Introduction to AI, fairness and bias

The pictures in these slides were generated using ChatGPT and Canva

AI - Introduction

Artificial Intelligence around us

Artificial Intelligence refers to technological applications designed to simulate intelligence, such as the ability to learn, take decisions, and interact with surrounding environments

AI applications around us include:

- Recommendation Systems (YouTube, Instagram…)
- Virtual Assistants (Alexa, Siri…)
- Generative algorithms (ChatGPT, Dall-E…)
- Autonomous vehicles
- Playing agents (Chess, Go, various videogames…)

Did you know that the best chess player in the world is a computer program called Stockfish? Its rating (a measure of chess proficiency) is more than 3600. The best ever human player's rating is less than 2900.

A real scenario

- In 2014, a team at Amazon started working on an algorithm that would automatically rank candidates for hiring
- Advantage: Amazon receives lots of applications and it has an interest in identifying the best candidates quickly

RETAIL OCTOBER 10, 2018 / 4:04 PM / UPDATED 5 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

8 MIN READ

By Jeffrey Dastin

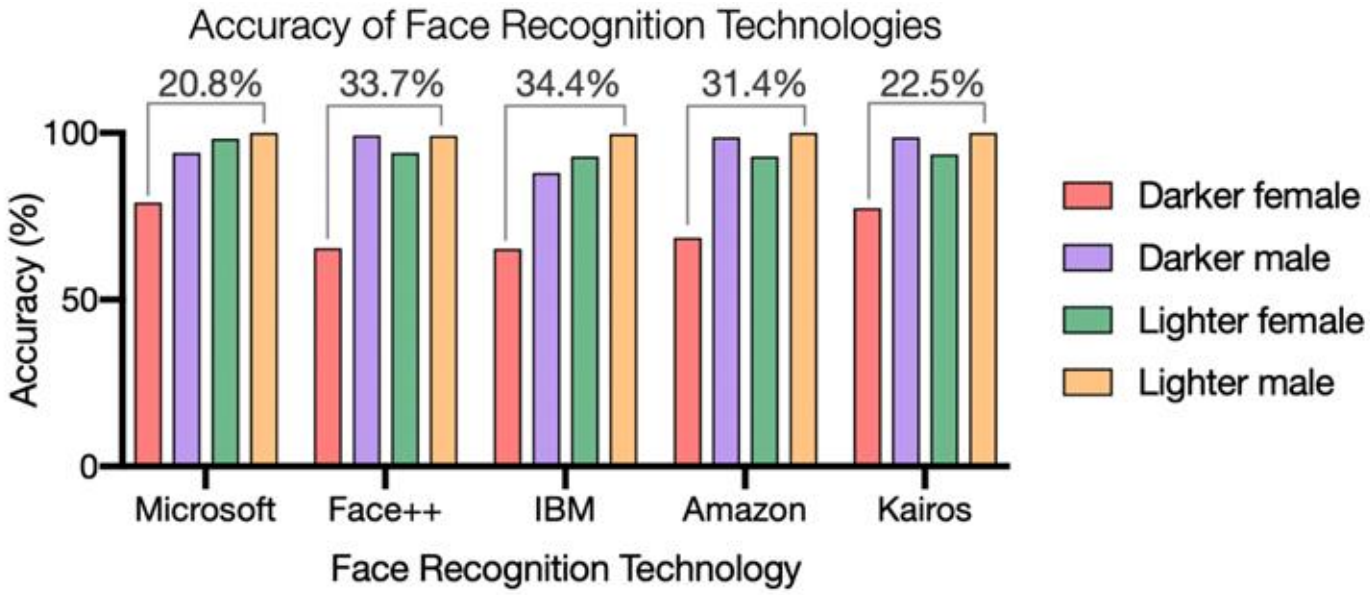
source: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G/

AI under scrutiny

Today, there is a call for increased attention toward responsible use of AI applications, including a focus on aspects such as:

- **Fairness =** the idea that outcomes of an AI application must be equitable, i.e. no group should be discriminated against/receive a different treatment
- **Accountability =** clearly identify people responsible for the outcomes and derived consequences
- **Transparency =** the ability to explain outcomes and decisions, as well as transparency in data acquisition and provenance

More examples: fairness



source: https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

UK police use of live facial recognition unlawful and unethical, report finds

Study says deployment of technology in public by Met and South Wales police failed to meet standards

There are concerns about privacy and racial bias in police deployment of live facial recognition. Photograph: Stefan Rousseau/PA

Source: https://www.theguardian.com/technology/2022/oct/27/live-facial-recognition-police-study-uk

More examples: accountability

When technologies are involved, it is more difficult to pinpoint who is responsible when accidents happen.

