Kool Aid and Snake Oil: the challenges of ‘explaining’ AI

Professor Chris Baber

Chair of Pervasive & Ubiquitous Computing

School of Computer Science

What I am, and am not, going to talk about…

I will talk about:

- What we mean by an ‘explanation’ (and why should we expect AI systems to be able to do this when we, humans, often find it difficult to do?)
- Where we should place the responsibility and accountability for AI systems’ decisions and recommendations?

I am not going to talk about:

- How to code specific AI algorithms
- Detailed examples of analysis using AI systems

“Imagine what it would be like if we gave the pharmaceutical industry the leeway that we currently grant to tech companies.”

John Naughton, *The Observer,* 06/08/20

FATML

- Fairness, Accountability, Transparency in Machine Learning
  - Responsibility: who to approach when the ML fails
  - Explainability: nontechnical details for all stakeholders
  - Accuracy: identify sources of error and uncertainty
  - Auditability: allow third party scrutiny and checking
  - Fairness: not biased against different demographics

https:///www,faml.org.resoureces/ principles-for-accountable-algorithms

Example Cases (almost all statistical algorithms and ML rather than AI)

- Recommending which movie to watch
- Advising on a course of medical treatment
- Managing a domestic energy system
- Driving a vehicle
- Managing a transport network
- Selecting applicants for jobs
- Deciding on loan applications
- Deciding on likelihood of repeat offending
- Deciding on likely crime ‘hot spots’
- Predicting A-level exam grades

Comparing AI with ML

- Public accounts often conflate AI with ML
  - Machine Learning seeks structure in data
  - AI seeks to learn not only the structure but also the actions to perform on states which generate those data
  - AI seeks autonomous action
    - But AI can use techniques from ML, and ML techniques can be unsupervised and can learn – which makes distinction a bit messy
    - Most of my colleagues in Computer Science would probably agree that only a fool would try to find clear and obvious differences between ‘AI’ and ‘ML’

Mat Velloso @matvelloso · Nov 22

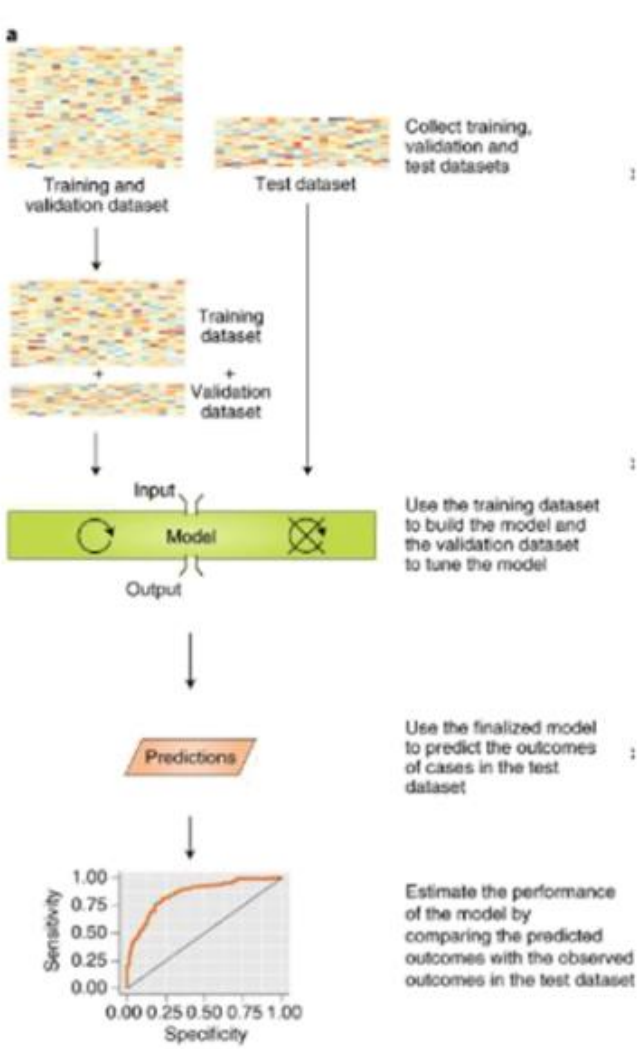
Difference between machine learning and **AI**:

