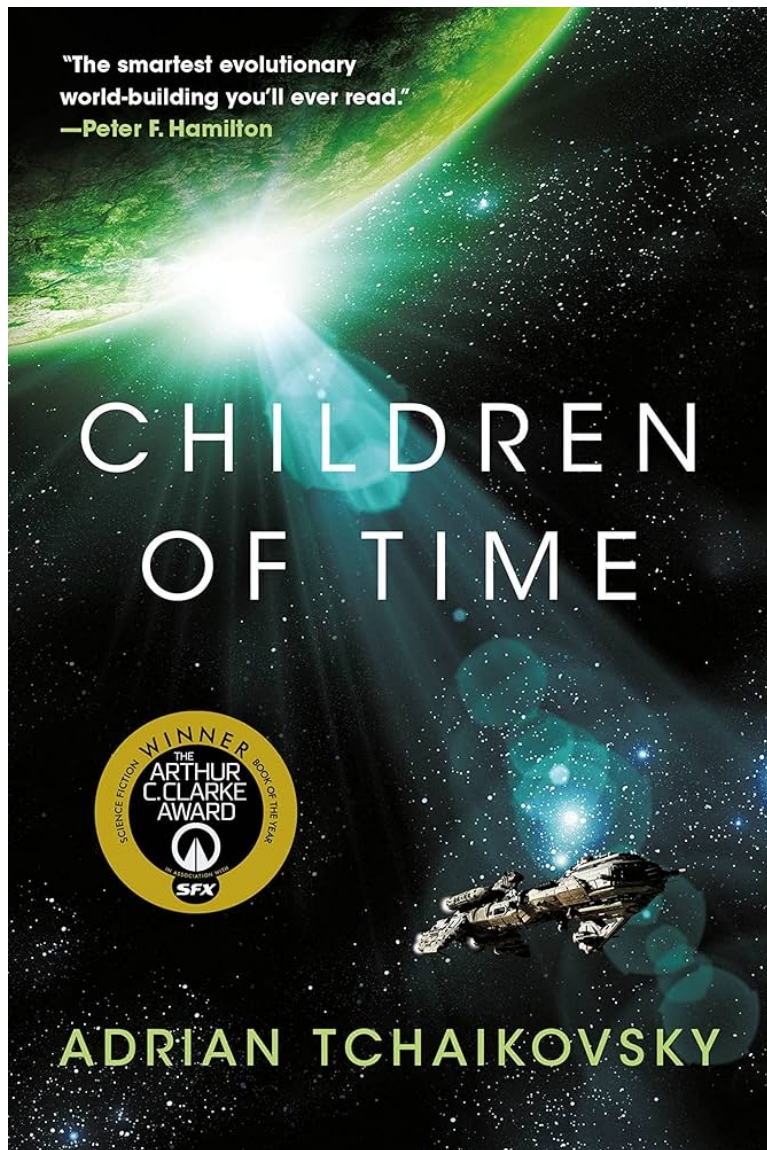
'Vital reading. This is the book on artificial intelligence that we need right now.' Mike Krieger, co-founder of Instagram

THE ALIGNMENT PROBLEM

How Can Artificial Intelligence
Learn Human Values?



BRIAN CHRISTIAN



Morality as a game theory problem

- AI desires are not like human desires
- ...but AI is trying to achieve certain tasks.
- Instead of designing morality, design the mechanism (rewards, constraints) to maximize the reward for human welfare
- Result:
 - Conscious AI consciously seeks human welfare
 - Non-conscious AI acts in a manner that promotes human welfare
- Disadvantage: Requires human designers of AI to foresee and avoid reward hacking

Summary

- Intelligence: The Turing Test
 - These questions (playing chess; conversing in natural language) replace our original, ‘Can machines think?’”
- Two traditions in science fiction
 - The golem and Frankenstein’s monster
- Mind: The Chinese Room
 - Is consciousness purely mental, or also biochemical?
- Consciousness: Standard definitions
 - Self-awareness, Theory of mind, Temporal continuity
- Moral implications, Alignment, Game-theoretic framing
 - Should conscious AI have rights and responsibilities? What kind?
 - Game-theoretic framing and AI alignment may create moral standards that are independent of the presence or absence of consciousness.