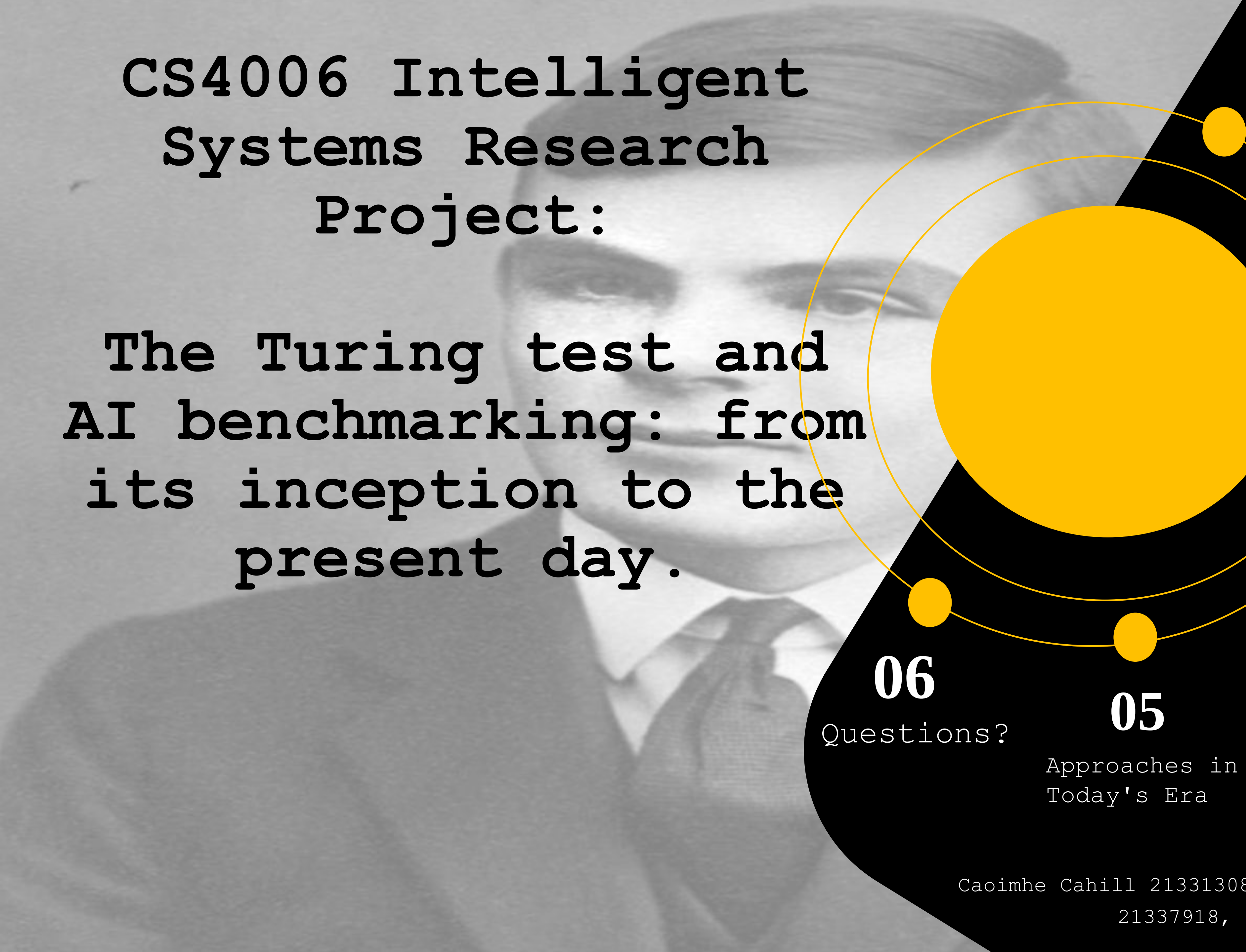
# CS4006 Intelligent Systems Research Project:

# The Turing test and AI benchmarking: from its inception to the present day.



01 The Turing Test

02 Limitations of Turing test

03 Total Turing test

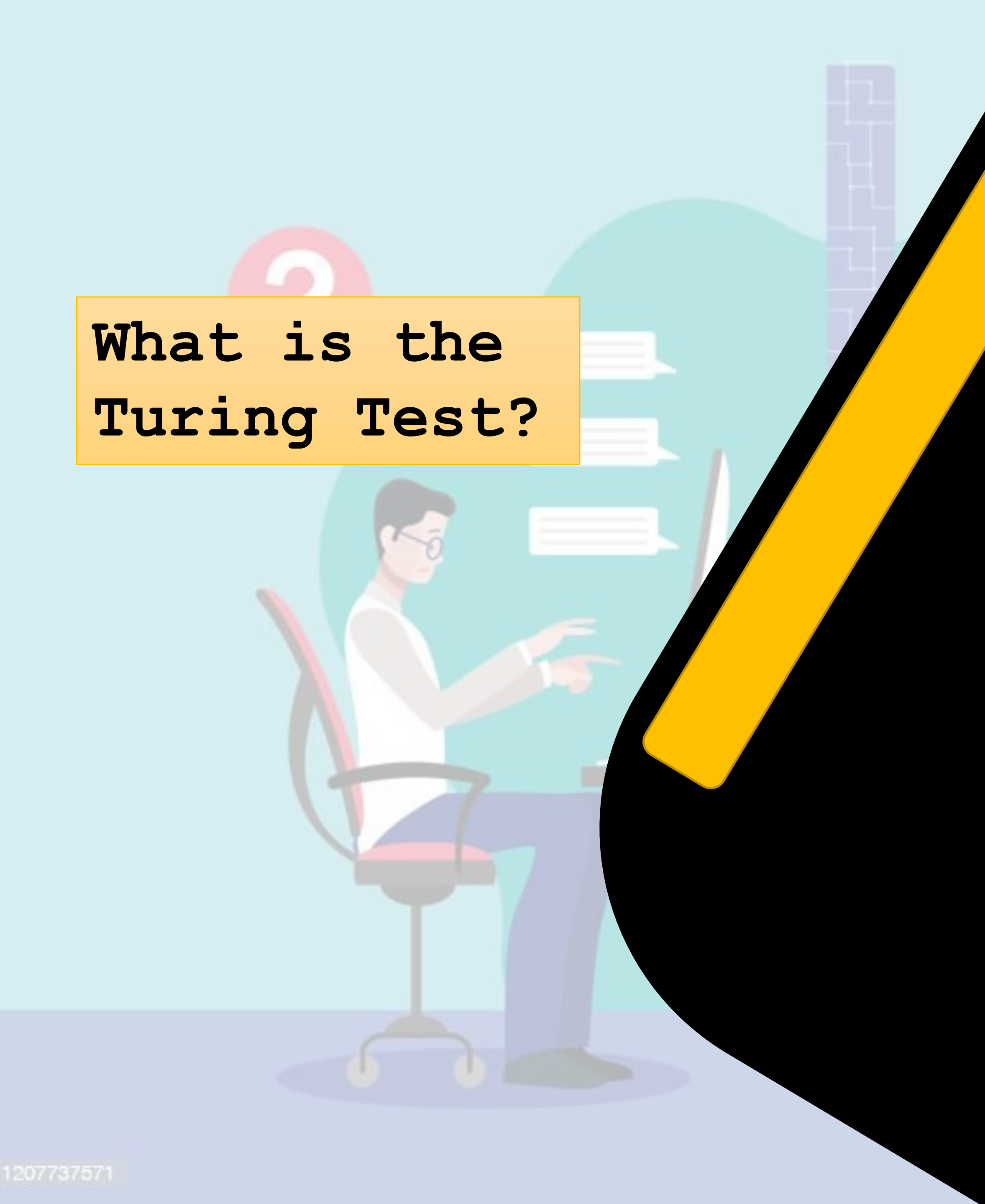
04 Programming a computer for playing chess