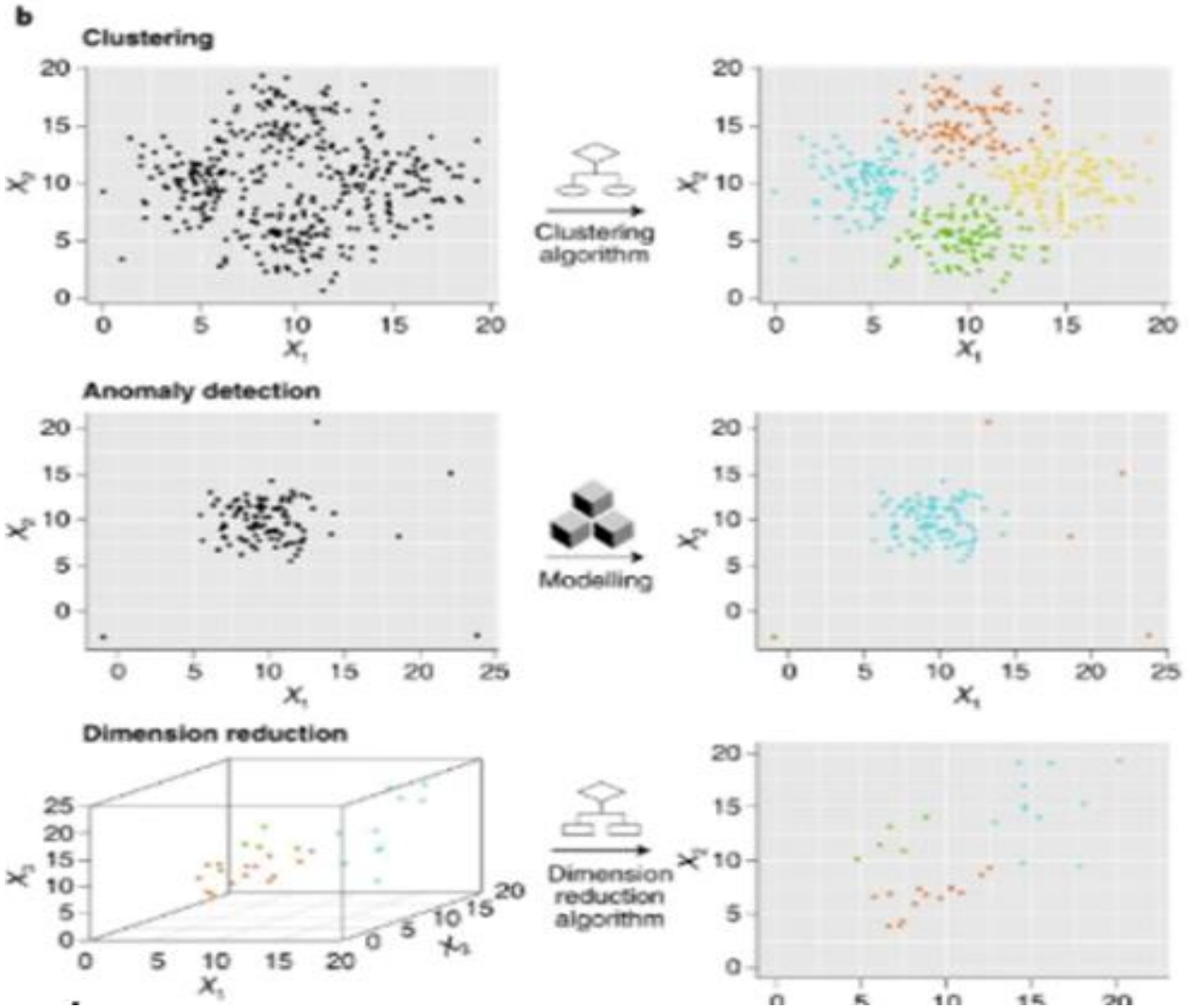
If it is written in Python, it's probably machine learning

If it is written in **PowerPoint**, it's probably **AI**

166 6.6K 19K

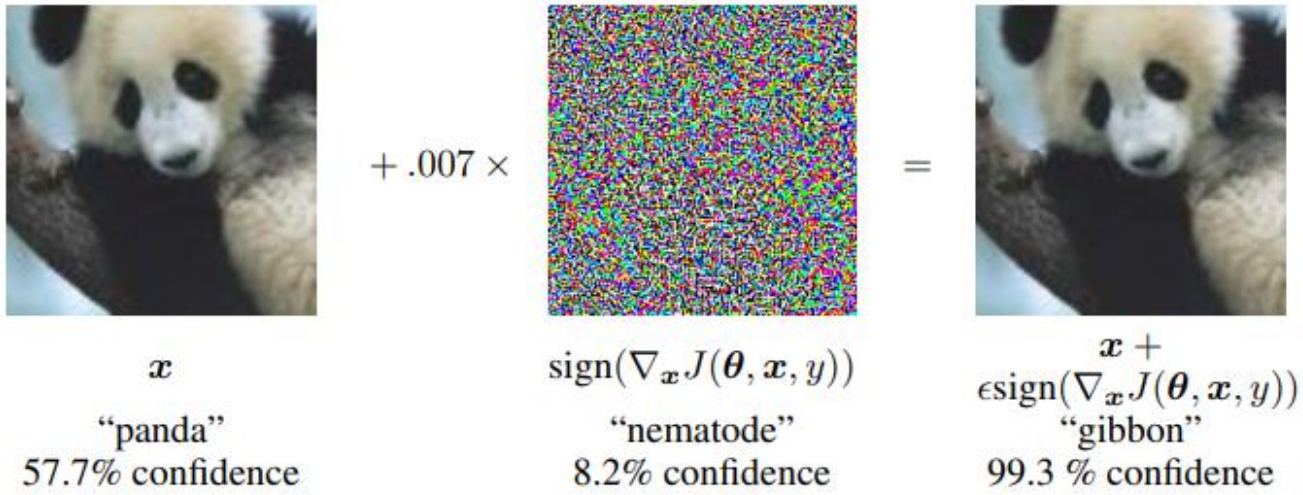
Show this thread





Kun-Hsing Yu, Andrew L. Beam & Isaac S. Kohane, 2010, Artificial Intelligence in Healthcare, Nature Biomedical Engineering

