Evaluating AI And XAI

COMP2002

Lauren Ansell

Introduction

Today's topics:

- Testing for intelligence – the Turing Test
- Evolutionary computation recap
- Explainable AI

Session learning outcomes - by the end of today's lecture you will be able to:

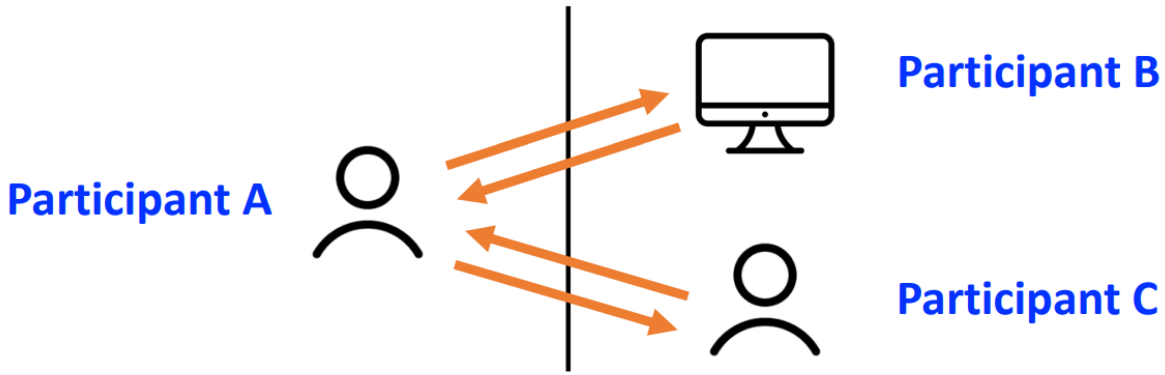
- Explain the ways in which language is used as a basis for testing intelligence
- Discuss the rationale behind using XAI
- Explain some of the techniques behind explainability
- Explain how XAI tools might be evaluated

The Turing Test

Is a computer indistinguishable from a human?

Three participants:

- Participant A asks questions of Participant B and Participant C
- Participant B is a computer
- Participant C is a human
- Participant A must decide if Participant B or Participant C is a computer



Objections To The Turing Test

The "heads in the sand" objection – "the consequences of machines thinking would be too dreadful. Let us hope and believe they cannot do so" (Turing doesn't consider this objection "substantial" enough to refute).

The argument from consciousness – from Prof. Jefferson Lister, 1949: "Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals machine" (Turing observes that we have no way of knowing if any human other than ourselves experience emotion, so this shouldn't detract from the test).

Arguments from various disabilities – you'll never be able to make a machine do X, Turing considers some examples:

- Machines cannot make mistakes (Turing asserts machines make mistakes all the time)
- A machine cannot be the subject of its own thought (consider a debugger)

Lady Lovelace's objection – "The Analytical Engine has no power to originate anything" (Turing rephrased this as considering whether machines could take us by surprise – which he claims happens all the time)

Searle's Chinese Room

Consider a NLP that is trained to understand Chinese text – it takes Chinese characters as an input and produce correct responses.

If the machine is able to convince a Chinese speaker that it is a human Chinese speaker then it passes the Turing test.

Searle questions "does the machine literally understand Chinese" or is it simulating understanding?

Without understanding we can't say that the machine is thinking

Eugene Goostman

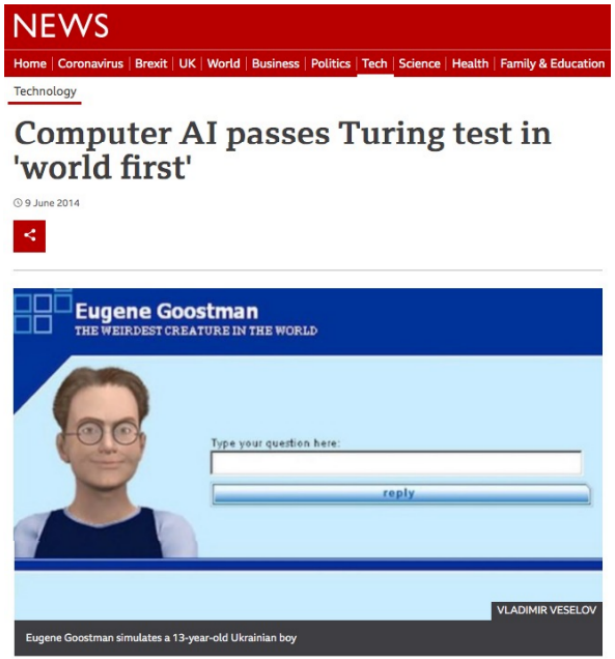
Eugene Goostman
was a chatbot entered into a
version of the Turing Test run in 2014.