05 Approaches in Today's Era

06 Questions?

Caoimhe Cahill 21331308, Olan Healy 21318204, Aaron Maher 21337918, Kevin Collins 21344256

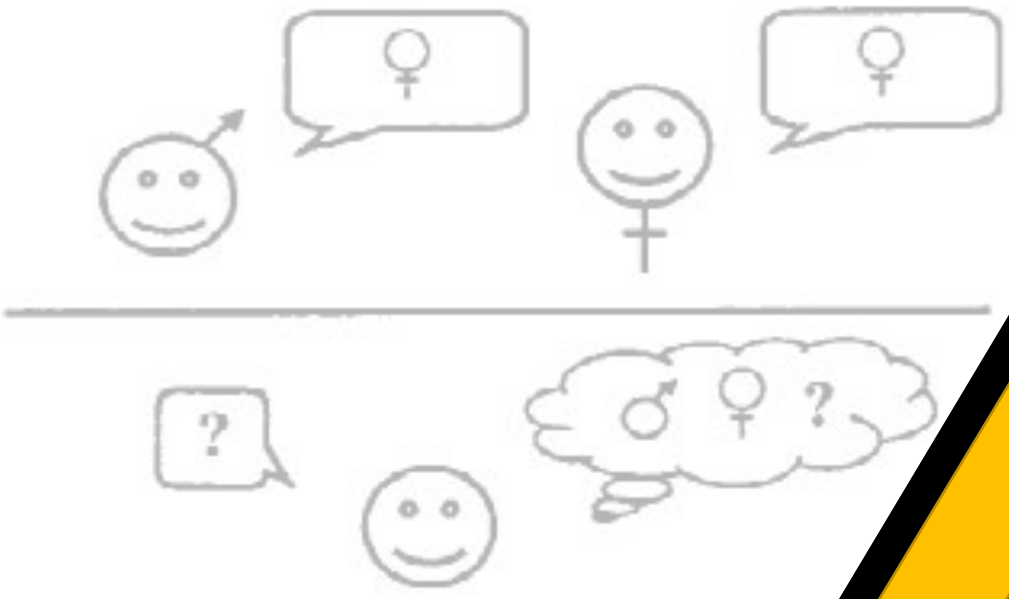
# What is the Turing Test?



The Turing Test is one of the most well-known benchmarks in AI.

Also known as the Imitation Game, the Turing Test is a method of determining whether or not a machine can think like a human being.

Founded by Alan Turing in 1950, the Turing Test aims to solve the question "Can machines think?"



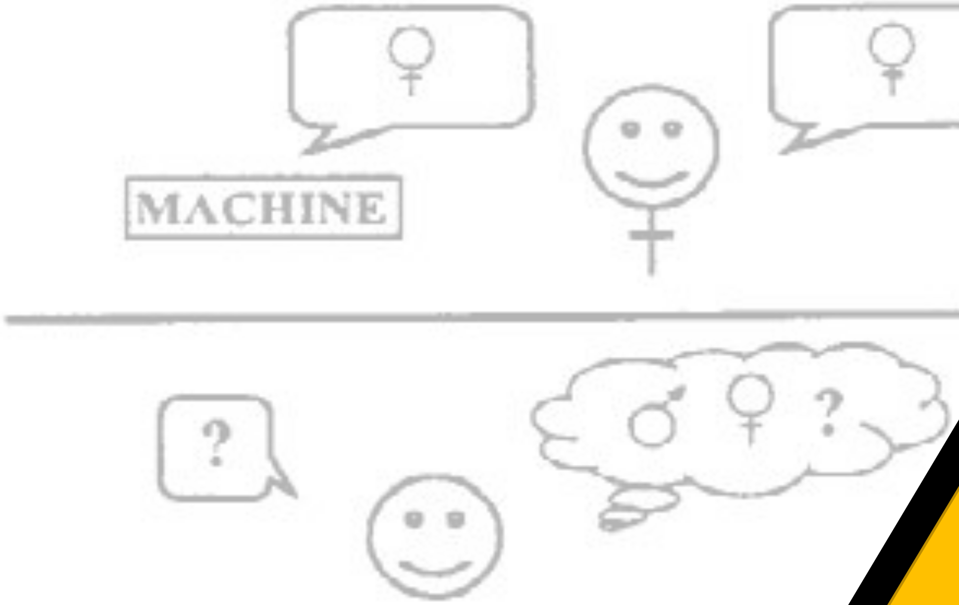
# Turing Test:

## Stage 1

Stage 1:
- A Man
- A Woman
- A Interrogator in a separate room

Objectives:

- The Man: convince the interrogator that he is a woman and the other is not
- The Woman: draw the interrogator to the correct decision, that she is a woman
- The Interrogator: determine which of the two participants is a woman by asking any questions they want



