Explainable AI

Introduction to Artificial Intelligence and Explainability

Anne Coull
Objective Insight
Sydney, Australia
Email: anne.objectiveinsight@gmail.com

Abstract— Artificial Intelligence (AI) applies algorithms to make decisions or support human decision-making. AI has the ability to transform every industry sector. While AI cannot yet reason abstractly about real-world situations or interact socially, it is responsible for transforming the online consumer industry, facilitating biometric access to mobile phones, and for bringing science-fiction into reality with driverless cars. Explainability is a major barrier to acceptance and utilisation of AI. This is most apparent in more conservative industry sectors, such as banking and finance, health and security, where the penetration of AI is nominal. Engendering greater user acceptance of AI requires an understanding of its stakeholders, who they are, and what they need to understand. Analysis of the current machine learning models identifies three main groups in the context of explainability: Those models that are transparent and easy to understand from their logical processes; Those models that can be adjusted to take a more human-logical approach that explains itself; and those models that are so complex they need to be explained post-hoc by interpreting their behaviour. Human-centred performance measures for explainability will facilitate continuous improvement and corresponding increased acceptability of AI models.

Keywords- Artificial intelligence; machine learning, explainability, stakeholder; acceptance; transparent; model; post-hoc; human-centred explanation; measure.

I. INTRODUCTION

Artificial Intelligence (AI) utilises algorithms to make decisions or support human decision making by analysing huge data sets, finding patterns, and proposing courses of action, and they do this at scales beyond human capability [7][8]. AI has the ability to transform every industry sector by imitating and augmenting human intelligence and removing inconsistencies in human decision making [8][9][17]. While AI is responsible for the providing biometric access to mobile phones through face recognition, and for turning science-fiction into reality with driverless cars [8][17].

Explainability is a major barrier to acceptance and utilisation of AI [1][20]. "The current generation of AI systems offer tremendous benefits, but their effectiveness is constrained by the machine's inability to explain its decisions and actions" [6]. This is most apparent in more conservative industry sectors, such as banking and finance, health and security, where the penetration of AI is nominal. Explainable AI will be essential if industry leaders, professional specialists and other AI stakeholders are to understand, appropriately trust, and effectively manage this upcoming generation of artificially intelligent partners.

Section II provides an introduction to Artificial Intelligence and Machine Learning. Section III looks at some of the areas where AI has been successfully applied, and analyses why AI is not being utilized more broadly. Section IV investigates user acceptance and Section V discusses different methods for making AI and ML explainable, and approaches for interpreting ML models and generating greater user acceptance amongst the AI stakeholders.

II. ARTTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Machine Learning is one tool in Artificial Intelligence that is largely responsible for its rise. The terms Artificial Intelligence (AI) and Machine Learning (ML) are used interchangeably in literature. In 1959, Arthur Samuel defined Machine Learning as the "Field of study that gives computers the ability to learn without being programmed." Mathematical models are utilized to "train" the AI system based on large volumes of training data so that when it is presented with an input it has not seen before, it can make its own assessment and provide a predictive response [19].

More recently, in 1998, Tom Mitchel [19] described AI as the well-posed learning problem, which outlines how computer systems learn: "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**." [19]

A computer or system is said to have artificial intelligence when it has the ability to process large amounts of data, reason, and identify patterns that a human could or could not discern, at a scale unattainable by humans [1].

The science fiction stories of AI self-learning to a degree that its intelligence enables it to operate completely autonomously, while exciting, have little to do with reality. The general AI capability of these autonomous computers would necessitate them have strong capabilities in multiple intelligences and to co-ordinate these concurrently. Narrow AI, where machines are good at one capability, is now part of every-day life, and is highly lucrative for consumer

 ISBN: 978-1-61208-893-8

internet where online advertising is targeted to those consumers more likely to click [18]. More recently the self-driving car has progressed to the point of prototypes demonstrating the conflicting decisions required by an autonomous vehicle.

III. AI POTENTIAL VERSUS REALITY

A. AI Capabilities

AI can transform every major industry [17]. Many fields, particularly those with huge volumes of reliable data, are already benefiting from the support ML offers to decision making and the inferring of relations far beyond human cognitive capability [1].