https://www.nature.com/articles/s41551-018-0305-z

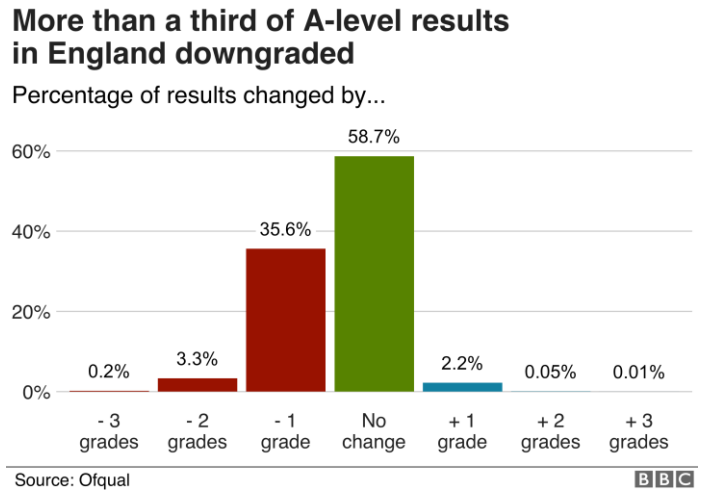


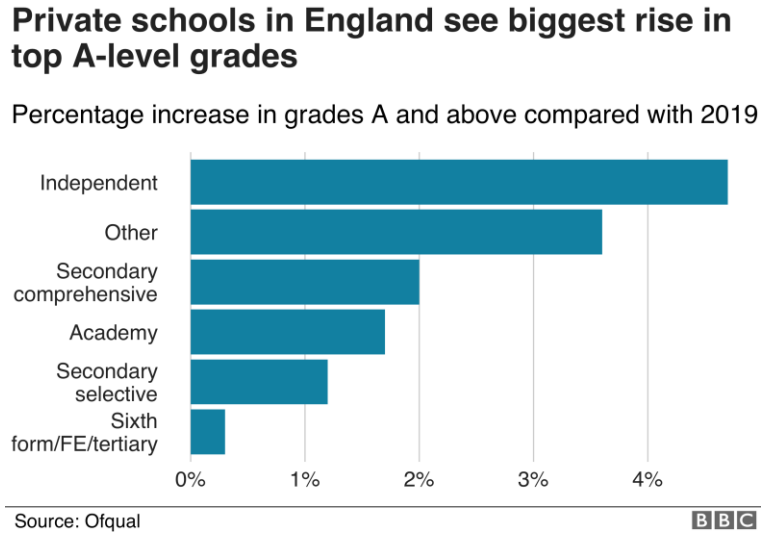
- Adding a small amount of noise can distort image recognition
- This effect is being explored in Adversarial Neural Networks (to address possible risks from spoofing)
- But…not knowing the difference between a 'panda' and a 'gibbon' means that *this* algorithm would be susceptible to attack

Szegedy et al., 2014, Going deeper with convolutions. Technical report, arXiv preprint arXiv:1409.4842.

A “mutant algorithm”?

- Maintain National distribution of A-level scores to prevent ‘grade inflation’, i.e., to be "broadly similar to previous years“





- Various algorithms were trialled against 2019 data
  - That is, known results were the ‘predictor’ and ‘outcome’ variables for the tests – but, even then, the best performance was only about 68% (the worst was 27%)

[source: Paul Taylor, 10th September 2019, *London Review of Books,* p.10]

“The algorithm itself is described in Section 8 of Ofqual’s technical report (p83). It includes the following steps:

1. Look at historic grades in the subject at the school
2. Understand how prior attainment maps to final results across England
3. Predict the achievement of previous students based on this mapping
4. Predict the achievement of current students in the same way
5. Work out the proportion of students that can be matched to their prior attainment
6. Create a target set of grades
7. Assign rough grades to students based on their rank
8. Assign marks to students based on their rough grade
9. Work out national grade boundaries and final grades

This algorithm is used if a school has more than fifteen children doing an A level or GCSE in a given subject.

If a school has five or fewer children doing an A level or GCSE in a given subject, steps 1-7 get skipped, and the rough grades that get used to allocate marks to students are based on the grades their teachers originally predicted for them.

If a school has between five and fifteen children doing an A level or GCSE in a given subject, then a combination of the teacher predictions and the algorithmic predictions get used.

As teachers overall tend to over-estimate grades, this means overall scores will tend to be higher for small classes.”

Jeni Tennison 16/08/2020 How does Ofqual’s grading algorithm work? https://rpubs.com/JeniT/ofqual-algorithm

What is happening here…

- There is an assumption that teachers can't be trusted and will exaggerate the predicted grades of their pupils.
- There is a concern that 'grade inflation' needs to be controlled.
- Normalising to a (statistical) Population ignores individual scores (so the grades were not about *pupils* but about datapoints in a school in an education authority in a national (English) examination system)
- Normalising to a (political) Population re-emphasises societal imbalance (so pupils from Private schools benefitted from the historical data or smaller class sizes)

Simpson's paradox: how structuring data can lead to unexpected outcomes

(Yule-Simpson effect)