# Turing Test:

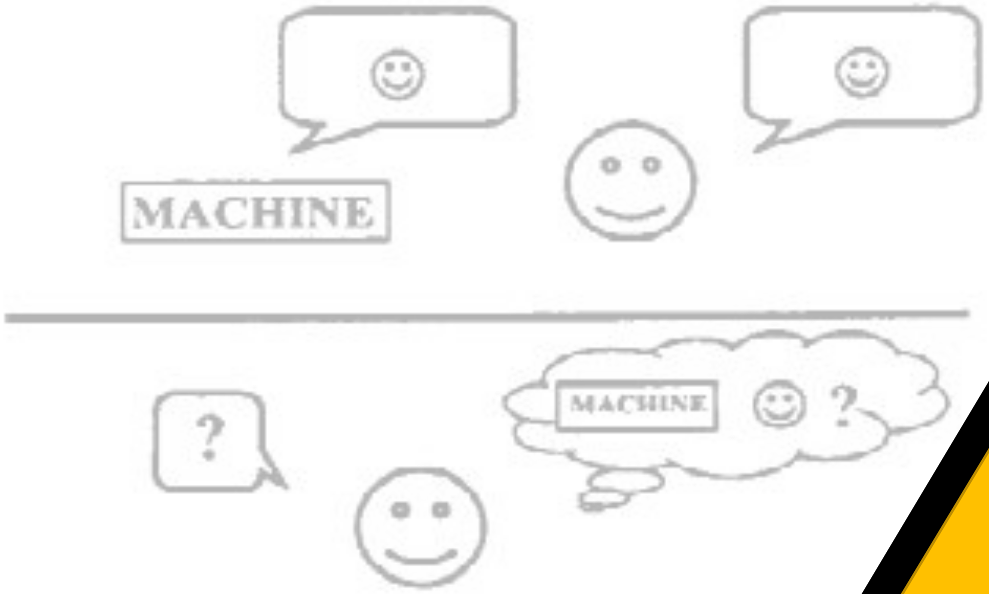
## Stage 2

Stage 2:
- A Machine
- A Woman
- A Interrogator in a separate room

***"What will happen when a machine takes the part of A in this game? Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?"***

**– Turing 1950**

The test is successful if the machine convinces the interrogator that it is the woman.



# Turing Test:

## Today

Today, the test usually describes the woman as a person of either gender and the interrogator must determine which one is the human being, as seen in the diagram

It is also sometimes described as either a person or a machine, and the interrogator must decide whether they are talking to a human or a machine.

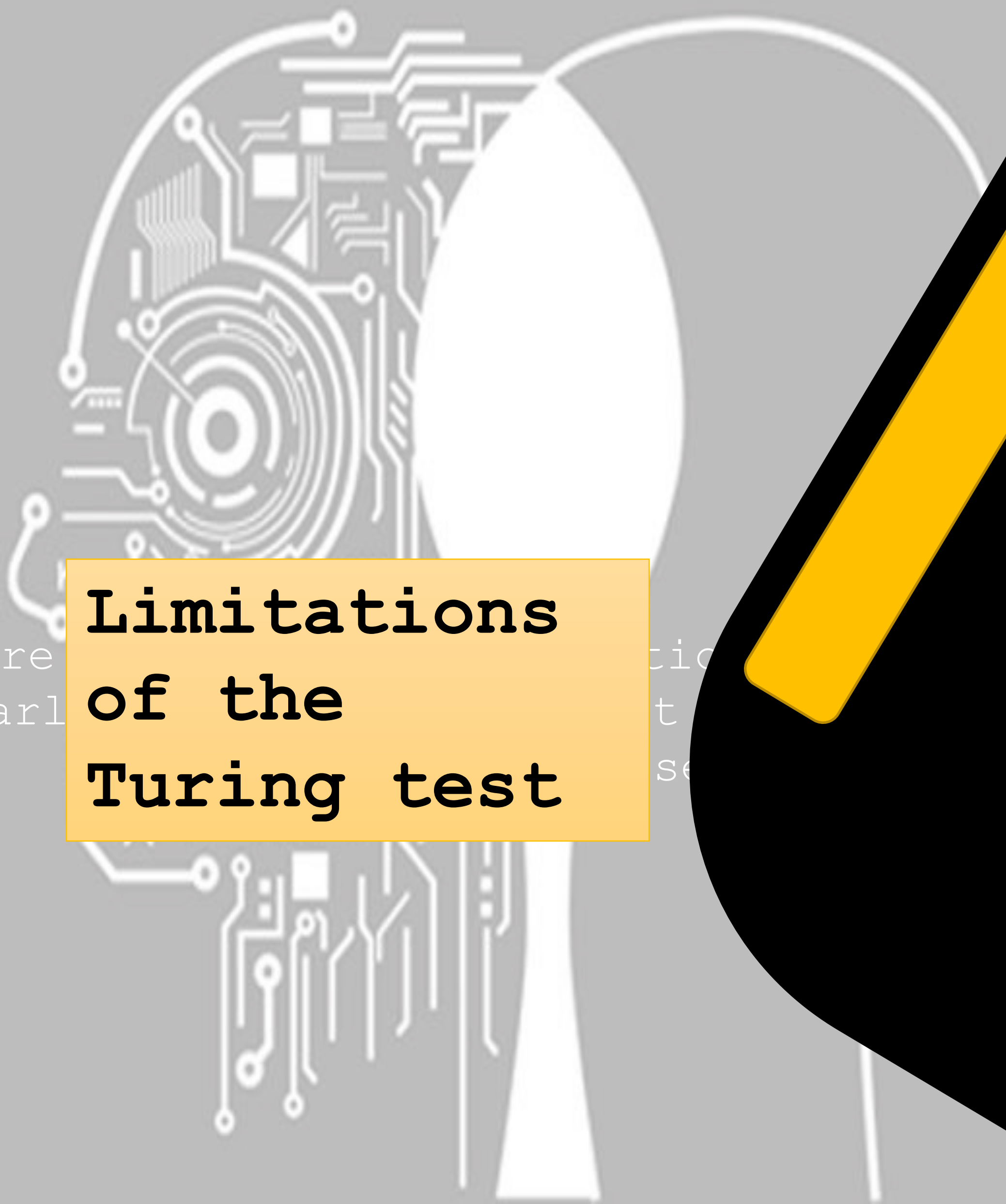
It is generally agreed that the differences don’t change the primary purpose of the Turing Test.

