

CSU33061 Artificial Intelligence I

Introduction

`www.scss.tcd.ie/Tim.Fernando/AI`

Key Phrases:

Can machines think?

- Turing test & ELIZA effect
- ChatGPT ... (generative AI)

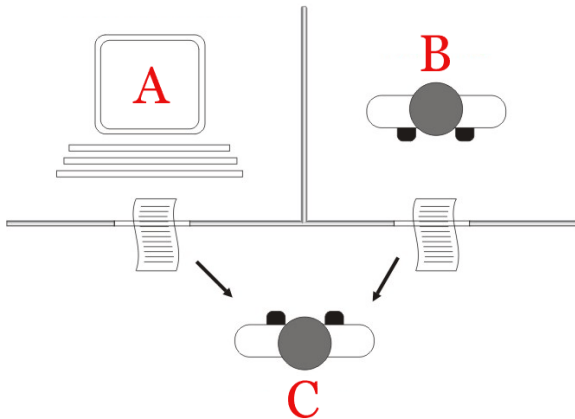
Agent & environment

- Cognitive Revolution & Big Data

Levels of intelligence

Can machines think? (Turing 1950)

Turing test: can C tell A from B?

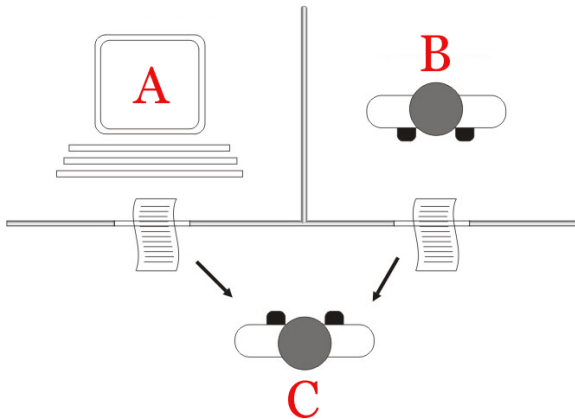


From Wikipedia, (Juan Alberto Sáánchez Margallo)

Intelligence operationalized: subject to testing

Can machines think? (Turing 1950)

Turing test: can C tell A from B?



From Wikipedia, (Juan Alberto Sáánchez Margallo)

Intelligence operationalized: subject to testing and cheating

ELIZA (Weizenbaum, 1964-66): artful deception

- use pattern matching and substitution to fake it (e.g. Prolog)

ELIZA (Weizenbaum, 1964-66): artful deception

- use pattern matching and substitution to fake it (e.g. Prolog)

ELIZA effect: humans anthropomorphise computers

ELIZA (Weizenbaum, 1964-66): artful deception

- use pattern matching and substitution to fake it (e.g. Prolog)

ELIZA effect: humans anthropomorphise computers

CAUTION: programs are quite different from humans

Siri rage (Urban dictionary):

When you get enraged because Siri just doesn't get it.

ELIZA (Weizenbaum, 1964-66): artful deception

- use pattern matching and substitution to fake it (e.g. Prolog)

ELIZA effect: humans anthropomorphise computers

CAUTION: programs are quite different from humans

Siri rage (Urban dictionary):

When you get enraged because Siri just doesn't get it.

Fooling fools is easy

ELIZA (Weizenbaum, 1964-66): artful deception

- use pattern matching and substitution to fake it (e.g. Prolog)

ELIZA effect: humans anthropomorphise computers

CAUTION: programs are quite different from humans

Siri rage (Urban dictionary):

When you get enraged because Siri just doesn't get it.

Fooling fools is easy

though even fools can get angry.

ELIZA (Weizenbaum, 1964-66): artful deception

- use pattern matching and substitution to fake it (e.g. Prolog)

ELIZA effect: humans anthropomorphise computers

CAUTION: programs are quite different from humans

Siri rage (Urban dictionary):

When you get enraged because Siri just doesn't get it.

