Explainable AI

Anne Coull
UNSW
anne.objectiveinsight@gmail.com

October 2021

Anne Coull

Anne is a courageous, committed and strategic leader with a track record of leading high performing teams to streamline business processes and turn around failing programs.

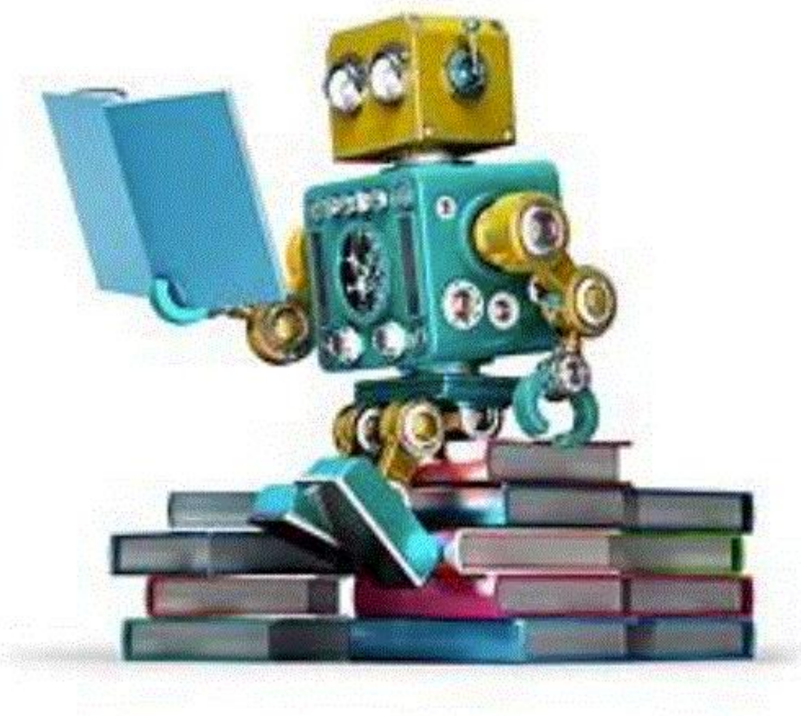
She applies her deep knowledge and experience in Program Management, Cyber Security, SDLC, ITSM, Lean Operational Excellence, Agile and Organisational Change to deliver Business, Cultural, and Technology Transformations at scale.

Dedicated to continuous learning and research Anne is co-founder of Women in Cyber Security (Wicys) Australia, a member of the NSW School of Engineering & Information Technology (SEIT) External Advisory Committee, an active contributor to the development of technical research papers and conference presenter for the International Academy, Research and Industry Association (IARIA). Topics include: Four testing types core to informed ICT governance for cyber-resilient systems; How much cyber security is enough; Most Essential of Eight; and Explainable AI.

Projects and Areas of Interest

1. Artificial Intelligence and Machine Learning
2. Explainable AI
3. Cyber Security, Resilience, and Anti-Fragility
4. Cyber Anti-Fragility through Explainable AI

Artificial Intelligence & Machine Learning



Machine Learning

Field of study that gives computers the ability to learn without being programmed.

(Arthur Samuel 1959)

Well-posed learning problem

A computer program is said to ‘learn’ from experience **E**
with respect to some task **T** and some performance measure **M**,
if its performance on **T**, as measured by **P**, improves with experience **E**.

(Tom Mitchel 1998)

Narrow AI Machines are good at learning a narrow task Consumer internet Google Self driving car	**General AI** Good at multiple things Science Fiction Take over the world

([19])

"AI can transform every major industry"

Input	AI	Output
Email	SPAM filtering	SPAM
Audio	Speech recognition	Text transcript
English	Machine translation	Chinese
Ad, user info	Online advertising	Click?
Owner's face	Facial recognition	Unlock the iphone
Applicants financial info	Loan application risk forecasting	Will you repay the loan
Scene in front of car, lidar reading	Self driving car	Positions of other cars
Image of a product (eg. A phone)	Visual inspection	Identify manufacturing defects

([18])

Page 4

Why is AI not used more broadly?