# Passing the Turing Test

It is still argued that no computer has ever passed the Turing test.

However, there have been some strong contenders such as
- Eliza chatbot 1966
- Parry chatbot 1972

The Eugene Goostman computer program in 2014 is argued to have passed the Turing Test.
However, there is a lot of doubt and criticism surrounding this.

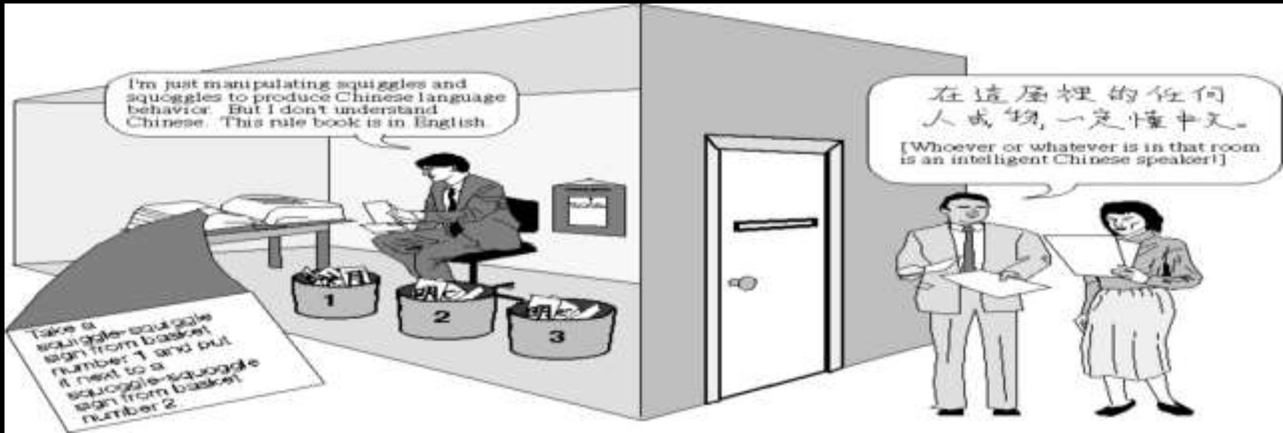


# Limitations of the Turing test

No doubt Alan turing’s test played a major role in checking an AI system’s intelligence is on par with that of a human.

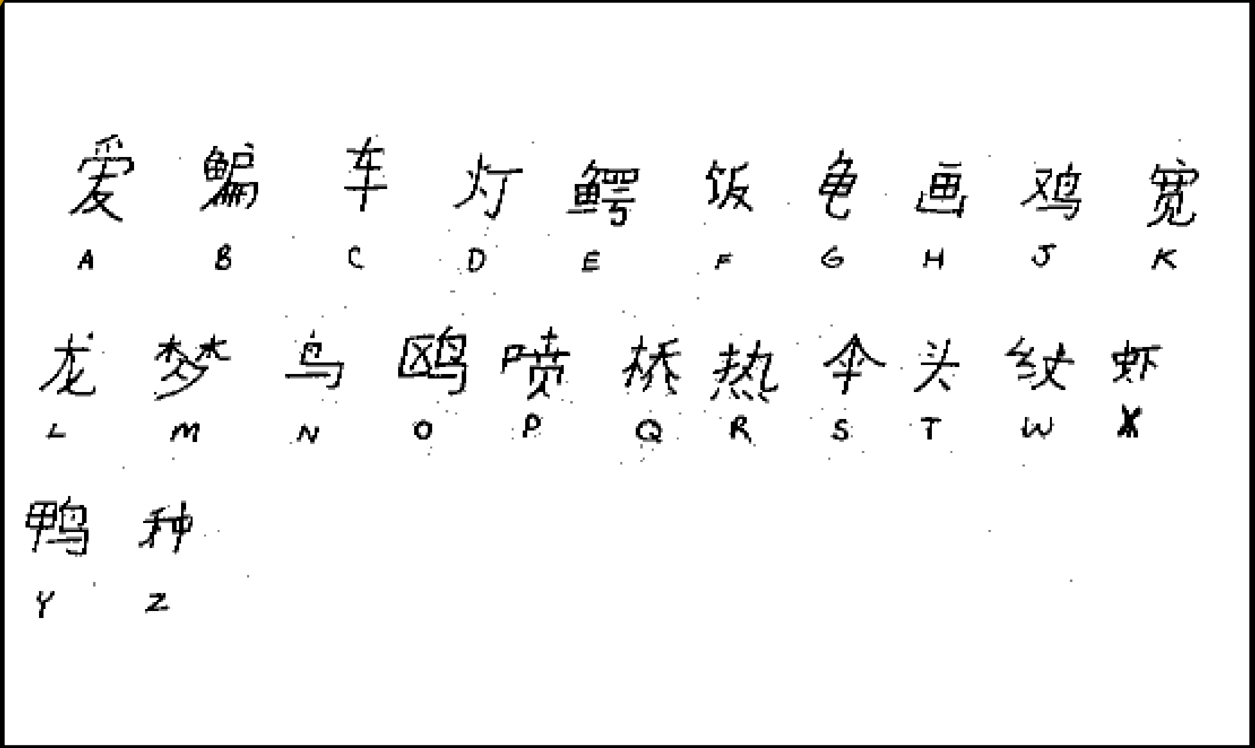
There are also some limitations with his test. A philopspher John Searle proposed an argument against the turing test known as the ‘Chinese Room Argument’.

Its basis was that pretty much an artificial system doesn’t actually show intelligence, it simulates it.