Fooling fools is easy

though even fools can get angry.

Programs have improved hugely

ELIZA (Weizenbaum, 1964-66): artful deception

- use pattern matching and substitution to fake it (e.g. Prolog)

ELIZA effect: humans anthropomorphise computers

CAUTION: programs are quite different from humans

Siri rage (Urban dictionary):

When you get enraged because Siri just doesn't get it.

Fooling fools is easy

though even fools can get angry.

Programs have improved hugely,
humans largely not.

ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

The New York Times | <https://www.nytimes.com/2022/12/05/technology/chatgpt-ai-twitter.html>

THE SHIFT

The Brilliance and Weirdness of ChatGPT

A new chatbot from OpenAI is inspiring awe, fear, stunts and attempts to circumvent its guardrails.

From OpenAI:

Limitations

- ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers. Fixing this issue is challenging

From OpenAI:

Limitations

- ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers. Fixing this issue is challenging

From the NY Times:

ChatGPT isn't perfect, by any means. The way it generates responses — in extremely oversimplified terms, by making probabilistic guesses about which bits of text belong together in a sequence, based on a statistical model trained on billions of examples of text pulled from all over the internet — makes it prone to giving wrong answers, even on seemingly simple math problems. (On Monday, the moderators of Stack Overflow, a website for programmers, temporarily barred users from submitting answers generated with ChatGPT, saying the site had been flooded with submissions that were incorrect or incomplete.)

From OpenAI:

Limitations

- ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers. Fixing this issue is challenging

From the NY Times:

ChatGPT isn't perfect, by any means. The way it generates responses — in extremely oversimplified terms, by making probabilistic guesses about which bits of text belong together in a sequence, based on a statistical model trained on billions of examples of text pulled from all over the internet — makes it prone to giving wrong answers, even on seemingly simple math problems. (On Monday, the moderators of Stack Overflow, a website for programmers, temporarily barred users from submitting answers generated with ChatGPT, saying the site had been flooded with submissions that were incorrect or incomplete.)

From The Guardian:

Its current knowledge base ends in 2021, rendering some queries and searches useless.

ChatGPT can also give entirely wrong answers and present misinformation as fact, writing “plausible-sounding but incorrect or nonsensical answers”, the company concedes.

OpenAI says that fixing this issue is difficult because there is no source of truth in the data they use to train the model and supervised training can also be misleading “because the ideal answer depends on what the model knows, rather than what the human demonstrator knows”.

Over capacity

ChatGPT is at capacity right now

[Get notified when we're back](#)

Write an inspirational speech about the status of ChatGPT.

My fellow citizens, the time has come. The future is here, and it is called ChatGPT. This revolutionary AI chatbot website is here to change the world, and we must embrace it with all our strength and determination.

I know that many of you are eager to try ChatGPT for yourselves. And I assure you, we are doing everything in our power to make that happen. Our team is working tirelessly to accommodate the incredible demand for this groundbreaking technology.

But I want to remind you that greatness is never achieved without sacrifice. And the road to the future of AI will not be without its challenges. So please, be patient. Trust that the right time for you to try ChatGPT will come.

And when it does, let us seize this opportunity with all our might. Let us embrace the power of AI and use it to build a brighter, more prosperous future for ourselves and for generations to come. Thank you.



Do you have a theory of the world?



I am an AI trained to process and generate natural language text. I do not have beliefs, theories or desires. My function is to assist users in answering questions and completing tasks to the best of my ability using the knowledge and capabilities that have been programmed into me. Is there anything else I can help you with?



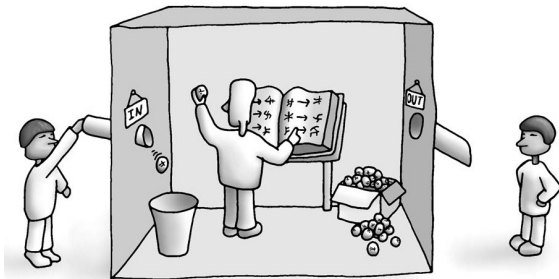
When you think, do you think based on a conception of the universe and how it works, or are you simply working via association?