Availability of Training Data

Speech recognition
50,000 hours of data
and transcripts

Generalisability across Data Sources

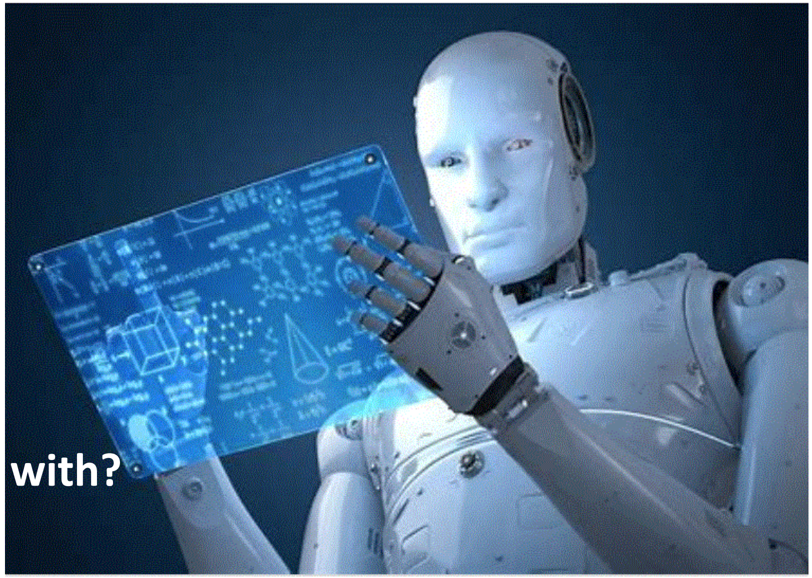
Effective diagnosis
with one scanner
but not transferable
to another

Acceptance by those affected

Change
Management
&
Explainability

([19])

User Acceptance



Change Management

+

Explainability

↓

Increased understanding and acceptance for stakeholders

What is in it for me (WIIFM)?

- What problem is this AI solving

- Will this help, hinder, or confuse me

How will this affect my job and those I work with?

Trustworthiness: the model acts as intended

Causality: of relationships between variable

Transferability: model boundaries and alternative uses

Informativeness: the problem the model is solving

Confidence: stability and reliability of the model

Fairness: explainability facilitates an ethical analysis of the model

Accessibility and Interactivity: user involvement in ML model development

Privacy awareness: how data is captured and used by the ML model

Cybersecurity: identifies vulnerabilities in the ML model

([1][21])

Classic vs Explainable AI: Classic

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

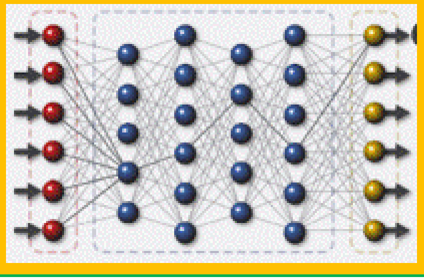
What is this?

Training Data

Learning Process

$$L = d(tv, pv) = \sum_{i=1}^{T} d(tv, pv)$$

Learned Function



Output

This is a cat
P=.92

(based on [5][6])

Classic vs Explainable AI: Explainable

- I understand why
- I understand why not
- I know when you'll succeed
- I know when you'll fail
- I know when to trust you
- I know why you erred

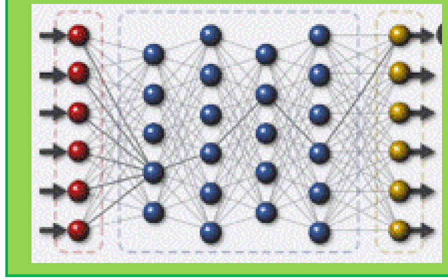
What is this?