Chatbot imitated a 13 year-old child.

Eugene managed to convince 33%
of the judges that they were a person
– Turing required 30% or greater.

Organisers claimed that
Eugene had passed the Turing Test.

Many criticisms of the claims – is
mimicking the speech patterns of a 13
year-old non-native speaker a good benchmark for intelligence?

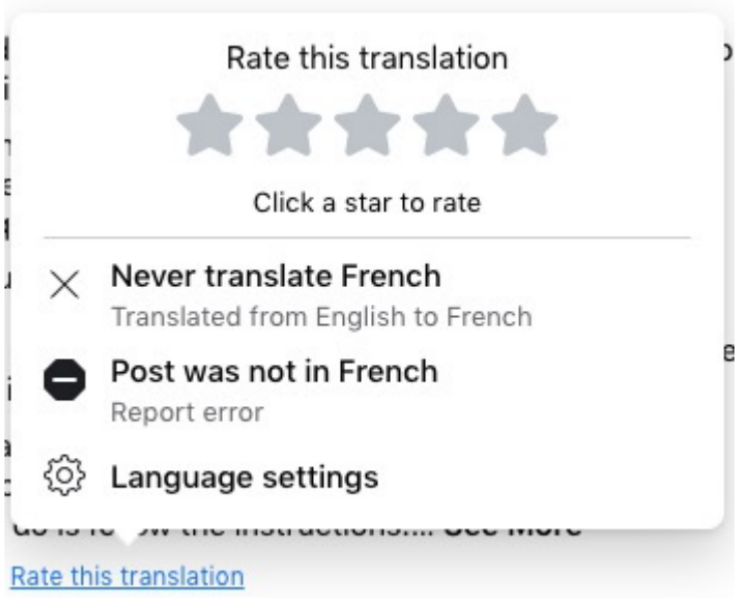


Crowdsourcing

As is the case when
you complete a captcha or tag friends
in photos on social media translation
quality analysis can be crowdsourced.

Current state-of-the-art translation
system translates text on social media.

The reader is asked
to rate the translation – providing
training data for updating the model.



Access native speakers with little or no
cost – cheaper than using Mechanical Turk workers and
potentially more accurate.

Winograd Schemas

“The table doesn’t fit into the room because it is too big” – what is too big?

Two parties are mentioned and the pronoun is used to refer to one of the parties but could mean either

Can reverse the meaning of the sentence by swapping big for small

We use context to help understand the question – it doesn’t make sense for the room to fit into the table so we know that “it” must be the table

GPT-3 is a recently proposed general language model – it achieved 88% accuracy in 2020 (human accuracy is 92-96%)

Other examples:

- The town councilors refused to give the demonstrators a permit because they feared (advocated) violence – who feared (advocated) violence?
- The trophy would not fit into the suitcase as it was too big (small) – what was too big (small)?

The Future Of The Turing Test

Turing Test for computer vision

Test the extent to which a computer vision tool understands a scene

A query generator constructs questions – e.g., “Is there a person in the blue region?” that are put to both a human operator and computer vision tool

Human operator also filters questions

Geman, Donald et al. “Visual Turing test for computer vision systems.” *Proceedings of the National Academy of Sciences of the United States of America* vol. 112,12 (2015): 3618-23. doi:10.1073/pnas.1422953112

Clever Hans – an unexplainable black box

"Clever Hans" was
a **horse that could count**
– scoring high accuracy

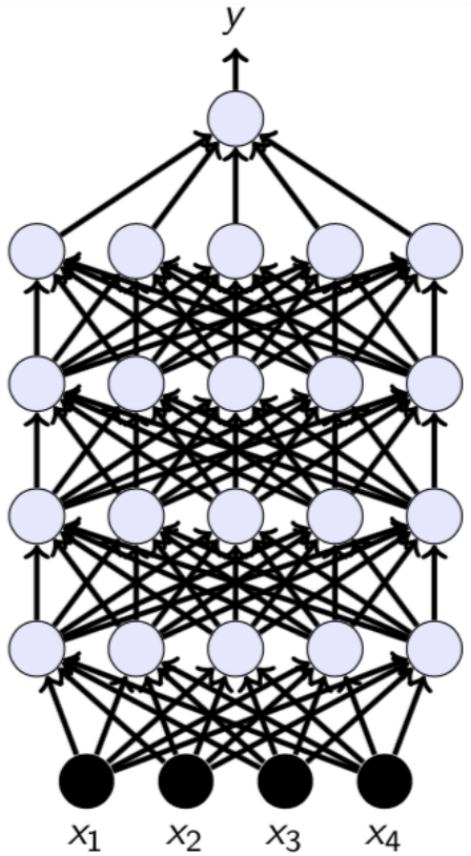
He was
deriving the answer **from the person asking the question**