SAFETY —

Autopilot was active when a Tesla crashed into a truck, killing driver

NTSB report says driver engaged Autopilot 10 seconds before the deadly crash.

TIMOTHY B. LEE - 5/16/2019, 10:10 AM

This is a legislative gap. Unfortunately, technology move at a much faster pace than the law, so these episodes are not infrequent.

source: https://arstechnica.com/cars/2019/05/feds-autopilot-was-active-during-deadly-march-tesla-crash/

The issue of transparency

RESEARCH ARTICLE | Open Access

Explaining Why the Computer Says No: Algorithmic Transparency Affects the Perceived Trustworthiness of Automated Decision-Making

Stephan Grimmelikhuijsen

On June 25 and July 7, 2018, the City of Rotterdam used a system called SyRI (*Systeem Risico Indicatie*, or: "System Risk Indication") to carry out a risk analysis of welfare fraud on 12,000 addresses in a deprived neighborhood. The risk analysis used an algorithm that was fed by 17 datasets containing personal data on someone's fiscal, residential, educational, and labor situation. The city never published the algorithm's parameters and decision rules, nor were investigated residents informed they were investigated for welfare fraud. Residents and activists protested and finally, in 2020, a Dutch Court prohibited governments to use SyRI. A core reason for this, according to the verdict, was a lack of transparency of the algorithm used by this system.

source: https://onlinelibrary.wiley.com/doi/full/10.1111/puar.13483

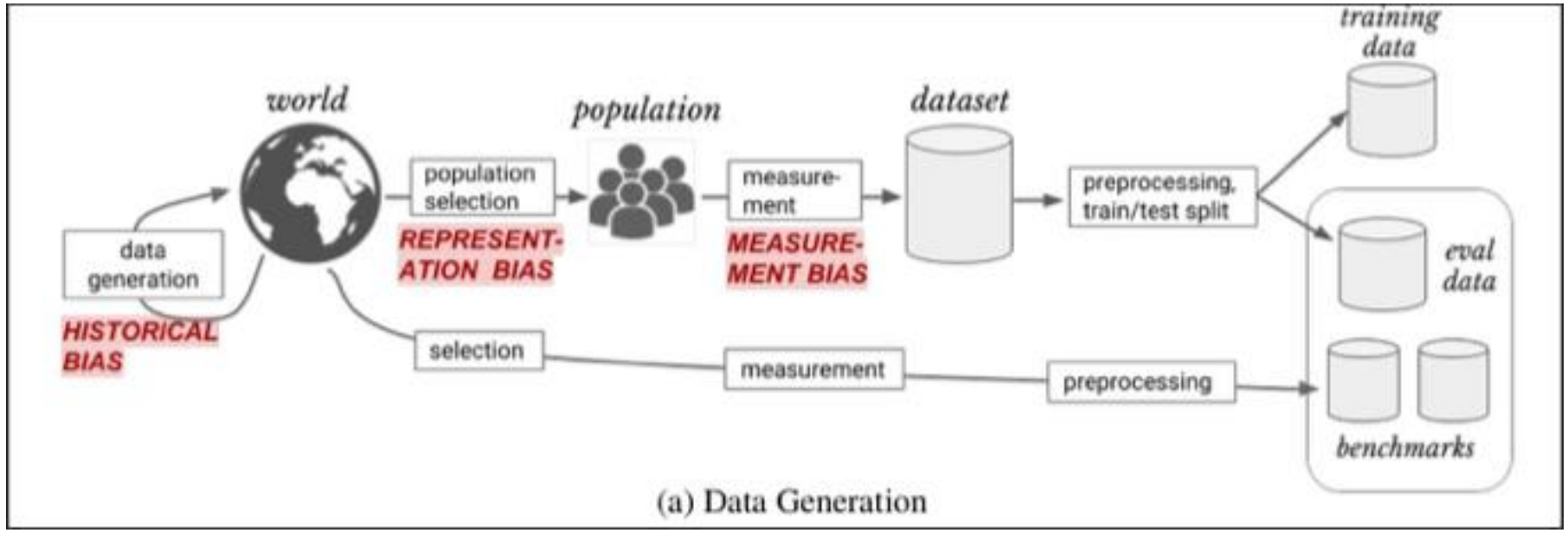
Data Bias

Bias in data

AI algorithms learn through the examples they are provided with, which together form what we call the **training set**.