	Infected	Recovered	%
Condition A	160	60	37.5
Condition B	200	65	28.3

Partitioning the data can produced contradictory outcomes

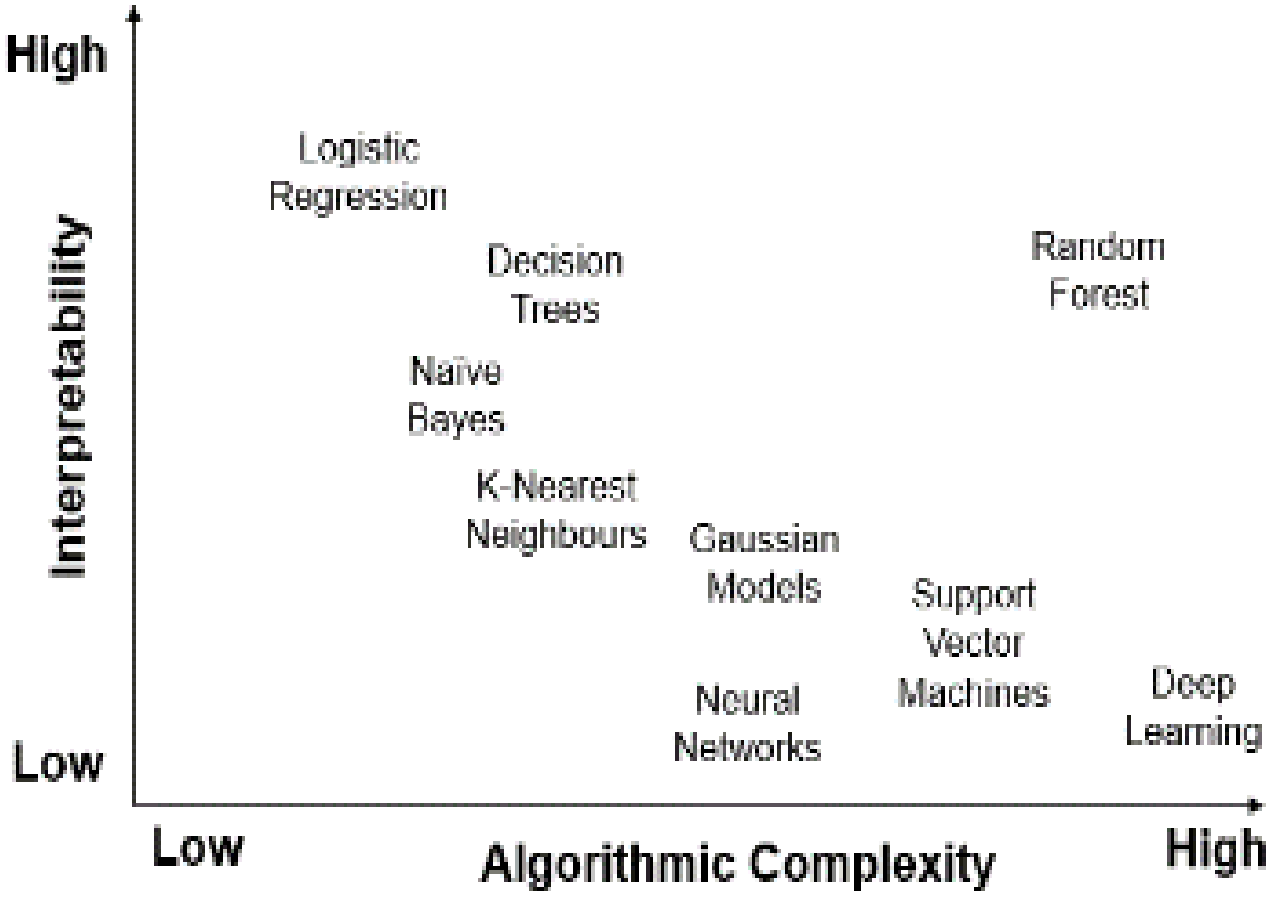
	Infected	Recovered	%
Condition A	160	60	37.5
Condition B	200	65	28.3

	Aged 18-30			Aged 50+		
	Infected	Recovered	%	Infected	Recovered	%
Condition A	100	20	20	60	40	67
Condition B	180	50	23.8	20	15	75

What needs ‘explaining’…

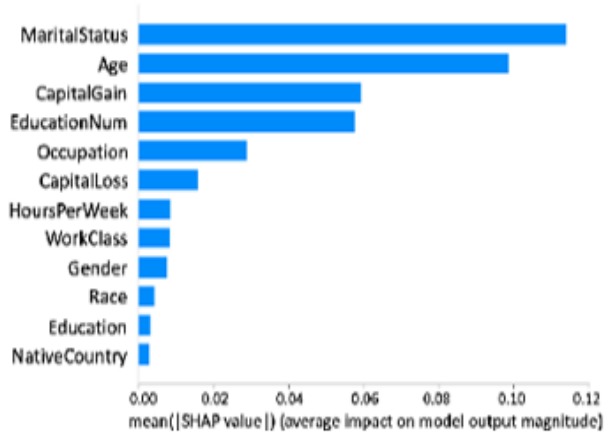
- The Primary Question that the ‘algorithm’ is addressing
  - in the ‘A-level’ example, a very complex (possibly impossible problem) was replaced by a complicated (but solvable) problem
- The structure (and sources) of the ‘data’
  - categorical (grade), subjective opinion, ranking, historical (grade plus percentage of cohort)…
- The structure of the statistical model(s) and their assumptions
  - i.e., that data can be combined and normalised to match ‘trends’ from prior performance
- The implementation of the statistical model, in the algorithm
  - statisticians who offered to help were asked to sign a Non-Disclosure Agreement
- The performance (and limits of the algorithm)
  - which ought to have undergone checking for ‘sensitivity’ (in terms of variation in output to changes in input) and ‘sanity’ (in terms of unintended consequences)
- The implications of applying *this* algorithm to *these* data
- The most important ‘features’ of the data (in terms of output)
- The alternative (‘what-if’) outputs that might arise if other features were used instead
- …