Supervised learning is the most common form of Machine Learning, or AI that can perform simple A input to B output mappings [18]:

TABLE I. SUPERVISED LEARING APPLICATIONS [18]

Table Column Head		
Input	***AI***	***Output***
Email	SPAM filtering	SPAM
Audio	Speech recognition	Text transcript
English	Machine translation	Chinese
Ad, user info	Online advertising	Click?
User'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

- If the type of input is email and the output required is 'is it SPAM or not?' then the AI performs SPAM filtering, or
- If the input is an audio recording, and the output required is a text transcript then the AI is speech recognition.
- If the required input is English and output another language, Chinese, French, then this is machine translation.
- The most successful form of this type pf Machine Learning is online advertising, where all the large online advertising platforms have a piece of AI that inputs a piece of information about the ad, and some information about the user, and they try to determine the likelihood that you will click. By showing the ads the user is most likely to click on, this has become very lucrative.
- Similarly, facial recognition is used as a form of biometric access control for unlocking mobile phones.
- For a self-driving car, one of the key pieces of AI is one that takes in an image of the surroundings along with other positional input from a radar, and outputs the position of other cars and makes a decision to avoid the other cars.
- Or, in manufacturing, AI is being used to Identify manufacturing defects by taking a picture of the product being manufactured and using AI to perform visual inspection and identify any defects such as scratches [18].

This simple form of Machine learning has taken off in the last few years with the availability of large volumes of digitized data, and has proven very valuable [18].

B. Why is AI not used more broadly

ML / AI have proven useful when applied to critical areas of health care, fraud detection, and criminal justice. In healthcare, AI accelerates diagnosis and recommends treatments [22] more consistently than doctors [9]. In criminal justice it facilitates greater consistency in sentencing [22] by reducing the effects of cognitive bias [9].

Deep neural networks (DNN) have been particularly successful due to their ability to efficiently find an answer by extrapolating highly complex learning algorithms with millions of parameters [1]. The complexity of DNNs makes it difficult for a human to understand the path that was taken by the machine to get to the answer presented [1][8] and this raises questions around the trustworthiness of these AI-based systems [22]. AI has made incredible inroads into the software industry but has failed to penetrate more broadly for three key reasons: Availability of huge volumes of data to train the machine learning, ability for AI to generalise across different data sources, and the lack of acceptance by those affected.

To achieve high levels of performance at, speech and image recognition, machine learning approaches require vast quantities of training data [12] [20]. Each training data element includes an example, and a translation of what that training example means. For example, Speech recognition models require 50,000 hours of data and transcripts of that audio. Generating this training data is a significant undertaking with the data itself becoming a valuable differentiating asset. [17][20]. The areas that have extracted great benefit from AI are those with access to that data, such as Google, Facebook, and Apple.

In fields such as health, where the data volumes relate directly to the disease occurrence, the variance in data can drive variance in results. The AI may perform to human level when presented with diseases for which there is a high volume of data, but is ineffective at identifying less-common diseases. From a lung scan, for example, the AI may recognise pneumonia, peristalsis, lung cancer, or pulmonary thrombosis, but it may not recognise tuberculosis, for example.

Generalisability fails when a researcher uses data from only a few data sources. They may work closely with

radiologist from the local hospital, for example. They test their machine learning, refine their algorithms based on this limited data input and may get human-level results, but as soon as they take their model to a new context, to another hospital with a different radiologist, their AI algorithm doesn't perform so well. Here lies the gap between research and the real-world. In the health sector, different radiologists have different techniques for scanning patients. The ML algorithm may perform as well as a human with scans from a small sample set radiologists, but misdiagnose when presented with scans from another radiologist [20].

The medical practitioner will not trust the AI to diagnose her patient if she cannot rely on it to perform consistently under different circumstances [20], nor can she be confident of the AI system's diagnosis if there is no explanation for why the AI thinks there might be cancer present. This lack of transparency is fueling the gap between the research community and industry sectors, and our scans continue to be diagnosed by a human [1][16].

There has been resistance to AI in more risk averse industry sectors such as banking, finances, security and health where lack of trust in how these AI models work, along with heavy oversight by regulators, is impeding inhibiting the uptake of AI [1]. Results and metrics from AI systems may be impressive, but explainability will continue to be a barrier to AI adoption in practical implementations [1]. As Michael I Jordan explains: "We will need well-thought-out interactions of humans and computers to solve our most pressing problems" [8].