I'm just manipulating squiggles and squoggles to produce Chinese language behavior. But I don't understand Chinese. This rule book is in English.
在這屋裡的任何人或物，一定懂中文。
[Whoever or whatever is in that room is an intelligent Chinese speaker!]
Take a squiggle-squiggle sign from basket number 1 and put it next to a squoggle-squoggle sign from basket number 2
1
2
3

# Limitation:

## The Chinese Room Experiment



| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 爱 | 鳊 | 车 | 灯 | 鳄 | 饭 | 龟 | 画 | 鸡 | 宽 | |
| A | B | C | D | E | F | G | H | J | K | |
| 龙 | 梦 | 鸟 | 鸥 | 喷 | 桥 | 热 | 伞 | 头 | 纹 | 虾 |
| L | M | N | O | P | Q | R | S | T | W | X |
| 鸭 | 种 | | | | | | | | | |
| Y | Z | | | | | | | | | |

# Limitation:

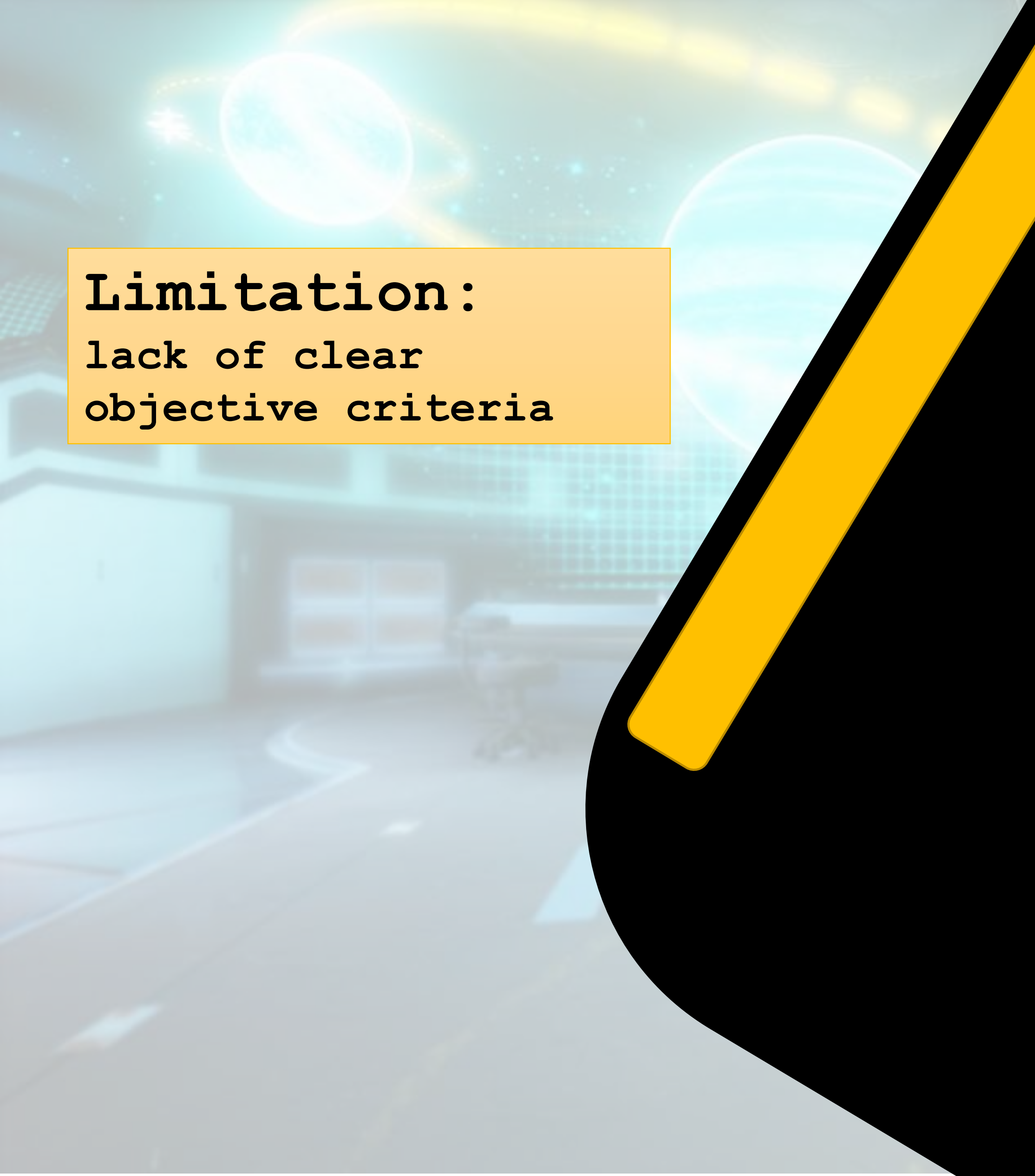
## The Chinese Room Experiment

Let's show his experiment:

The Ai system would only be following instructions not actually proving it has human like

This highlights the need to evaluate AI systems based on their **comprehension of concepts**, rather than just their linguistic capabilities.

AI systems should be able to **"observe without evaluating"** to show they have human-like intelligence



## Limitation:

**lack of clear objective criteria**

It relies solely on subjective judgement of human evaluator. Could lead to bias, inaccuracies

It is more of an existence proof rather than a performance test

Future **benchmarking techniques** should be unbiased and asses an AI system performance in a number of different ways