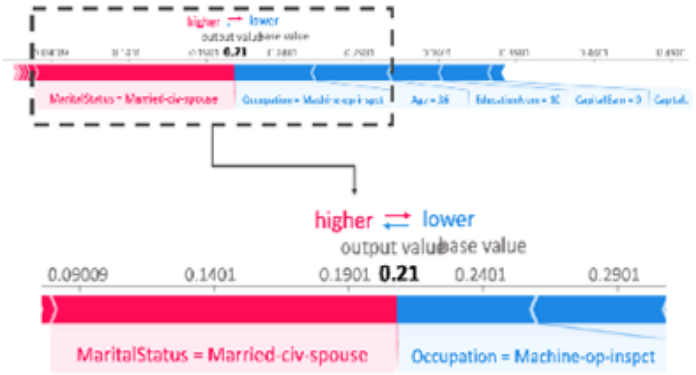
How can we look inside a ‘black box’?



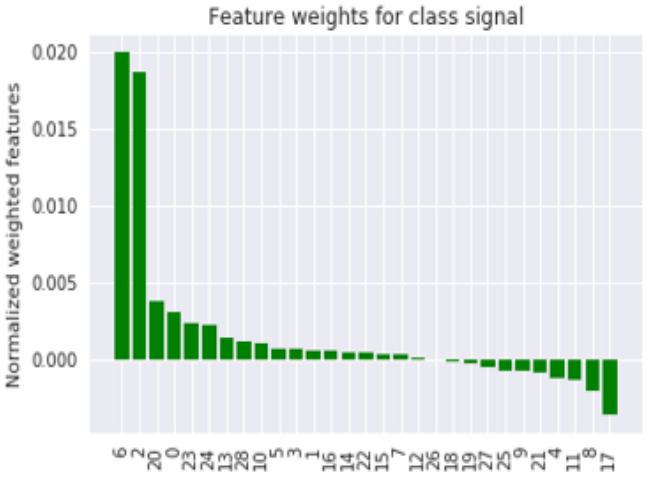
- Creating ‘local’ models that explain specific outcomes
  - Local Interpretable Model-agnostic Explanations (LIME)
  - SHapley Additive ExPlanations (SHAP)
- BUT…we don’t know whether the Local explanation (or features used) will generalise to any other situations

SHAP



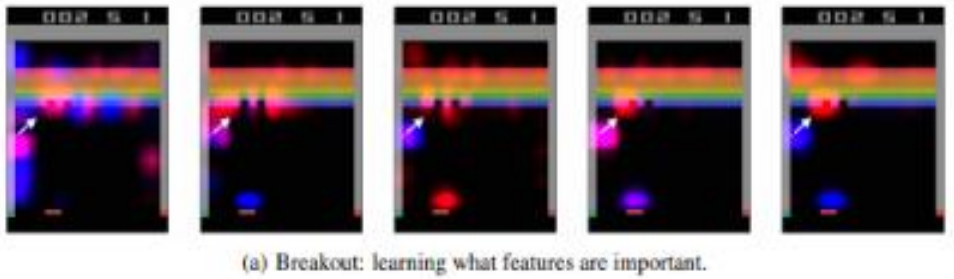


LIME

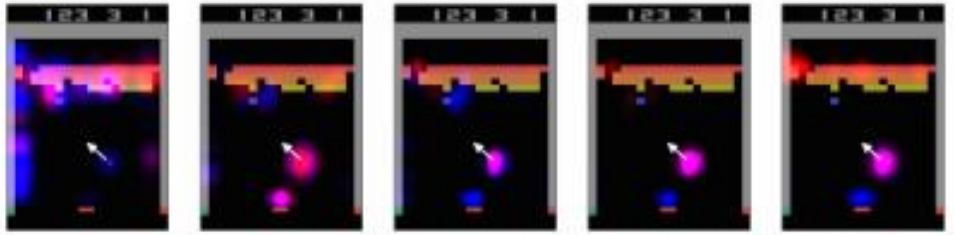


Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., Vaughan, J.W. (2020) Interpreting interpretability: understanding data scientists’ use of interpretability tools for machine learning, *CHI 2020,* New York: ACM, paper 92.

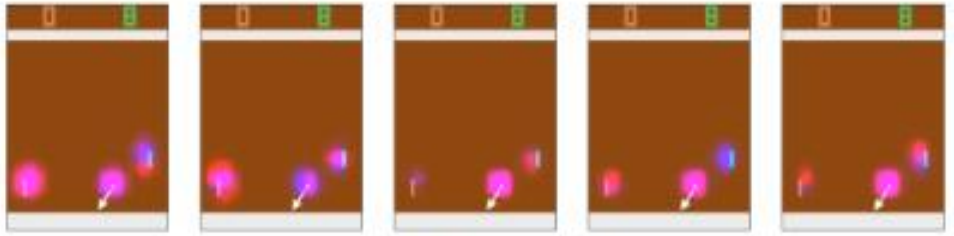
- Agents trained with Asynchronous Advantage Actor-Critic algorithm
- Creating post-hoc Saliency Maps of ‘actions’ (for Actor and Critic models) in Atari video-games, e.g., Pong, Breakout, Space Invaders…
- This allows humans to *infer* which features that AI systems were most likely to be using and the strategy that the AI system might be using
- In Space Invaders, the Agent seemed to apply an aiming strategy but not clear how precise this was. So, from Saliency Maps it looked as if Agent “…had learned a sophisticated aiming strategy the actor…would ‘track’ a target…[and] to monitor the area above the ship.”
- BUT we don’t know whether our inferences of strategy or features are correct



(a) Breakout: learning what features are important.