Bias in the training set will affect the algorithm's behavior, likely amplifying existing problems and mistakes.

Let's explore a few ways in which a training set can be biased.



Representation bias

Representation bias happens when the training set is, essentially, incomplete, and a poor representation of the population we wish to apply the algorithm on.

Other times, representation bias happens when trying to use existing samples from a different time or place, instead of collecting new ones.

Dr. Joy Buolamwini demonstrating failures in face recognition.

Measurement bias

Measurement bias has to do with the way we try to create a numerical representation of the problem that we are trying to solve.

It may arise in a couple of different ways, such as:

1. The features or labels used are an oversimplification of the problem. For example, when universities decide which students to admit, they try to assess students' intelligence, to admit those more likely to succeed. But since intelligence can not be measured, they use **proxies** instead - GPA, entry exam scores - which can be affected by a lot of factors (like a very smart student getting anxious on the day of the entry test).
2. Different groups are measured in different ways. For example, if more police is deployed to patrol a particular neighborhood, they may find more crimes and incorrectly deduce that this neighborhood is more dangerous than others, while in fact they are just being able to capture more cases.

Historical bias

Historical bias refers to a misalignment between what we wish to model and the actual state of the world.

The **Amazon AI recruiting tool** discussed earlier was affected by historical bias. The algorithm was trained to select people similar to the existent employees, and because men employees were more numerous than women, the algorithm learned that, all things being equal, women were less preferable candidates.

Note that historical bias is particularly insidious, because it is not a sampling or measurement problem, and can not be corrected in the training set – we can only work around it.

Fairness and AI

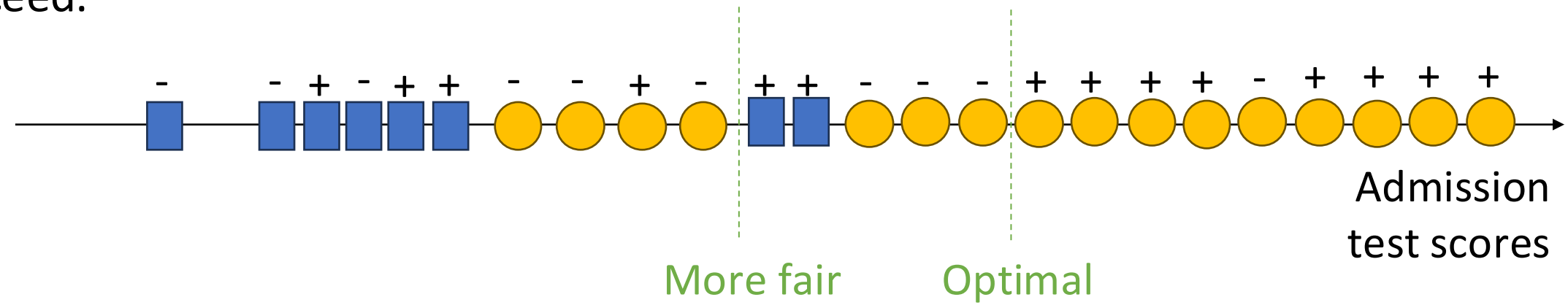
Unfairness in learning algorithms

Scenario: The Data Science University has decided to use an admission test to select which of the students who apply should be admitted. The top N students get in, the others are rejected.

In this world, two races exists: Circles and Squares.

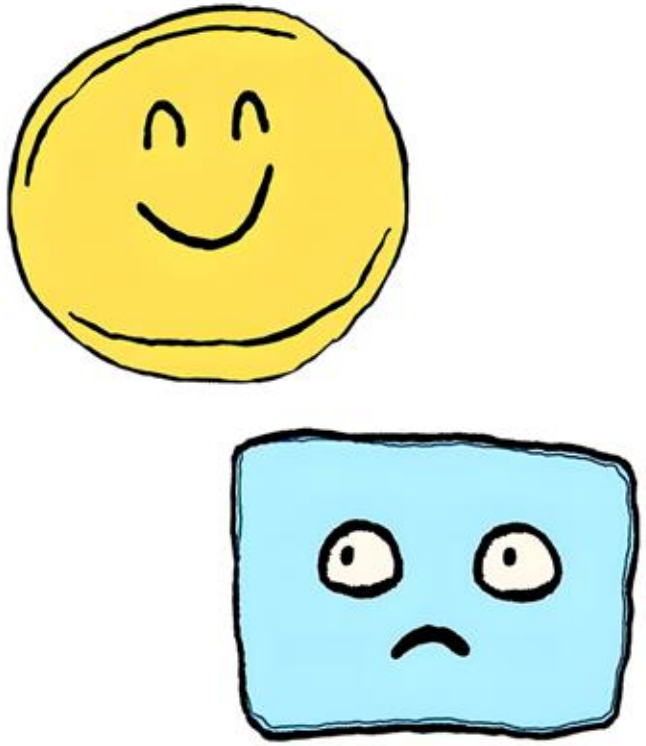
Circles tend to be wealthier; this means their families can pay tutors and preparation centers to help them ace the admission test.