Training Data

Learning Process

$L = d(tv, pv) = \sum_{i=1}^{T} d(tv, pv)$

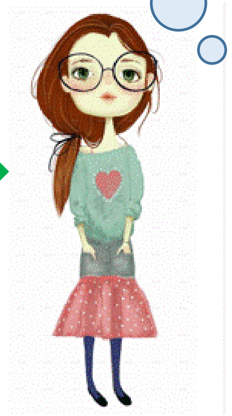
Learned Function



Output

This is a cat.
It has fur, whiskers, claws, eyes like this, &ears like this

Explanation Interface



(based on [5][6])

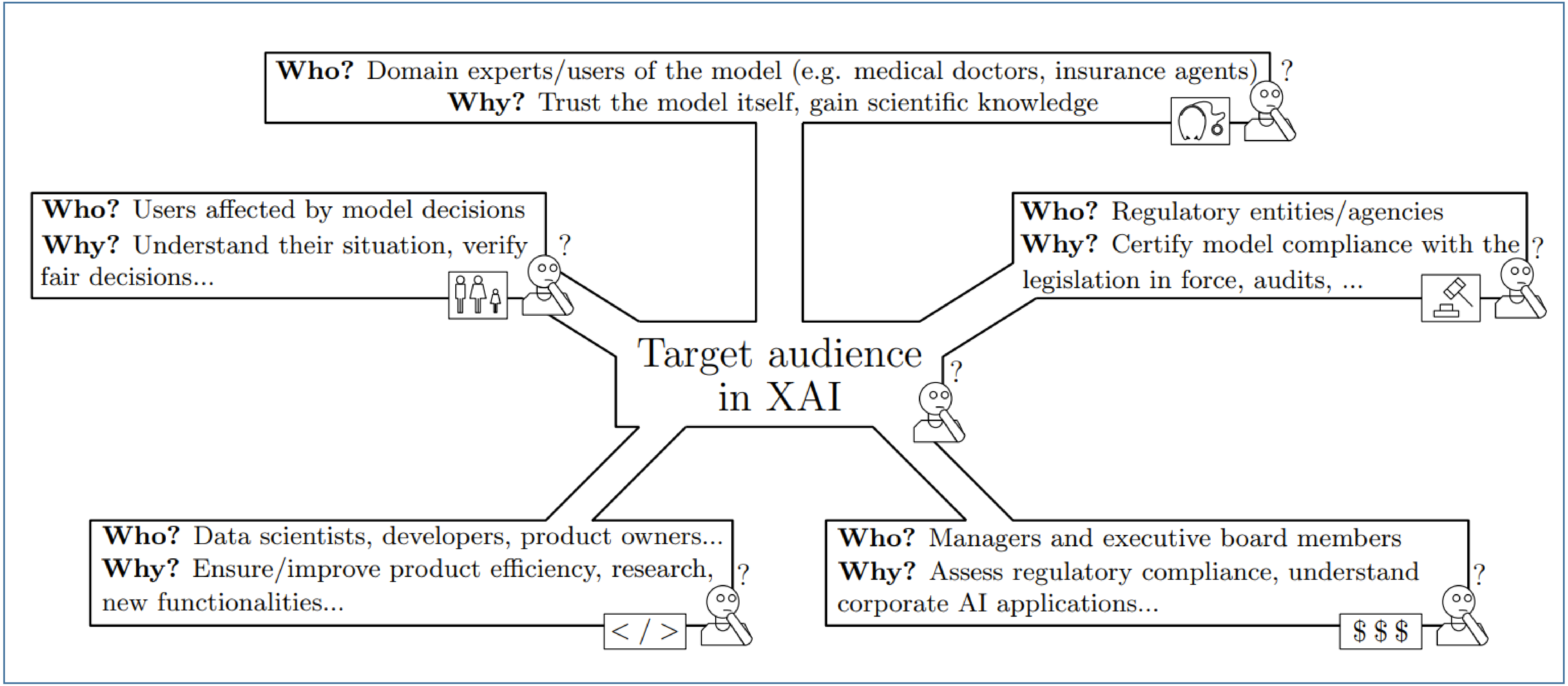
Explainable Artificial Intelligence (XAI)

1. Explainability facilitates impartiality in decision-making by making bias, generated from the training set, transparent so this can be corrected
2. Explainability facilitates robustness by highlighting conflicting outcomes that could destabilise the predictions and make them unreliable.
3. Explainability can verify that only meaningful variables drive the output, providing assurance that the model's reasoning is solid and reliable.