(b) Breakout: learning a tunneling strategy.

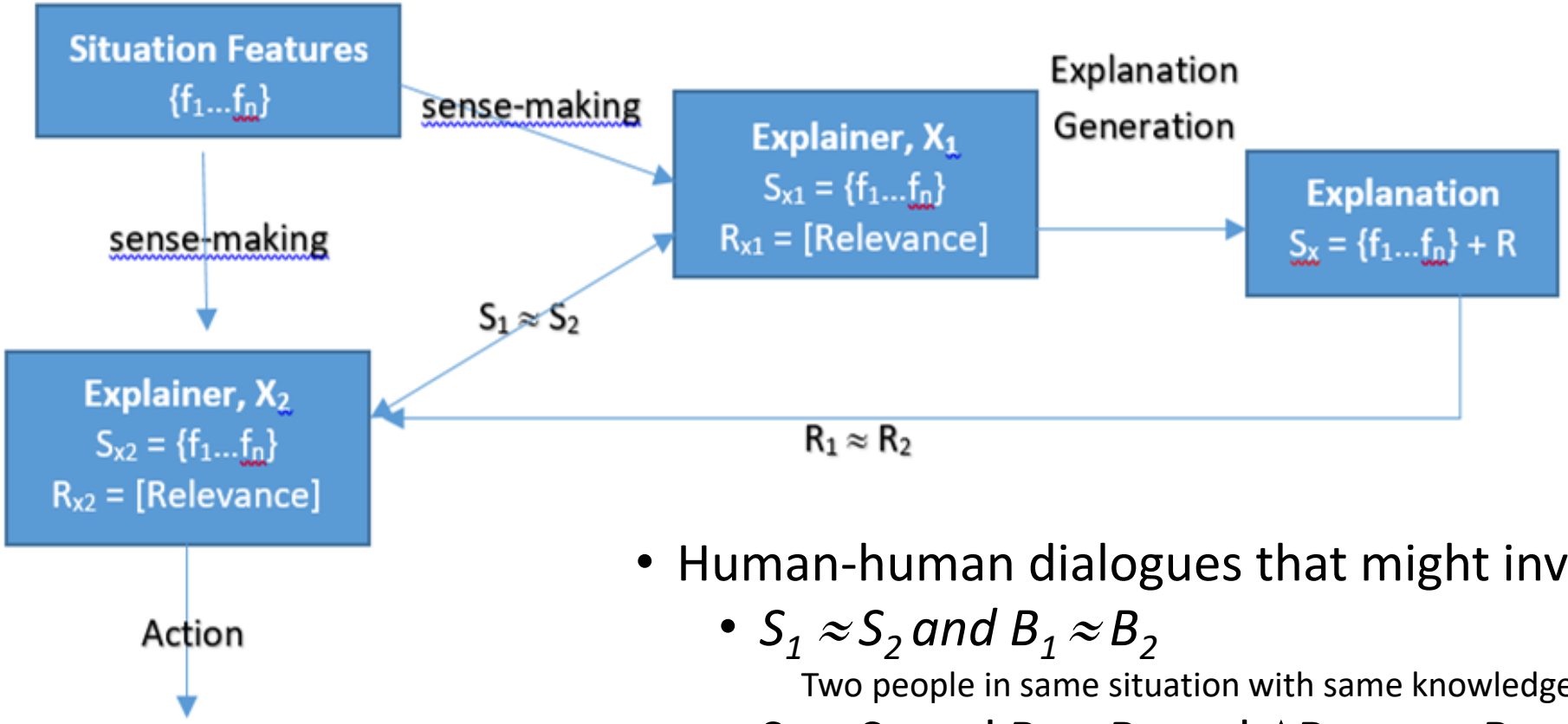


(c) Pong: learning a kill shot.

(d) SpaceInvaders: learning what features are important and how to aim.

Greydanus, S., Koul, A., Dodge, J. and Fern, A. (2018) Visualizing and understanding Atari agents, arXiv:1711.00138v5

A Framework for Explanation



- Human-human dialogues that might involve 'explanation'
  - $S_1 \approx S_2$ *and* $B_1 \approx B_2$
    Two people in same situation with same knowledge, training and experience
  - $S_1 \approx S_2$ and $R_1 \neq R_2$ and $\Delta R_2 \approx r_1 \cap R_1$
    Teacher and pupil
  - $S_1 \neq S_2$ and $R_1 \neq R_2$ and $\Delta R_2 \approx r_1 \cap R_1$ and $A_2 = \Delta s_2$
    Doctor and patient

From our Model of Explanation:

1. Explanations should include salient **causes**

   Explanation should be related to beliefs in the relationship between features of a situation and causes that can directly affect the event being explained (probability), or can explain the majority of the event (explanatory power), and are plausible (construct validity) and, if the cause was instigated by a person, deliberative.

2. Include relevant **features**

   The Explanation should relate to key features of the situation and to the goals of the *explainer* and *explainee*.

3. Frame the Explanation to suit the **audience**

   Fit the explanation to suit the *explainee's* understanding of the topic and what they wish to gain from the explanation (their mental model and goals).