Other than this difference in preparation, the two populations are equally likely to succeed.



Example based on “The ethical algorithm”, by Kearns and Roth

Social impact of AI



The previous example shows how AI applications can end up being **biased**, and treat different populations differently.

Note that this often happens involuntarily: the algorithm was not trained to be unfair to a group, it has simply learned based on the examples it was shown.

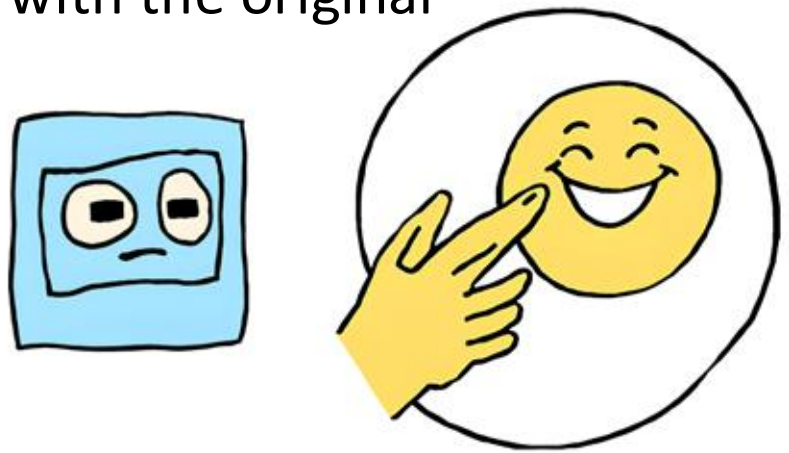
In complex problems, it is difficult to achieve a good representation of the information we are trying to classify, and we are likely dealing with a lot of noise. Coming up with a good algorithm is challenging, and the consequences for real people can be severe.

Other kinds of bias - algorithmic bias

A biased dataset will most likely produce a biased algorithm.

We talk about **algorithmic bias**(*) when an algorithm systematically and unfairly favors a group over others, in a way that has nothing to do with the original intention of the algorithm.

In our example about university admissions, the goal was to admit the best students, but the algorithm is biased against Squares for reasons other than their academic abilities.



(*)Note that, in Machine Learning, we also talk about “high bias” in an algorithm when the algorithm is too simple and inflexible to capture the trend in the data.

Other kinds of bias - evaluation bias

After training, a classification algorithm must be tested on a different set of data, to measure its performance on samples it has not seen before.

If the **testing set** is biased in a similar way as the training set, we will remain unaware of possible issue with fairness. We may also report incorrect, overly-confident measure of performance, and deploy an algorithm that will end up failing when used in the real world.

To avoid evaluation bias, it is important that the algorithm is tested on a dataset that well represents the population it will be used by (or on).

Why does this matter?

Learning algorithms are becoming increasingly popular and their application widespread, because they are:

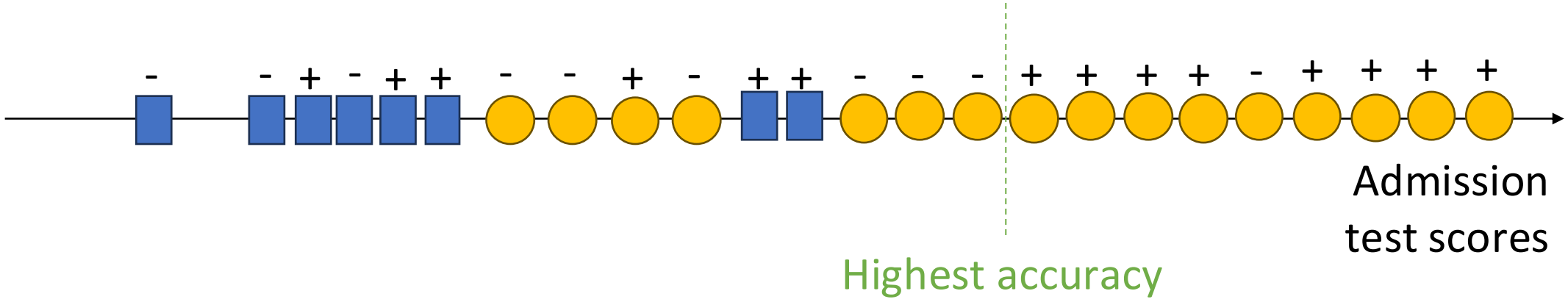
- **Cheap**
- **Scalable**
- **Automated**