As an AI, I don't think or experience in the way that humans do. I am a machine that processes input based on a set of rules and algorithms that have been programmed into me. These rules and algorithms allow me to process input, like text, and generate output, like a response to a question. I can understand and process natural language, but I don't have a personal experience of the universe, consciousness or any personal beliefs. My understanding of the world is based on the information and knowledge that has been programmed into me by humans.

My understanding is also limited by the parameters and information that I had at the time my knowledge cut off, which is 2021, I will not be able to make claims on new events or new discoveries that happened after that date.

Chinese room argument (Searle's thought experiment)

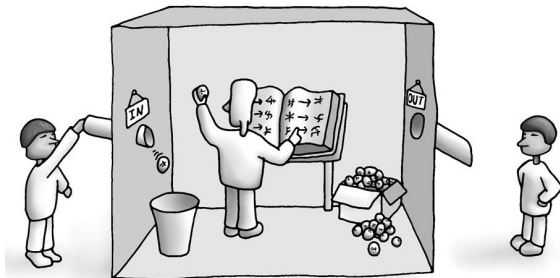


From <http://america.pink/images/9/6/3/2/5/4/en/2-chinese-room.jpg>

- a clerk can follow instructions for communicating in Chinese without understanding Chinese

Strong vs Weak AI

Chinese room argument (Searle's thought experiment)



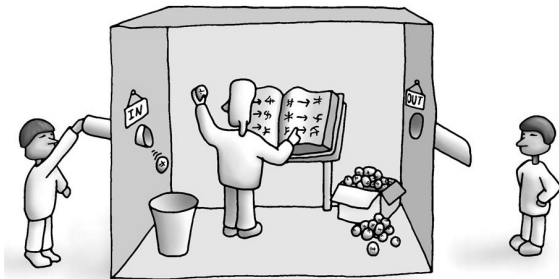
From <http://america.pink/images/9/6/3/2/5/4/en/2-chinese-room.jpg>

- a clerk can follow instructions for communicating in Chinese without understanding Chinese

mindless obedience \neq mind from biological processes

Strong vs Weak AI

Chinese room argument (Searle's thought experiment)



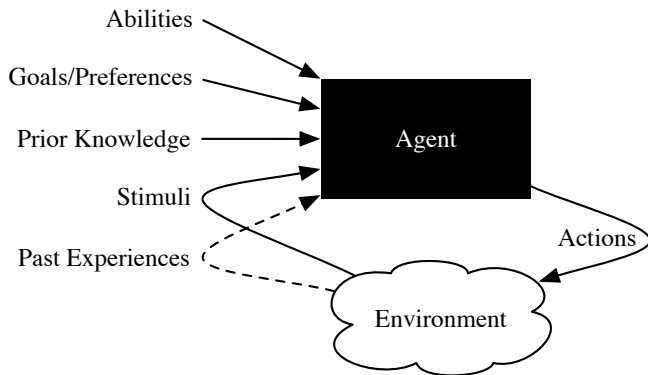
From <http://america.pink/images/9/6/3/2/5/4/en/2-chinese-room.jpg>

- a clerk can follow instructions for communicating in Chinese without understanding Chinese

mindless obedience \neq mind from biological processes

Just do it \rightsquigarrow Black Box, judged by its actions

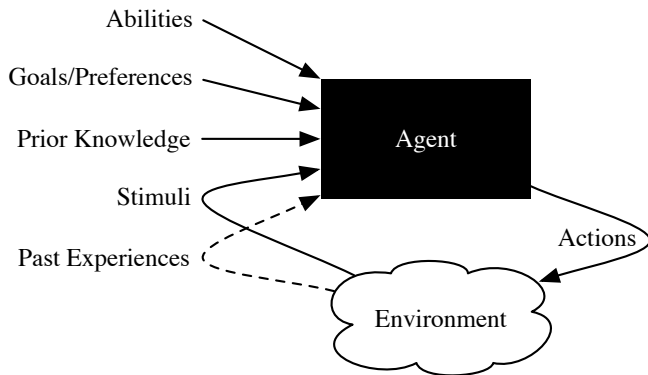
Locating intelligence (black box)