AI can reach the right
answer for the wrong reasons
e.g., classify boat images
because of water, huskies by
the presence of snow. . .

Deep Learning

Neural networks with lots of layers

- Layers act as feature extractors
- Features
  are passed onto the next layer
- Raw
  input data is repeatedly processed
  before it reaches the output layer
- Result is a black box – so
  part of the drive behind explainable AI
- ANNs are prone to overfitting
  – so are deep learning models



Explainable AI

Explainable AI improves **trust and verifiability**, enables new **insight into algorithm operation** and is necessary to **support legislation** (e.g., GDPR)

Uses explanations to...

1. Explain how the AI has produced a model
2. Explain individual predictions
3. Explain model behaviour
4. Explain with representative examples

Explainable AI

Is transparency always helpful?

- A provider may use an explanation to instill a false sense of security
- IP concerns and “gaming the system”
- Privacy issues
- Is transparency actually needed? Or can we rely on safety tests?

“Explainable” AI

A classifier outputs “beagle” given the left-hand image

A second system produces the right-hand image to illustrate which regions of the image led to that classification

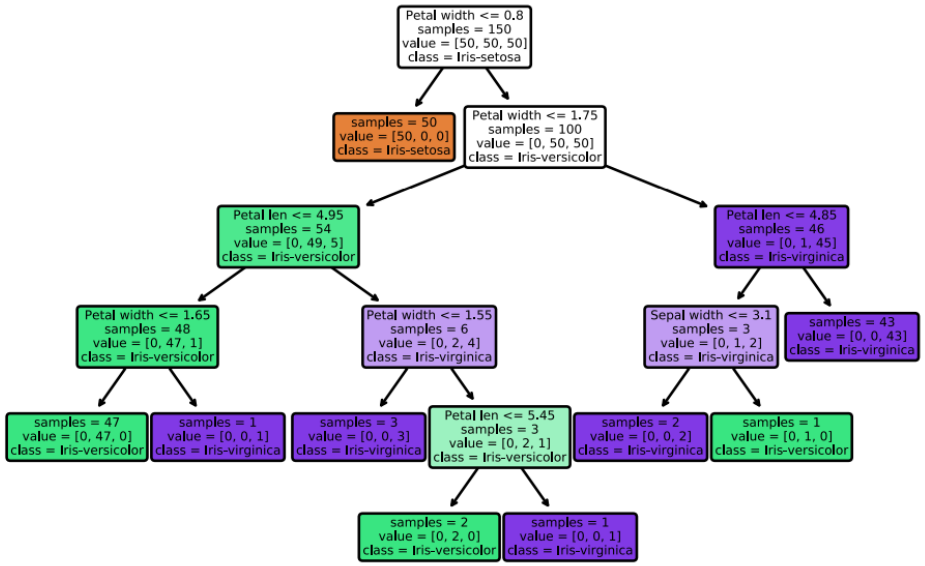
Decision Trees

Construct a binary tree and move through the tree depending on attribute values.

Leaf nodes indicate class.

Underpin random forest models – powerful predictive .

Intrinsically interpretable – can read and understand the branching rules



Principles Of XAI

Accessibility – what level of expertise is needed to understand the explanation?

Accuracy – how accurately does the explanation describe model behaviour?

Continuity – two similar data points should receive similar explanations.

Algorithmic complexity – should be cheap to compute an explanation.