IV. USER ACCEPTANCE

Machine learning models are opaque, non-intuitive, and difficult for people to understand [5][6] and as they expand into transportation, medicine, manufacturing and defence, no-one wants to be in the position of thinking the machine is wrong, and not understanding why [2]. Explanations of AI system's reasoning is paramount for users to collaborate and trust them [16].

Acceptance requires Change Management and Explainability [20]: an understanding of how the AI model works, and how it will impact those around it.

We, as researchers and IT professionals need to face into the fact that jobs will be lost through the implementation of AI. AI can automate any one-minute task, and many jobs are made up of a sequence of one-minute tasks [17].

"Explainable AI refers to models that take action to clarify their internal functions so that a human can understand the basis of their decision making" [1]. In order to be understood, an AI model needs to be transparent, interpretable, comprehensible and intelligible by the human audience [4][13].

Elements of Explainability include:

- Trustworthiness: confidence that the model will act as intended when facing a given problem [1].
- Causality: explainable models might should ease the task of finding relationships that could be tested for a stronger causal link between the involved variables [1].
- Transferability: clarity of the boundaries that affect a model provide insight to its limitations, and to other problems. Can the model be applied in different contexts [1]?
- Informativeness: The ML model is only solving part of the problem: The problem being solved by the ML model is only a subset of the problem being addressed by its human user [1].
- Confidence: ML models need to demonstrate stability and reliability [1].
- Fairness: ML models can have built in biases. Explainability facilitates an ethical analysis of the model by exposing these biases [1].
- Accessibility & Interactivity: Explainability opens the door, in certain situations, for users to interact with the model and to be more involved in the processes of improving and developing the ML model [1].
- Privacy awareness: To satisfy GDPR and other regulatory privacy legislations, there is a need to demonstrate to customers how their data is being used [13]. Explainability highlights what data has been captured by the ML model enabling potential privacy breaches to be prevented [1].
- Cybersecurity: Engineers regularly make trade-offs between functionality and cybersecurity. As they design functionality into systems, invariably they introduce cyber vulnerabilities, knowingly and unknowingly [21]. Explainability can make cyber vulnerabilities transparent during the model development so these can be addressed or mitigative controls implemented prior to release.

Change management, along with the increased understanding given by explainability facilitates acceptance by stakeholders of the ML model.

V. EXPLAINABLE ARTIFICIAL INTELLIGENCE

A. *Classic vs Explainable AI:*

1) Classic AI

Classic AI provides the response to the question or task requested, with a confidence level in terms of a probability. It provides no details as to how it reached this outcome, merely the result it came up with. The user has no comprehension of how or why the model produced this response nor the conditions under which this response could be questionable or invalid.

2) Explainable AI

Explainable AI provides the same recommendation that the model produced but in a more understandable form, in plain English, along with the reasoning for why and how the model determined this outcome.

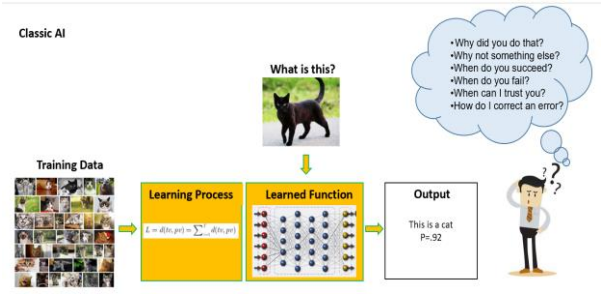

Figure 1. Classic AI [5][6]

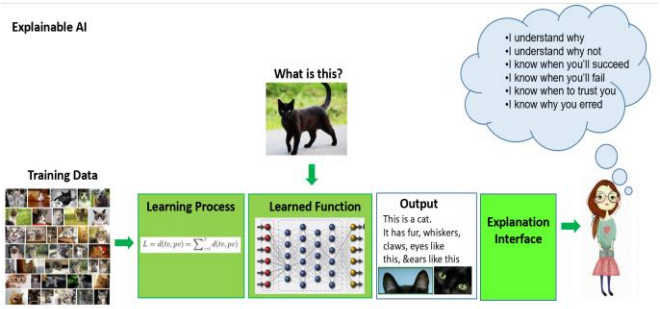

Figure 2. Explainable AI [5][6]

In the example provided, the differentiators were ear shape, eye type, and whiskers, along with other attributes associated with a four-legged furred animal. In addition, explainable AI presents this information through an interface comprehensible by humans.