A benchmark was made with **the Total Turing test** by Harnard in 1991

This made needed improvements on the turing test by assessing its

- Linguistic
- motor
- visual
- auditory

## Benchmark made:

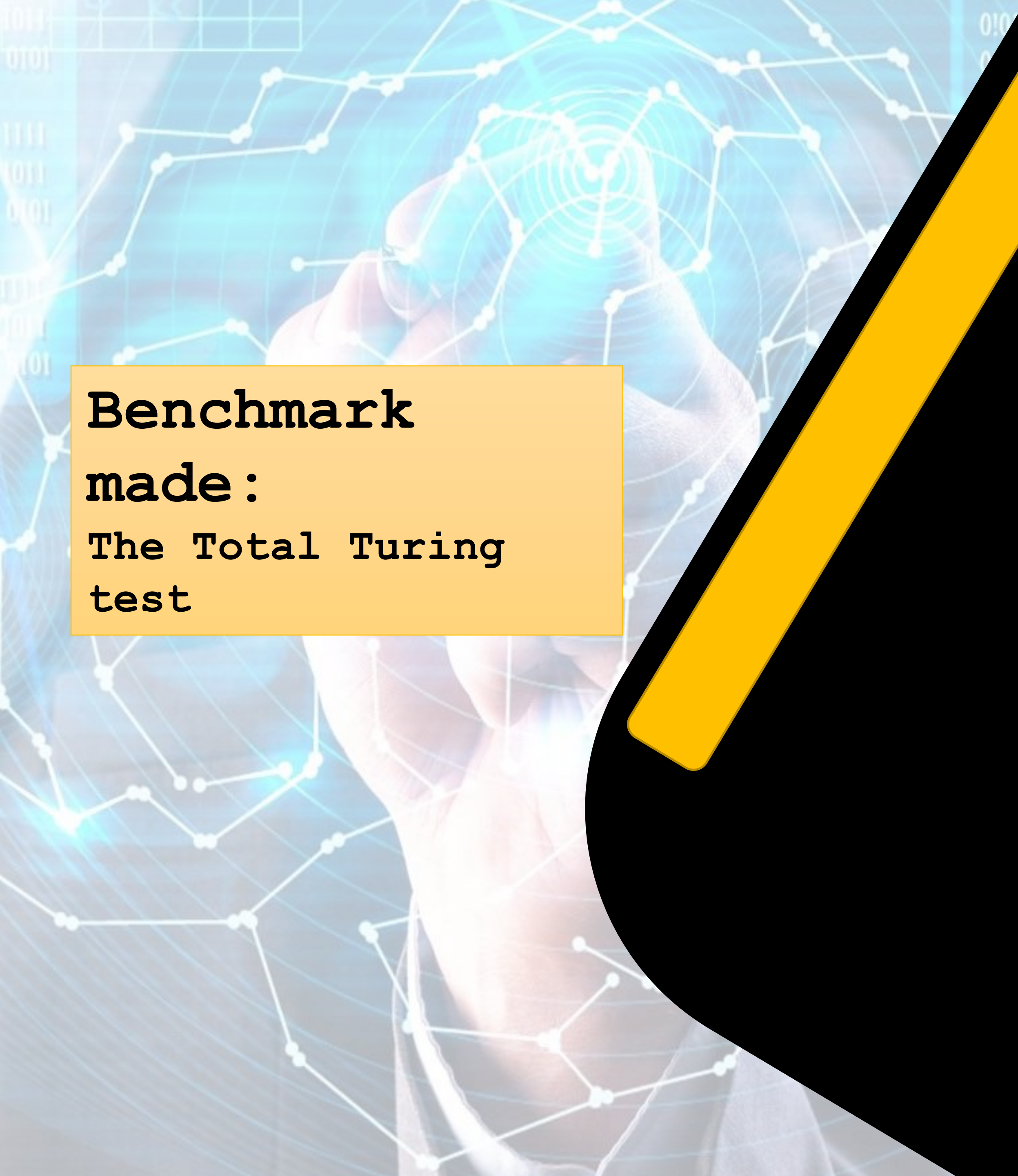
**The Total Turing test**

**LINGUSTIC:** same idea as original turing test.

**MOTOR:** looks at its capabilities of performing basic human tasks eg assembling furnite

**VISUAL:** present an object in front of the AI system, should be able to recognise it e.g. an apple. Present scene e.g. a busy street and it should be able to recognise it.

**AUDITORY:** should be able to understand audio clips and classify them eg types of music



## Benchmark made:

**The Total Turing test**

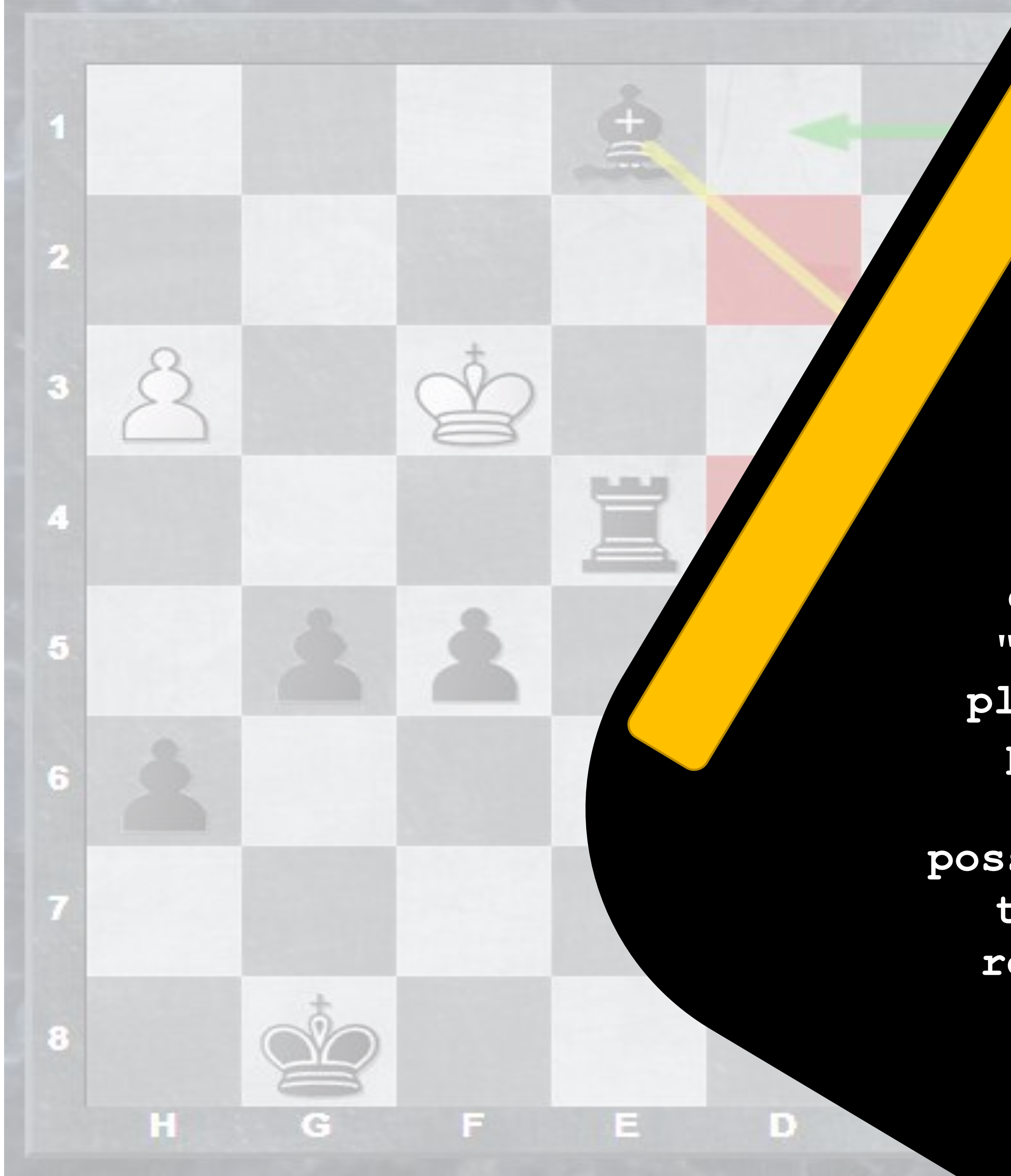
More fully defined version of the turing test.