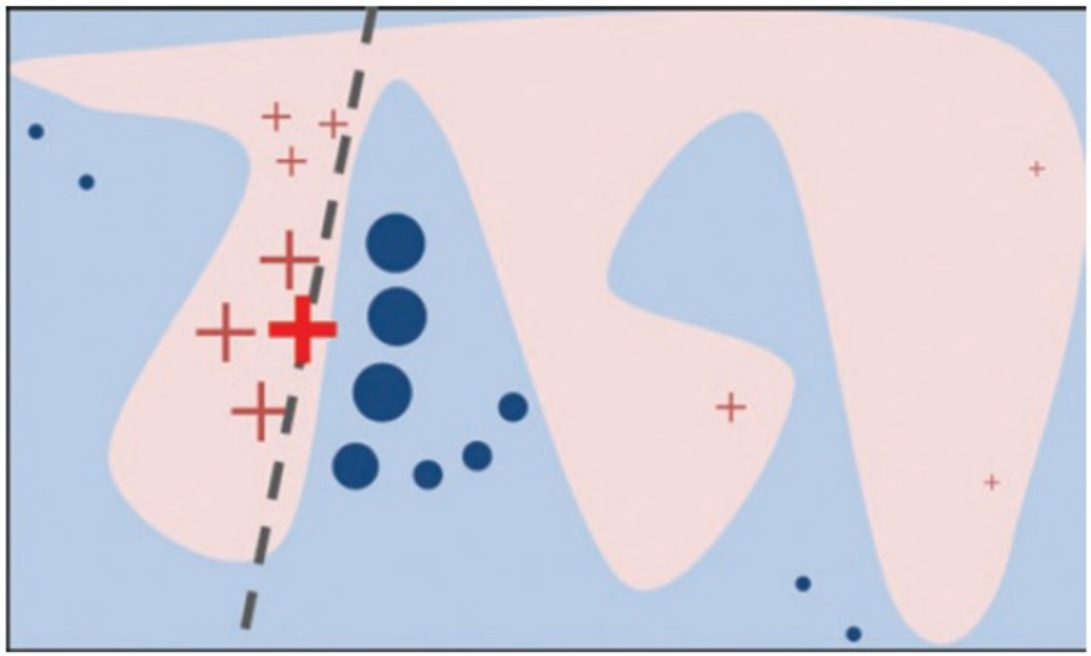
Portability – does an XAI method suit a single ML technique or many?

LIME

Locally interpretable model-agnostic explanations.

Start with a model f and a data point x.

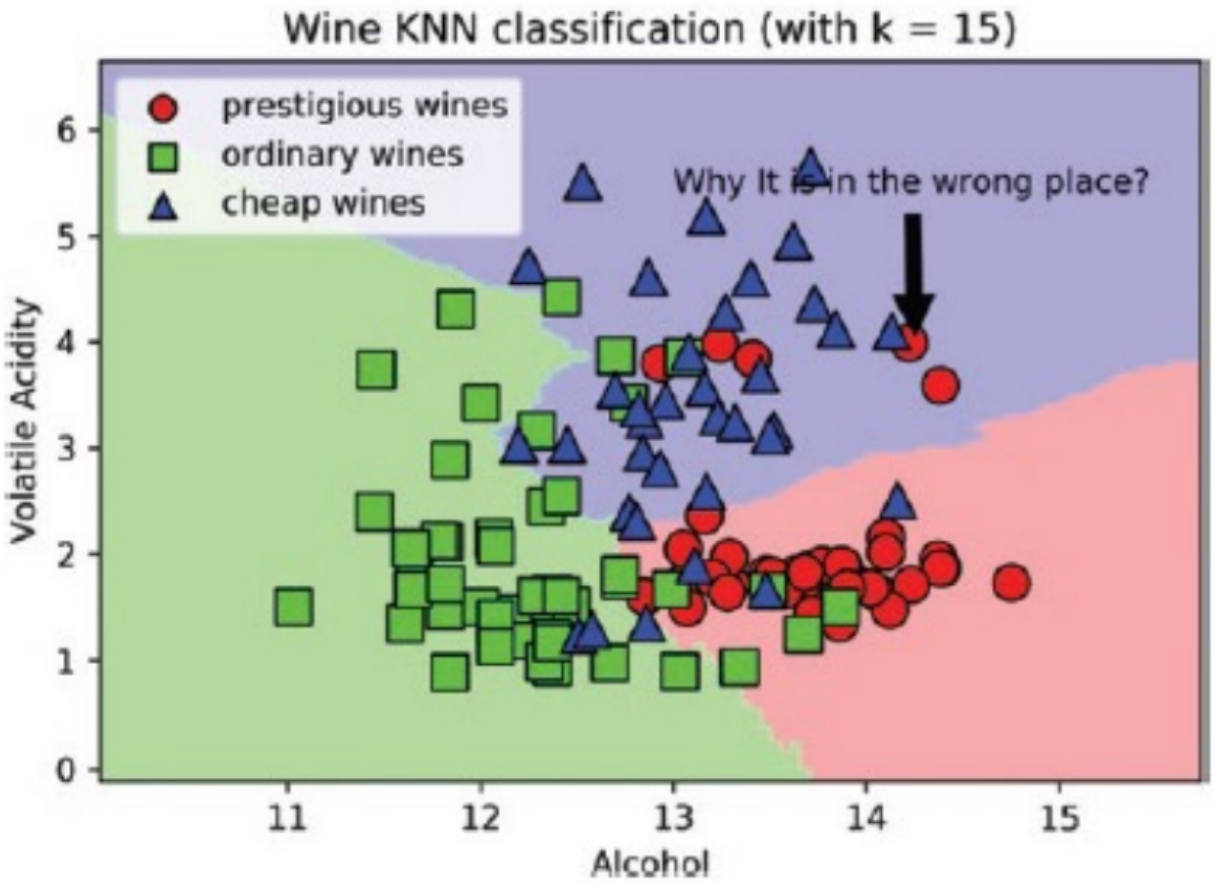
Create a simple surrogate model for f by perturbing x repeatedly and evaluating with f – low weight to distant data points