B. *Explainability connects research to reality*

1. Explainability facilitates impartiality in decision-making by making any bias generated from the training set transparent so this can be corrected [1].

2. Explainability facilitates robustness by highlighting conflicting outcomes that could destabilise the predictions and make them unreliable [1].

3. Explainability can verify that only meaningful variables drive the output, providing assurance that the model's reasoning is solid and reliable [1].

Explainability requires that algorithms demonstrating accuracy in the research laboratory also perform well in practice and provides explicit evidence for this. This will raise the confidence and trustworthiness of the claims made about the system [22].

"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" [5][6].

C. *Explainable to Whom: AI Stakeholders*

The key to making an intelligent system explainable, is to understand the needs and comprehension styles of the different human stakeholders for that system [1][16].

Different audiences have different requirements. The domain experts, medical practitioners, legal judges, and the decision makers relying on the AI's recommendations will want to understand what problem the model is solving and why the model gives the answers it does. They will need to gain confidence that the model it is stable and reliable, and behaves as expected in every situation. Government regulators external to the organisation, and internal compliance will be interested in the model's ability to meet legislative compliance obligations, data privacy appetite, and cyber security objectives. Internal technology risk will want to understand vulnerabilities and data breach risks the model introduces, and how these can be mitigated. Managers, executives, and board members who may have personal liability responsibilities will need to verify that their risk and compliance exposure is being contained, and that the system is robust and reliable. They may also be interested to understand the opportunities the ML model provides and whether it is transferable to enable new business opportunities.

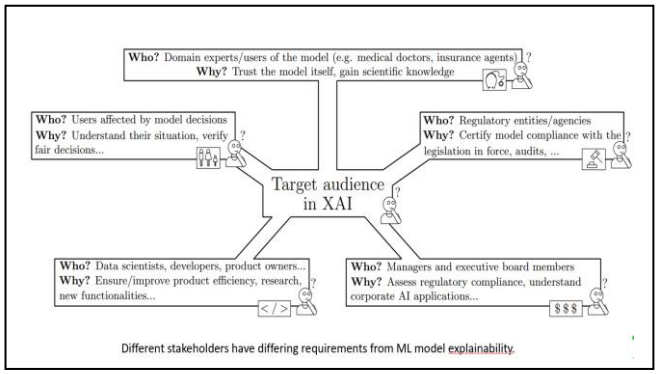

Figure 3. Stakeholders of Explainable AI [1]

Researchers, data scientists, and developers will be interested to understand how the learning model works, the training data that was used, causality between its data elements, what its boundaries and limitations are, and how this model can be applied in different contexts. The users will want to understand how the ML model benefits them and the way they work. They will want to understand the decisioning processes and how closely this aligns to their own. They may be interested in being actively involved in developing and improving the model, but they will most certainly will want to understand how this model will impact their job, their patients and/or clients [1].

D. *Explainability Approaches*

Machine Learning model explainability is categorized based on how easy it is to understand in its raw state:

 ISBN: 978-1-61208-893-8

1. Transparent interpretable models are easy for humans to understand, by reading the modelling logic;
2. Deep Explanations adjust a more complex model to incorporate explainability, and
3. Post-hoc explanations of opaque models that do not modify the model, but treat it as a black-box [1][6][20].

1) Transparent Interpretable Models

Transparent Models use simple computation structures, such as "if-then rules" within an interpretable architecture. These include:

- Logical / linear regression
- Decision trees
- K-Nearest Neighbours (KNN)
- Rule-based Learners
- General Additive Models
- Bayesian Models [1][16]

In situations where these models become extremely complex, dense, and difficult to decipher, they can be pruned or approximated with a simpler version. This involves identifying and the removing dependent support vectors and eliminating redundant parameters [16].

Bayesian Rule Lists (BRL), developed by Latham et al, provides 85-90% predictive accuracy. These models are structured as sparse decision lists consisting of a series of if… then… statements where the *if* statements list a set of features, and the *then* statements correspond to the predicted outcome. These are simple to follow and easy to understand. This list is built from the training data set: a comprehensive data set generates a comprehensive list of options and predictable outcomes. The example illustrated in Figure IV is based on the data set from the Titanic, and predicts survivability based on gender, adult-hood, and class. Given their high performance and ease of interpretability, BRLs are a preferred model for developing explainable AI [15].