***"XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners."*.**

(based on [1][3])

Explainable to Whom? AI Stakeholders



Different stakeholders have differing requirements from ML model explainability.

([1])

Self-Explaining: Transparent Interpretable Models

Logical / linear regression
Decision tree
K-Nearest Neighbours (KNN)
Rule-based Learners
General Additive Models
Bayesian Models

if…
then…
else…

Bayesian Rule Lists (BRL)

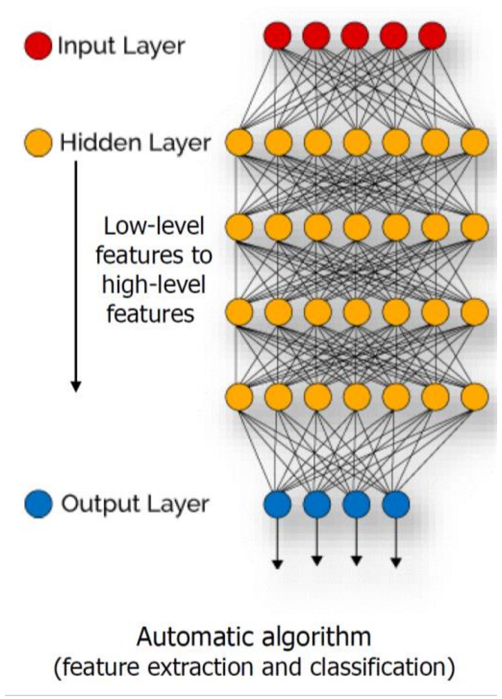
if male **and** adult **then** *survival probability* 21% (19%–23%)
else if 3rd class **then** *survival probability* 44% (38%–51%)
else if 1st class **then** *survival probability* 96% (92%–99%)
else *survival probability* 88% (82%–94%)

Decision list for the Titanic survivors. In parentheses is the 95% credible interval for the survival probability.

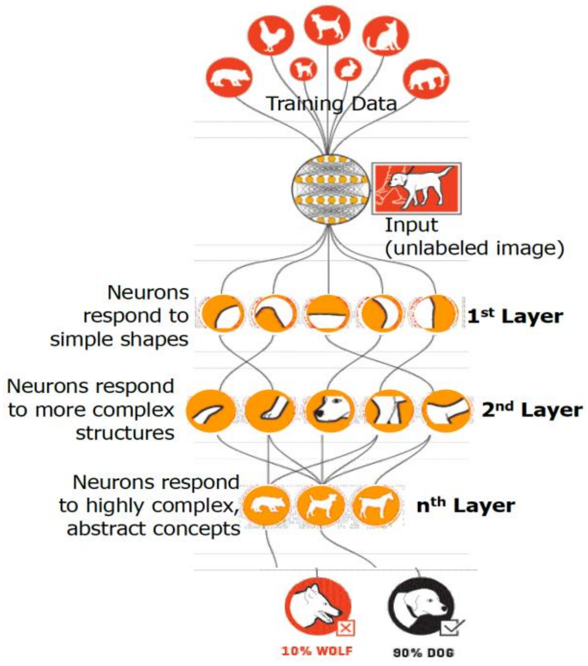
([1][15][16])

Deep Explanations

Deep Learning Neural Network



Modified deep learning techniques



Decision

This is a dog. It is not a wolf.