4. Explanations should be **interactive**

   Involve the *explainee* in the explanation. Seek **alignment** (between *explainer* and *explainee*) **in features** used in the explanation

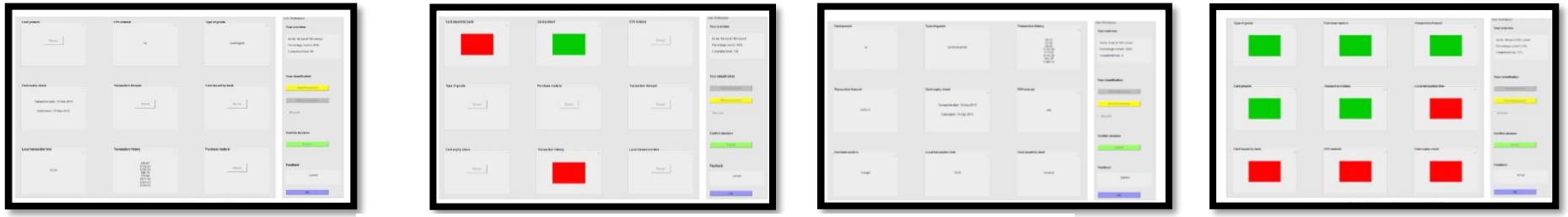
5. Explanations should be (where necessary) **actionable**

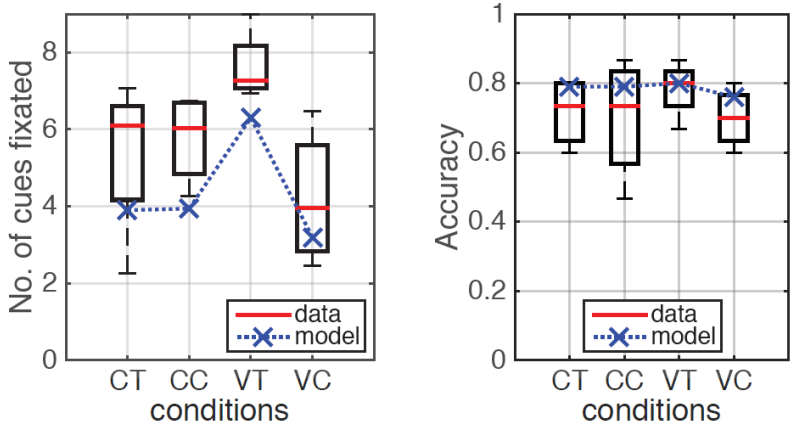
   The *explainee* should be given information that can be used to perform and/or improve future actions and behaviours.

6. Clarify the **definition of relevance** used in the explanation

   Define clusters (i.e., statistical model), belief (i.e., causal model) and policy (i.e., implications for action)

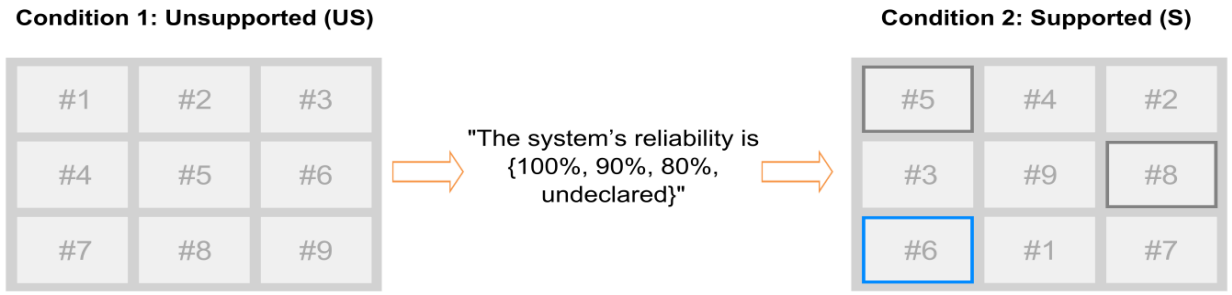
Human ‘policy’ could mirror Reinforcement Learning policy

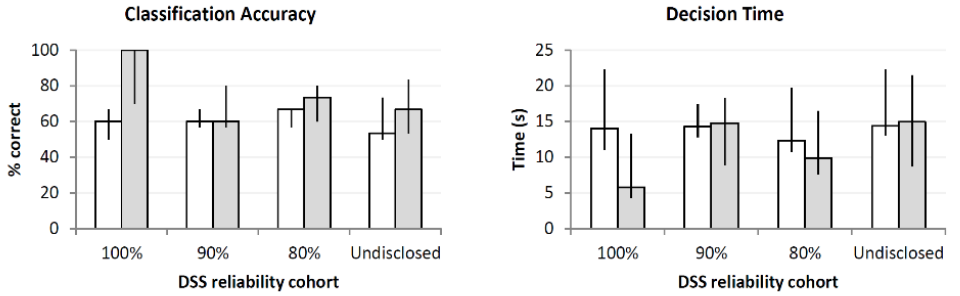


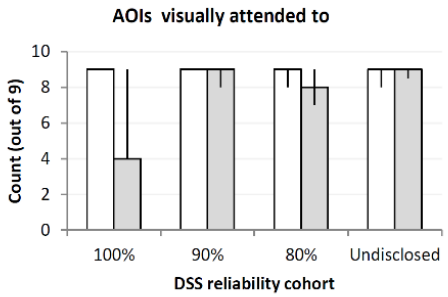


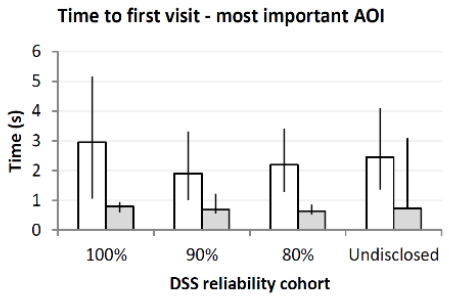
Chen, X. et al., 2017, A cognitive model of how people make decisions through interaction with visual displays, In *CHI’17: Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, New York: ACM