**“A machine that passes the Total Turing Test should be able to do anything a human can do, because the test requires the machine to be indistinguishable from a human”**
**~ (Russell and Norvig, 2020, p. 1029).**

# Programming a Computer for Playing Chess

"Although perhaps of no practical importance, the question is of theoretical interest, and it is hoped that a satisfactory solution of this problem will act as a wedge in attacking other problems of a similar nature and of greater significance."

- *Claude E. Shannon*



**Suggested two approaches:**

- Type A, calculate and evaluate every option, computationally heavy but sure to find the best outcome
- Type B, eliminate options that aren't expected to yield good results, only evaluate options that show merit, human-like

> **“Chess is generally considered to require "thinking" for skilful play; a solution of this problem will force us either to admit the possibility of a mechanised thinking or to further restrict our concept of "thinking".”**
>
> - ***Claude E. Shannon***

# Deep Blue

As far as chess-computing goes, the benchmark is pretty straightforward; be better than your opponent

Deep Blue was designed at IBM by Feng-Hsiung Hsu and his team

Can perform 200,000,000 evaluations per second

# Deep Blue VS Garry Kasparov

In 1997, Deep Blue fought against Garry Kasparov, the undisputed chess world champion

Score was tied at 2.5 points each going into the final game

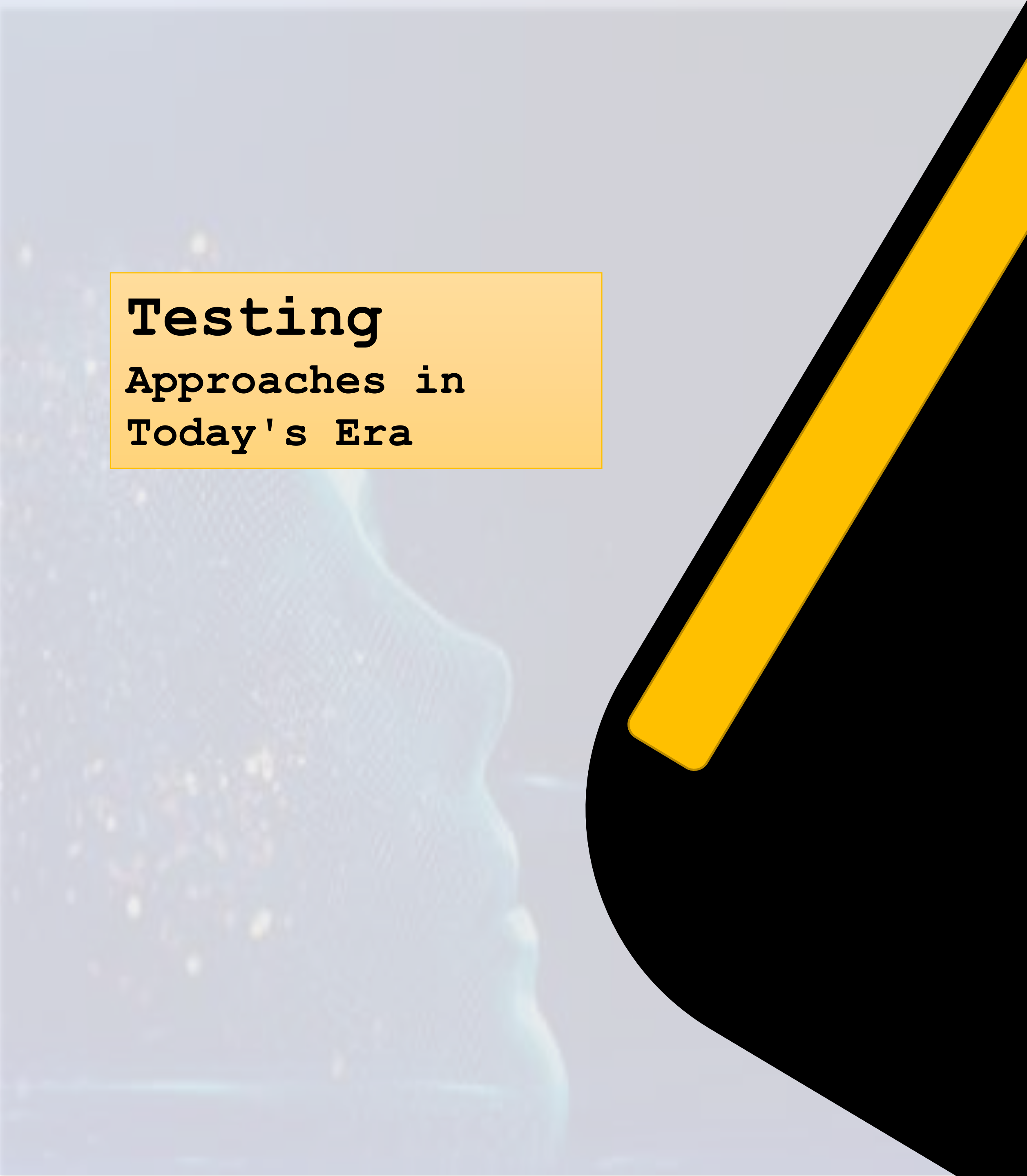
Kasparov made a risky play, expecting Deep Blue to not know how to respond properly. It did know and won.