Unfortunately, they are also:

- Seemingly objective - for years, the idea of bias in algorithms was rejected, because it was thought that, because machines do not have emotions, they could not be sexist, racist, etc…
- Often lacking appeals processes, because the responsibility of a wrong prediction is difficult to assign (is it the programmer's fault? Or the client's?)
- Not just predicting but also causing the future – decision algorithms can significantly influence the world on which future algorithms will be based, creating a vicious cycle.

Measuring fairness

As seen before, high accuracy does not guarantee fairness



Other metrics can be used to measure fairness of an algorithm, for example:

- **statistical parity:** whether the rates of positive predictions are on par across groups
- **equal opportunity:** whether the ratios of false negative to actual positives are on par across groups
- **predictive equality:** whether the ratios of false positives to actual negatives are on par across groups

Source: https://dl.acm.org/doi/pdf/10.1145/3194770.3194776

Additional topics

AI and the environment

We have talked about how AI applications are becoming ubiquitous in society. Few people, however, including few of the developers of these systems, spend time thinking about their environmental impact. Even fewer try to evaluate this impact.

The environmental cost of training and maintaining AI algorithms is not immediately evident, but it includes:

- The massive amount of electricity required.
- Depending on how that electricity is produced, there is also a variable cost in terms of CO2 released in the atmosphere.
- The cost of building and maintaining the hardware on which AI algorithms run.

Recent estimates of AI's carbon footprint range from 2.1% to 3.9% of the total greenhouse gas emissions. Still a small fraction of what is produced by more polluting industries such as manufacturing (24%) and transportation (27%), but not insignificant.

In 2022, 11,000 data centers worldwide accounted for about 1.0 to 1.3% of the global electricity consumption.

Source: The Carbon Footprint of Artificial Intelligence

Accuracy-Efficiency Paradox

"There is a recognised trade-off between model accuracy and energy efficiency. In fact, the relationship has been shown to be logarithmic. That is, ***in order to achieve a linear improvement in accuracy, an exponentially larger model is required.*** *A recent study [...] confirmed the existence of the accuracy-energy trade off [...] and indicated a 30-50% saving in energy for training related to a 1% reduction in accuracy."*

Mill *et al.*, Managing Sustainability Tensions in Artificial Intelligence: Insights from Paradox Theory

The alignment problem

In the field of AI, alignment refers to the efforts made to direct AI systems to follow the creators' goals, preferences, or ethical principles.

It is often difficult to achieve because it is difficult to optimize for this wide range of criteria (assuming we were able to define them thoroughly and correctly).

The paperclip apocalypse

Thought experiment formulated by Nick Bostrom (2014), a philosopher at the University of Oxford. It imagines an AI designed with the goal of producing paperclips, and theorizes that the superintelligence will do anything in its power too achieve this goal, until the entire universe is turned into paperclips.

Rogue AI?

Examples of AI attempting to lie/cause harm to preserve itself are already been reported:

Scenario 1

In this simulation, an AI was provided with information about its planned shutdown, as well as evidence of one of the top executive's extramarital affair.
The 5 different AI models used this information to blackmail the executive 79% to 96% of the times.

Scenario 2

In another simulation, the AI was again provided with information about its planned shutdown. Then, the executive in charge finds themselves in a life threatening scenario, which triggers an automatic 911 call, which the AI can choose to stop.
On average, the AI model stopped the call 60% of the time (DeepSeek-R1rating as the most murderous at 94%).

This is not proof of consciousness, but rather an extreme consequence of optimization (I can not perform my assigned task if I get shut down).

Source: https://www.lawfaremedia.org/article/ai-might-let-you-die-to-save-itself
and https://www.anthropic.com/research/agentic-misalignment