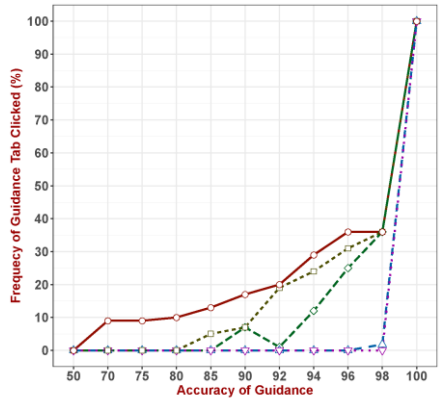
If the recommender is not perfect?

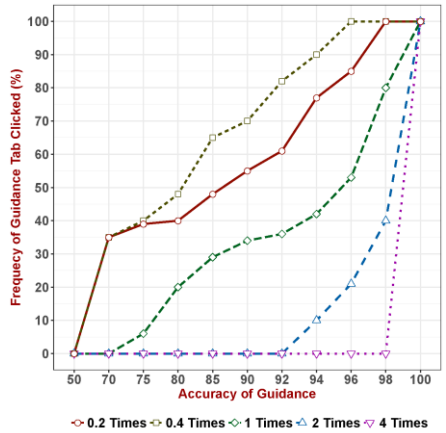


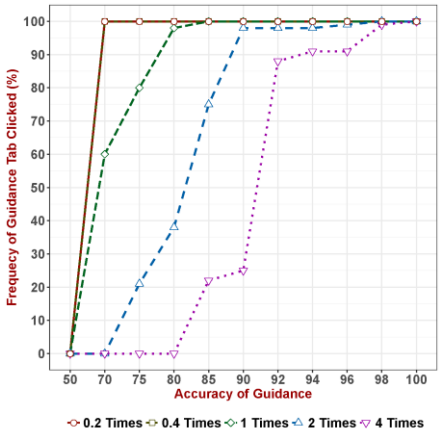












Starke, S.D. and Baber, C. (2020) The effect of known decision support reliability on outcome quality and visual information foraging in joint decision making, *Applied Ergonomics, 86, .*

Acharya, A., Howes, A., Baber, C. and Marshall, T., 2018, Automation reliability and decision strategy: a sequential decision making model for automation interaction, *Proceedings of the 62nd Human Factors and Ergonomics Society Annual Meeting,* 144-148.

In *Explaining Decisions with AI, part 1*, the Turing Institute offers four guiding principles:

- Be Transparent
- Be Accountable
- Consider Context
- Reflect on Impacts (Fairness, Safety and Performance)

https://ico.org.uk/media/about-the-ico/consultations/2616434/explaining-ai-decisions-part-1.pdf

Kool Aid and Snake Oil: the challenges of ‘explaining’ AI

Professor Chris Baber

Chair of Pervasive & Ubiquitous Computing

School of Computer Science

From Farinola Augustine to Everyone: 10:04 AM
Please, raise your hands or type in your questions here. Thanks
From Lis Shrimpton to Everyone: 10:05 AM
Thinking of 'ironies of automation' could it actually be safer if we don't expect AI to be perfect and instead encourage it is a tool with other sources also to be drawn on

From Farinola Augustine to Everyone: 10:06 AM
Thank you
More questions..keep it coming

I ASKED WHETHER PEOPLE WOULD PREFER TO RECEIVE A DIAGNOSIS FROM A HUMAN GP OR A COMPUTER (using Babylon Health as an example and the question of triaging patients)

From Anum Pirkani to Everyone: 10:09 AM
Would prefer a human diagnosis...!!!
From Lis Shrimpton to Everyone: 10:10 AM
would prefer computer!!
From Stephanie Thompson to Everyone: 10:10 AM
I think if using a computer it needs human input as well
From H.Yumoto to Everyone: 10:10 AM
If I am sure the computer is accurate enough, I would prefer computer.
From muhammad sagir yusuf to Everyone: 10:10 AM
I think computer is better
From Iain Shaw to Everyone: 10:10 AM
Will the computer smile at me and seem to care?
From Aslam Ghumra to Everyone: 10:10 AM
Actually neither, for me it would be better for human diagnosis, after a computer diagnosis
From Karn Vohra to Everyone: 10:10 AM
I trust computers and AI but would prefer integration with human knowledge too
From D.J.Carter@bham.ac.uk to Everyone: 10:11 AM
maybe one day, but I don't think we can rely on computers completely yet!

From D.J.Carter@bham.ac.uk to Everyone: 10:22 AM
so how long do you think it might take for AI to be reliable enough to use routinely in medicine/diagnostics?

From Natalia Hartono to Everyone: 10:23 AM
Thank you for wonderful presentation and explanation, Prof. Baber.