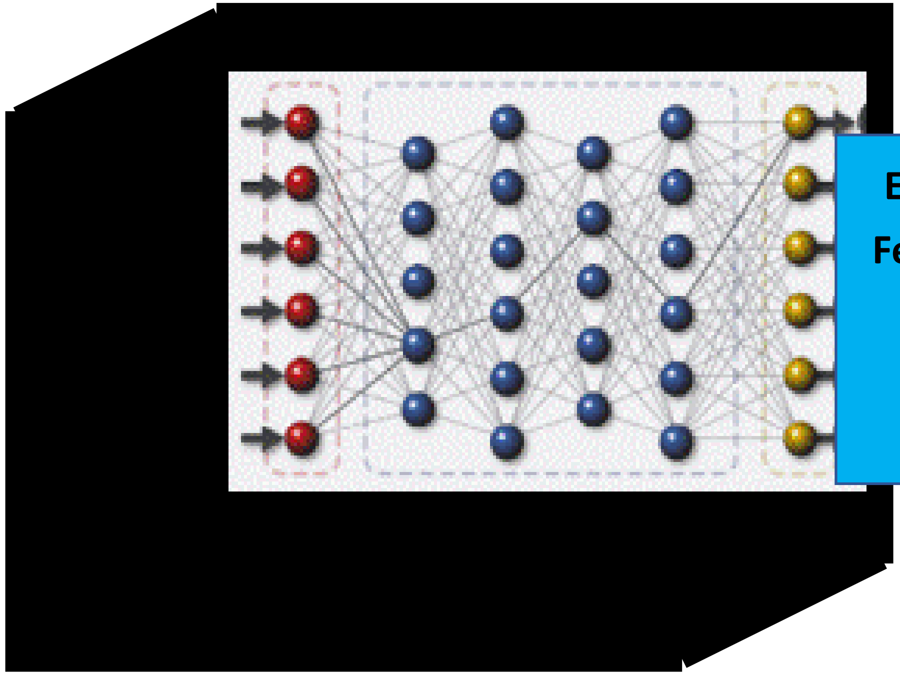
Explanation

Because it has short fur, narrow tail, round paws, a rounded nose, dark eyes and a round rump.

([5][6][23])