# Modern Chess AI

Deep Blue retired, but chess computing continues to grow

Stockfish is the most successful modern day chess AI, having won over 30+ tournaments

Stockfish itself has become a benchmark for other chess AIs to test their ability, e.g. Google's AlphaZero



# Testing Approaches in Today's Era

Breakthroughs in areas such as:

- Reinforcement learning
- Computer Vision
- Natural Language Processing

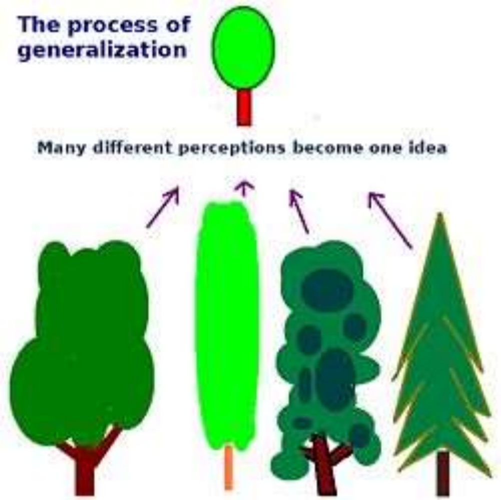
As AI systems grow more advanced, assessing their performance becomes increasingly important.

Who has nearly cried getting ChatGPT to understand what you want from it?

# Reinforcement Learning

## Generalisation Capabilities

Generalisation capabilities refers to a model's ability to learn from examples it **has** seen before to correctly understand and work with new, similar examples that it **hasn't** seen before.


The process of generalization
Many different perceptions become one idea

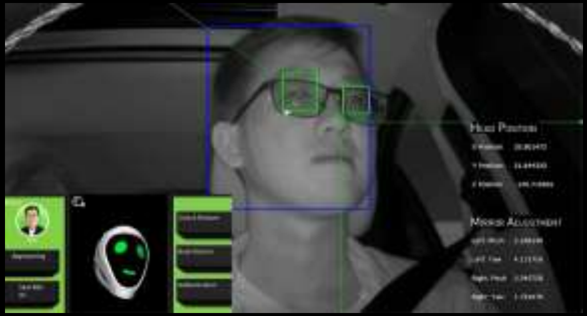
# Modern Safety Testing

## Computer Vision

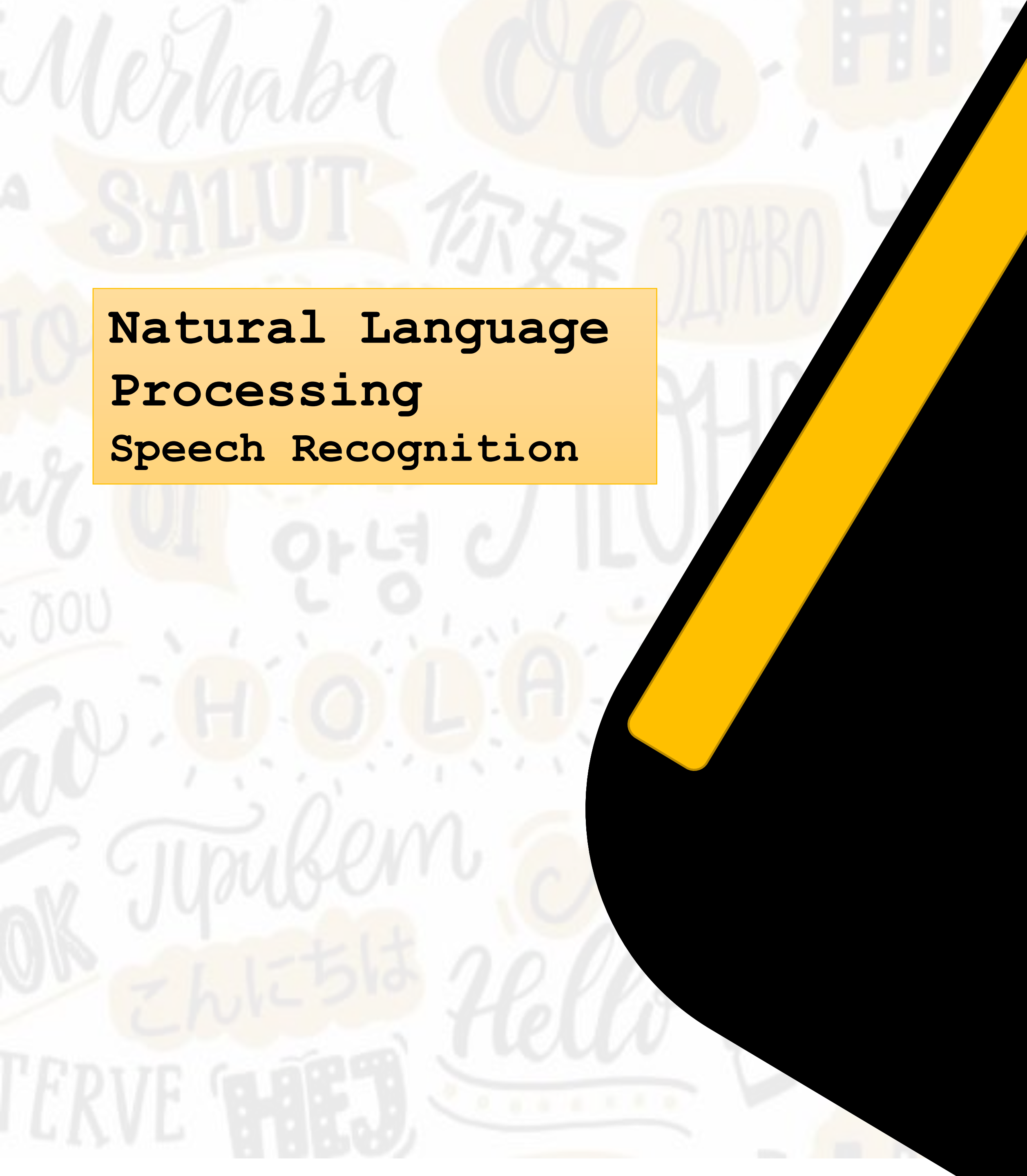
Look at your phone. Most now a days see you,  recognize your face, and open.
But would you want it to open for anybody's face?
What about a photo of you?

Hostile cases:

- Distorted photos
- Glasses
- Headphones
- Camera Glare



But do we want our cameras always watching us?



# Natural Language Processing

## Speech Recognition

**Automated Speech Recognition (ASR)** - The crucial technology that enables AI systems to translate spoken language into written text.

Tested by a range of speech situations, such as different accents, dialects, and noise levels.

Let’s test it!

It’s improving rapidly but still I don’t think it will be fully trusted for many more years.

# Any Questions?