Poole & Mackworth

Intelligence: (abilities, goals, ..., experience) \mapsto action

Locating intelligence (black box)



Poole & Mackworth

Intelligence: (abilities, goals, ..., experience) \mapsto action

Turing test: what to say \rightsquigarrow what to do

Between agent and environment

agent	environment
program	data
Cognitive Revolution	BIG DATA
hard-wired	experienced
rationalist	empiricist
nativist	behaviorist
innate	tabula rasa
nature	nurture

Turing machine &
specialized automaton

Between agent and environment

agent	environment
program	data
Cognitive Revolution	BIG DATA
hard-wired	experienced
rationalist	empiricist
nativist	behaviorist
innate	tabula rasa
nature	nurture

Turing machine &
specialized automaton

Learning (from environment)
trial & error: “data as oil”

Between agent and environment

agent	environment
program	data
Cognitive Revolution	BIG DATA
hard-wired	experienced
rationalist	empiricist
nativist	behaviorist
innate	tabula rasa
nature	nurture

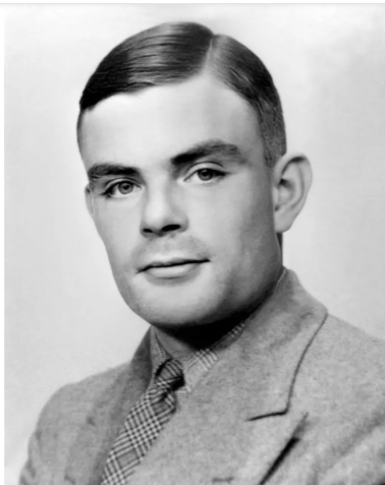
Turing machine &
specialized automaton

Learning (from environment)
trial & error: “data as oil”

Moving target: changing agent & environment

Intelligence comes in many levels

Intelligence comes in many levels from genius



Alan Turing, a British mathematician, proposed in 1950 that the test of machine intelligence would be an ability to conduct a conversation in an indistinguishably human way. Archivio GBB, via Alamy

Intelligence comes in many levels from genius to



H. Frankfurt, *On Bullshit* (Wikipedia)

bullshit is speech intended to persuade without regard for truth. The liar cares about the truth and attempts to hide it; the bullshitter doesn't care if what they say is true or false, but cares only whether the listener is persuaded...

H. Frankfurt, *On Bullshit* (Wikipedia)

bullshit is speech intended to persuade without regard for truth. The liar cares about the truth and attempts to hide it; the bullshitter doesn't care if what they say is true or false, but cares only whether the listener is persuaded...

When asked why he decided to focus on bullshit, he explained:

Respect for the truth and a concern for the truth are among the foundations for civilization. I was for a long time disturbed by the lack of respect for the truth that I observed... bullshit is one of the deformities of these values.

His book *On Bullshit* addresses his concern and makes a distinction between "bullshitters" and liars. He concludes that bullshitters are more insidious: they are more of a threat against the truth than are liars.



technology, with its god-like power ... has allowed a virus of lies to infect each of us, pitting us against each other, bringing out our fears, anger and hate, and setting the stage for the rise of authoritarians and dictators around the world.



technology, with its god-like power ... has allowed a virus of lies to infect each of us, pitting us against each other, bringing out our fears, anger and hate, and setting the stage for the rise of authoritarians and dictators around the world.

Without facts, you can't have truth. Without truth, you can't have trust. Without trust, we have no shared reality, no democracy, and it becomes impossible to deal with our world's existential problems