Opaque Model Induction

Opaque Model Induction



Explanation by simplification

Feature relevance explanation

Local Explanations

Visual explanation, and

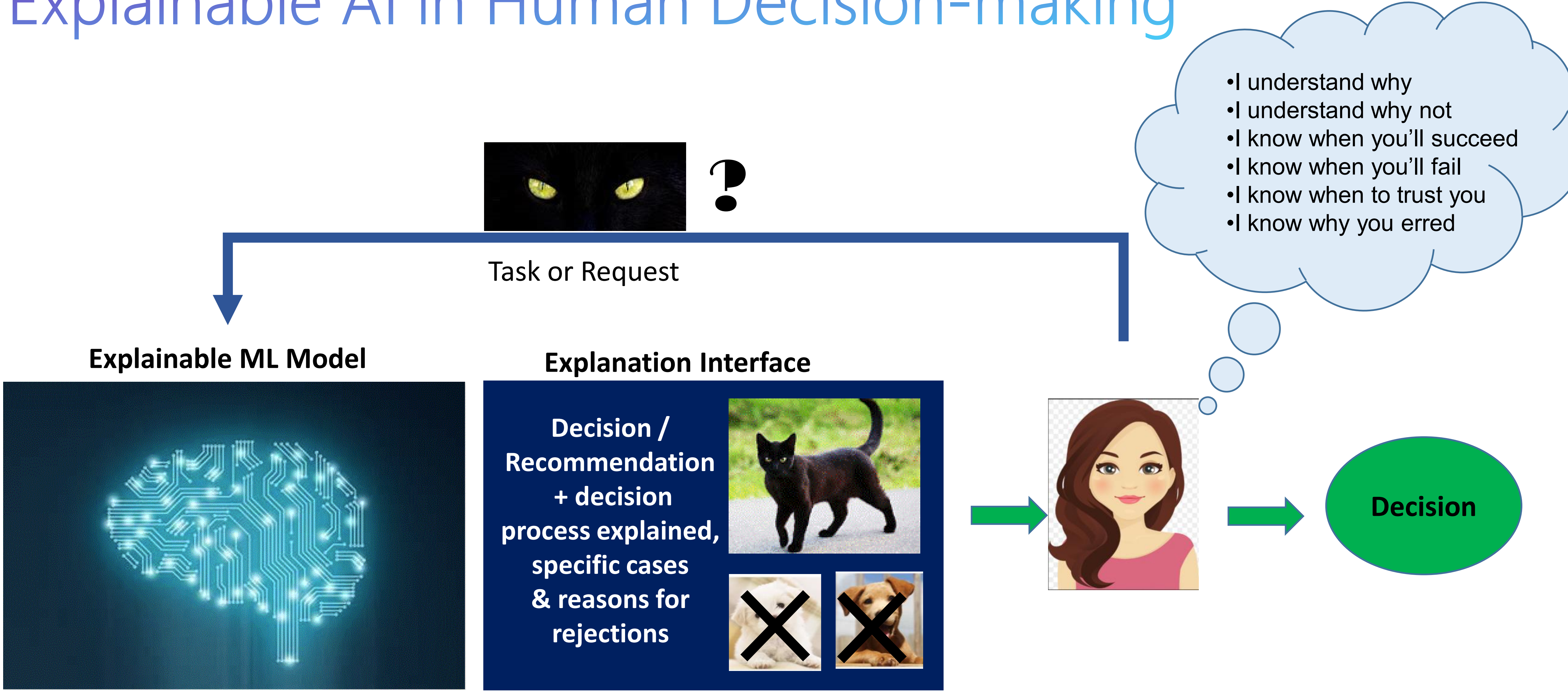
Architecture modification

Human-centred Explanation Interface

Natural Language Statements describe the elements, analytics, and context that support a choice	**Visualizations** directly highlight portions of the raw data that support a choice and allow viewers to form their own understanding
Specific Cases examples and/or stories that support the choice	**Reasons for Rejections** of alternative choices that argue against less preferred answers based on analytics, cases, and data

([1][6]))

Explainable AI in Human Decision-making



Makes recommendation, decision, or action

Explains & justifies the recommendation, decision or action to the user

([5][6])

Measuring Explainability Effectiveness

Model Understanding

Does the user understand the overall model & individual decisions, its strengths & weakness, how predictable is the models decisioning, and are there work-arounds for known weaknesses.

User Satisfaction

Based on explanation clarity & Helpfulness as measured by the user

Trustworthiness

Is the model trustworthy and appropriate for future use

Task & Decisioning Performance

Does the AI explanation improve the user's decision, Does the user understand the AI decisioning

Self-Correctability & Improvement

Does the model identifying & correct its errors. Does it undergo continuous training

([6])

Conclusion

“Life is by definition unpredictable. It is impossible for programmers to anticipate every problematic or surprising situation that might arise, which means existing ML systems remain susceptible to failures as they encounter the irregularities and unpredictability of real-world circumstances,” Hava Siegelmann. DARPA

References

[1] A. Arrieta et al, “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI”, available from: ResearchGate arXiv:1910.10045v1 [cs.AI] 22 Oct 2019, accessed March 2021.

[2] A. Bleicher, “Demystifying the Black Box That Is AI: Humans are increasingly entrusting our security, health and safety to “black box” intelligent machines”, Scientific American, August 2017, available from: https://www.scientificamerican.com/article/demystifying-the-black-box-that-is-ai/, accessed September 2021.

[3] DARPA, Researchers Selected to Develop Novel Approaches to Lifelong Machine Learning, DARPA, May 7, 2018, available from: http://ein.iconnect007.com/index.php/article/110412/researchers-selected-to-develop-novel-approaches-to-lifelong-machine-learning/110415/?skin=ein, accessed September 2021

[4] R. Guidotti et al, “A survey of methods for explaining black box models”, ACM Computing Surveys, 51 (5) (2018), pp. 93:1-93:42, accessed September 2021.