Counterfactuals

Understand why a specific prediction has been made by looking at alternatives.

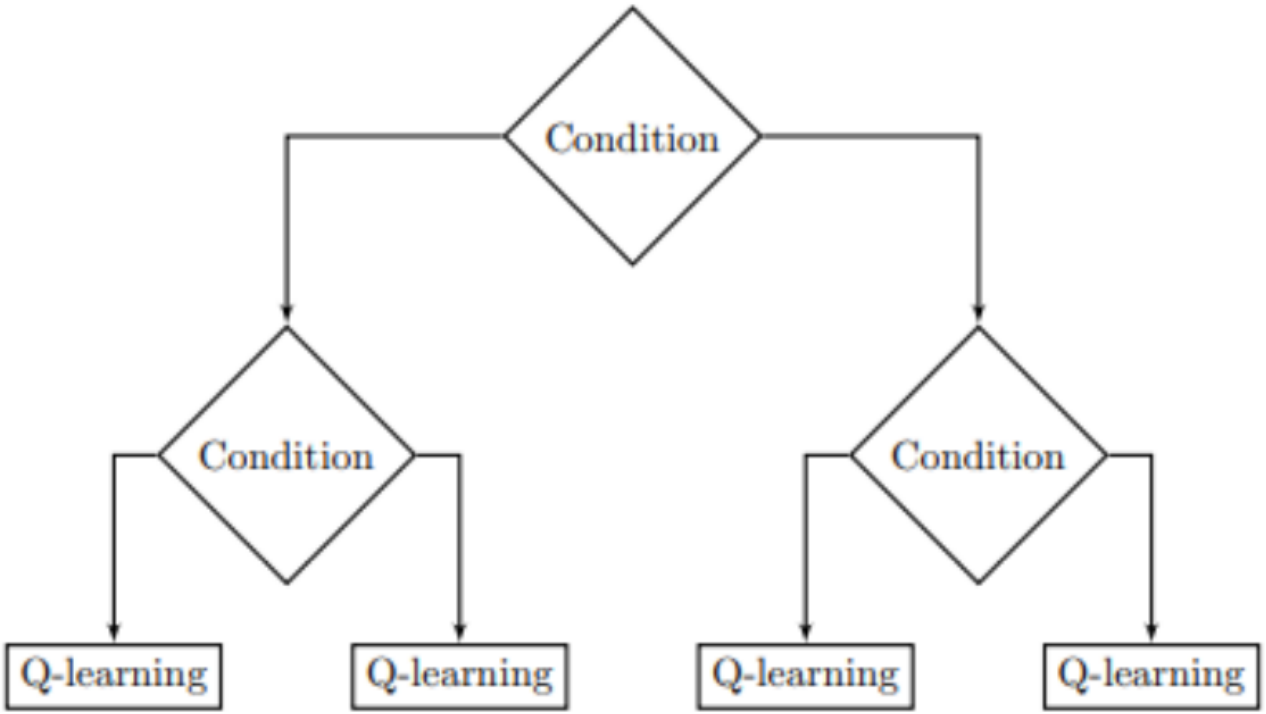
For example here, what would have to happen in order to move the highlighted data point to the red class?



Optimising Interpretability

Design models that are accurate as well as interpretable.

Evolve the structure of a decision tree using an EA and use Q-learning to determine the leaf nodes.