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%)

Figure 4. Decision list for the Titanic survivors. In parentheses is the 95% credible interval for the survival probability [15].

2) Deep Explanations

Deep Explanations involve adjusting the model itself so that is learns in a more human-logical way and can more easily explain the steps it took to reach its decision [5][6].

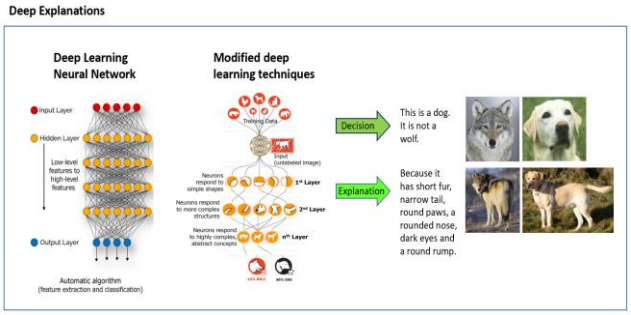

Figure 5. Deep Explanations [5][6]

3) Opaque Model Induction

The guidance from literature for classification of post-hoc explanatory models varies, but a logical method for aligning these complex ML models is to segregate them into:

- Explanation by simplification
- Feature relevance explanation
- Local Explanations
- Visual explanation, and
- Architecture modification [1].

These explanatory models deduce the decisioning of opaque ML by analysing the input to output alignments.

E. Human-centred Explanation Interface

The ultimate purpose of explainability is to enable humans to make informed decisions with valuable input from the AI system. The human-centred explanation interface translates the AI explanations, and presents them as:

- Statements in natural language that describe the elements, analytics, and context that support a choice;
- Visualisations that directly highlight portions of the raw data that support a choice and allow viewers to form their own understanding;
- Specific Cases, examples and/ or scenarios that support the choice the model made;
- Reasons for Rejection of alternative choices that argue against less preferred answers based on analytics, cases, and data [6].

 ISBN: 978-1-61208-893-8

F. *Explainable AI in Human Decision-making*

AI explainability, provided to humans through a suitable interface, will improve the uptake of AI by increasing trust and confidence in the responses the AI provides. Humans will own the decision process, with explainable AI becoming a tool commonly applied to enhance informed decision making [5][6].

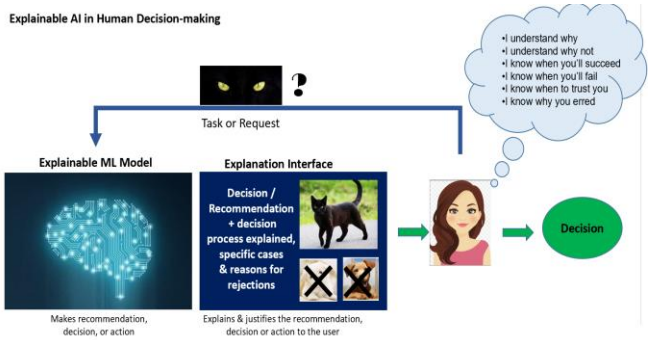

Figure 6. Explainable AI in human decision making [5][6]

The user will raise the request to the AI system, the system will provide a recommendation, along with the explanation for its recommendation, via an easy-to-understand Explanation Interface. The human can choose to take the additional information provided by the AI system into account when they make their decision, or not.

G. *Measuring Explainability Effectiveness*

To ensure explainability of AI improves over time, its effectiveness needs to be measured. Gunning proposed the following human-centric measures:

TABLE II. MEASURING EXPLAINABILITY EFFECTIVENESS

Explainability Effectiveness	
Metric	***Measure***
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 and improvement	Does the model identifying & correct its errors? Does it undergo continuous training?

VI. 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 L2M program manager [3].

It is not adequate for an AI system to merely perform its task and provide the answers. As Machine Learning and Artificial Intelligence evolve, their ability to enhance human capabilities in every industry will grow. Yet, their uptake will continue to rely on humans. AI systems will need to be able to explain themselves if humans are to trust them, understand them, and work *with* them on critical, life-affecting decisions and tasks.

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, 2015.

[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.