[5] D. Gunning, “Explainable Artificial Intelligence (XAI), DARPA/120, National Security Archive”, 2017, available from: https://ia803105.us.archive.org/17/items/5794867-National-Security-Archive-David-Gunning-DARPA/5794867-National-Security-Archive-David-Gunning-DARPA.pdf, accessed September 2021.

[6] D. Gunning, “Explainable artificial intelligence (XAI)”, Technical Report, Defense Advanced Research Projects Agency (DARPA) (2017), accessed March 2021.

[7] M. I. Jordan, “Artificial Intelligence—The Revolution Hasn’t Happened Yet.” Harvard Data Science Review, 1(1) 2019. Available from: https://doi.org/10.1162/99608f92.f06c6e61, accessed March 2021.

[8] M I. Jordan, “Stop calling everything Artificial Intelligence”, IEEE Spectrum March 2021, available from: https://spectrum.ieee.org/stop-calling-everything-ai-machinelearning-pioneer-says, accessed March 2021.

[9] D. Kahneman, “Thinking, fast and slow. Penguin Press, ISBN: 9780141033570, 2 July 2012”

[10] A. Korchi et al, “Machine Learning and Deep Learning Revolutionize Artificial Intelligence”, International Journal of Scientific & Engineering Research Volume 10, Issue 9, September-2019 1536 ISSN 2229-5518, accessed September 2021

[11] T. Kulesza et al. “Principles of Explanatory Debugging to Personalize Interactive Machine Learning”. *IUI 2015, Proceedings of the 20th International Conference on Intelligent User Interfaces* (pp. 126-137).

[12] B. Lake et al, “Human-level concept learning through probabilistic program induction”, 2015 Available from: https://www.cs.cmu.edu/~rsalakhu/papers/LakeEtAl2015Science.pdf, accessed September 2021.

[13] G. Lawton, “The future of trust must be built on data transparency”, techtarget.com, Mar 2021, available from: https://searchcio.techtarget.com/feature/The-future-of-trust-must-be-built-on-data-transparency?track=NL-1808&ad=938015&asrc=EM_NLN_151269842&utm_medium=EM&utm_source=NLN&utm_campaign=20210310_The+future+of+trust+must+be+built+on+data+transparency, accessed September 2021.

[14] G. Lawton, “4 explainable AI techniques for machine learning models”, techtarget.com, April 2020, available from: https://searchenterpriseai.techtarget.com/feature/How-to-achieve-explainability-in-AI-models, accessed March 2021.

[15] B. Letham et al. “Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model”. *IUI 2015, Proceedings of the 20th International Conference on Intelligent User Interfaces* (pp. 126-137).

[16] Y. Ming, “A survey on visualization for explainable classifiers”, 2017, available from: https://cse.hkust.edu.hk/~huamin/explainable_AI_yao.pdf, accessed September 2021.

[17] A. Ng, “The state of Artificial Intelligence, MIT Technology Review”, EmTech September 2017. Available from: https://www.youtube.com/watch?v=NKpuX_yzdYs, accessed September 2021.

[18] A. Ng, “Artificial Intelligence for everyone (part 1) – complete tutorial”, March 2019, available from: https://www.youtube.com/watch?v=zOI6Oll1Zrg, accessed September 2021.

[19] A. Ng, “CS229 – Machine Learning: Lecture 1 – the motivation and applications of machine learning”, Stanford Engineering Everywhere, Stanford University. April 2020. Available from: https://see.stanford.edu/Course/CS229/47, accessed September 2021.

[20] A. Ng, “Bridging AIs proof-of-concept to production gap”, Stanford University Human-Centred Artificial Intelligence Seminar, September 2020, available from: https://www.youtube.com/watch?v=tsPuVAMaADY, accessed September 2021.