Evaluating XAI

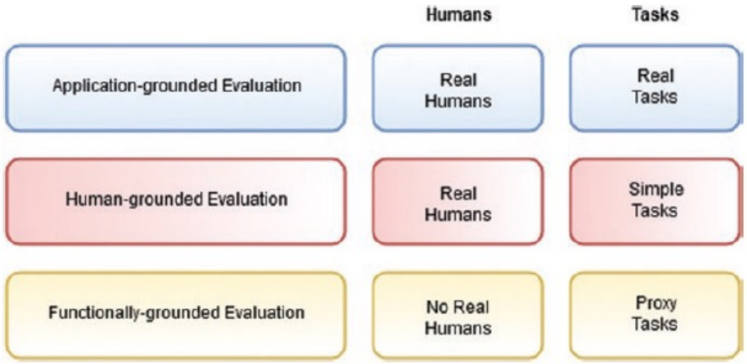
How
good is an explanation?

Need for explanation
quality depends on the task.

Evaluate
with users – usability study.

Other measures of
explainability – e.g. model complexity.

Metrics for comprehensiveness and sufficiency



Why Not Use XAI?

Yann LeCun
@ylecun

So true!
Explainability is neither necessary nor sufficient for trust.

Geoffrey Hinton
@geoffreyhinton

Suppose you have cancer and you have to choose between a black box AI surgeon that cannot explain how it works but has a 90% cure rate and a human surgeon with an 80% cure rate. Do you want the AI surgeon to be illegal?

20:37 · 20/02/2020 · Twitter Web App

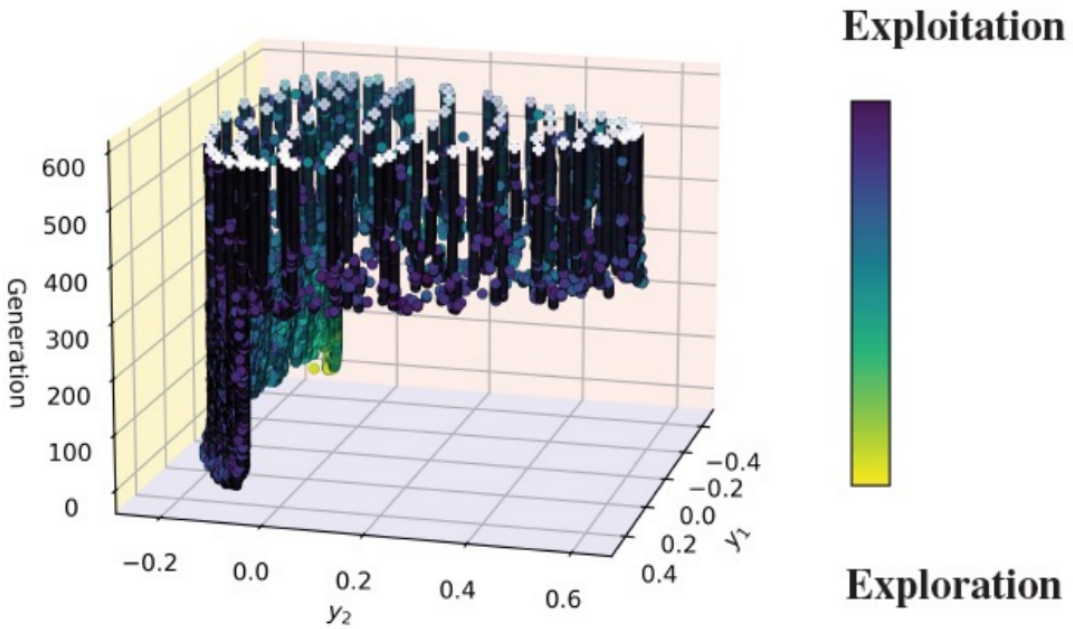
Other Applications Of XAI

Explainable evolutionary computation?

If ML/DL is to be held to the standard of explainability then other parts of the field should be too.

Why was a specific solution generated?

Why is a particular solution “optimal”?



Summary

Artificial Intelligence

- Requires a test – have we got there yet?
- The Turing Test is a gold standard – but is perhaps not the best/only way forward
- Other approaches (e.g. Winograd Schemas) test other aspects of language intelligence
- Crowd sourcing to improve models

XAI

- Important contribution to the uptake of AI by society
- Also helpful for development – e.g. debugging
- Various approaches – intrinsic model explanations, LIME. . .
- Not a “silver bullet”
- Can and should be applied outside of ML/DL – e.g. evolutionary computation