[21] D. Snyder et al, “Improving the Cybersecurity of U.S. Air Force Military Systems Throughout their Life Cycles”, Library of Congress Control Number: 2015952790, ISBN: 978-0-8330-8900-7, Published by the RAND Corporation, Santa Monica, Calif. 2015

[22] D. Spiegelhalter, “Should We Trust Algorithms?”. Harvard Data Science Review, 2(1). 2020, available from, https://doi.org/10.1162/99608f92.cb91a35a, accessed March 2021.

[23] 3brown1blue, “Neural Networks: from the ground up”, 2017, available from: https://www.youtube.com/watch?v=aircAruvnKk, accessed September 2021.

Discussion & Questions

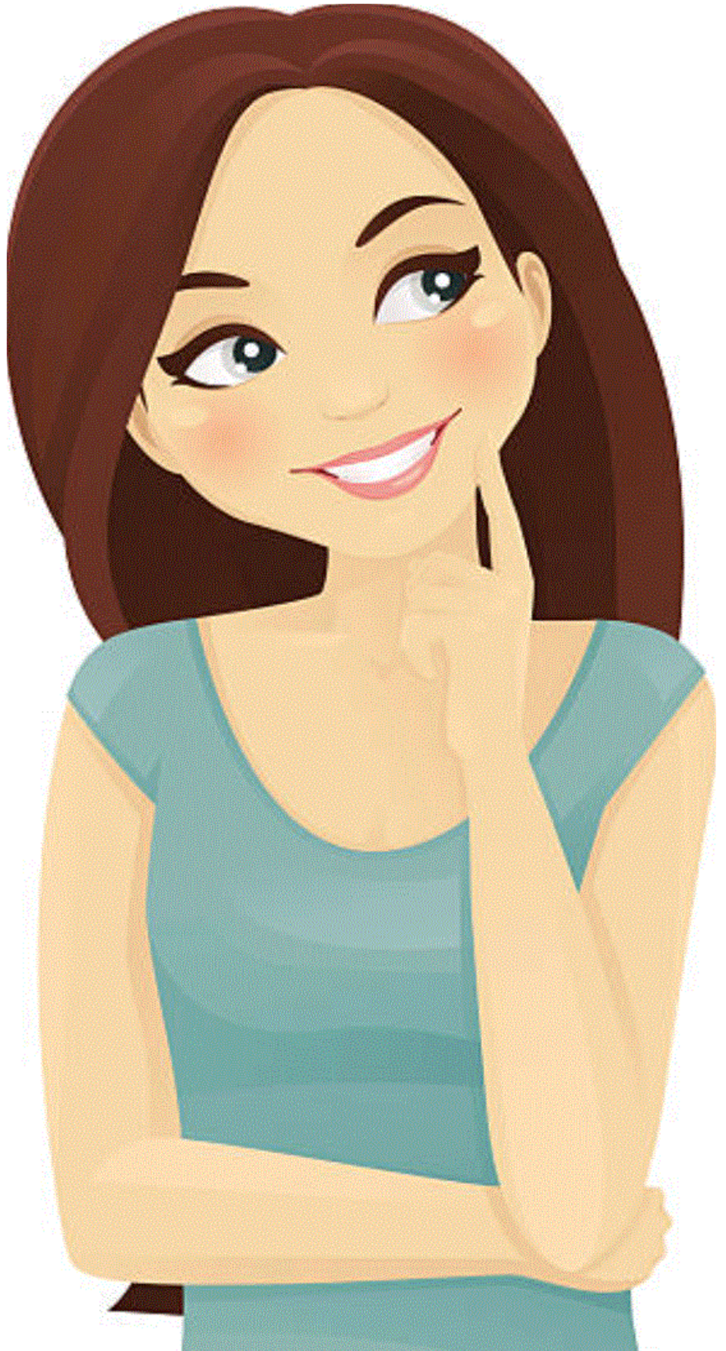
What are your experiences with classic AI and explainable AI?

How have you applied explainable AI?

Where can you see explainability will increase acceptability with your stakeholders?

email your responses to:

anne.objectiveinsight@gmail.com



Explainable AI