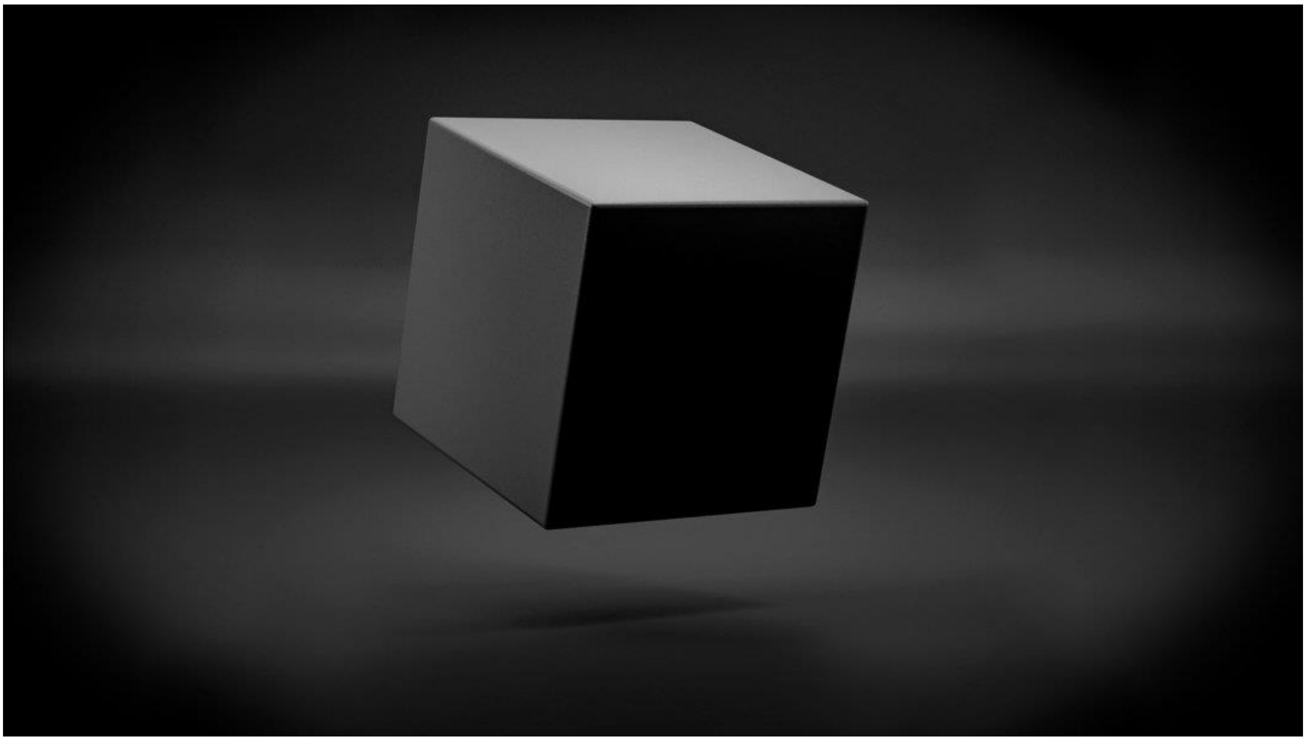
The Practical Problems of Explaining AI Black Box Systems

By Martin Anderson



First published ***May 28th, 2020*** *at:*

https://www.iflexion.com/blog/black-box-ai

Web-archived version

If your dog kills someone, it's probably not the dog that will be subject to a lawsuit. Likewise, if you empower a brilliant but unpredictable thinking system with the ability to implement critical decisions, you will have to take responsibility for its actions[1].

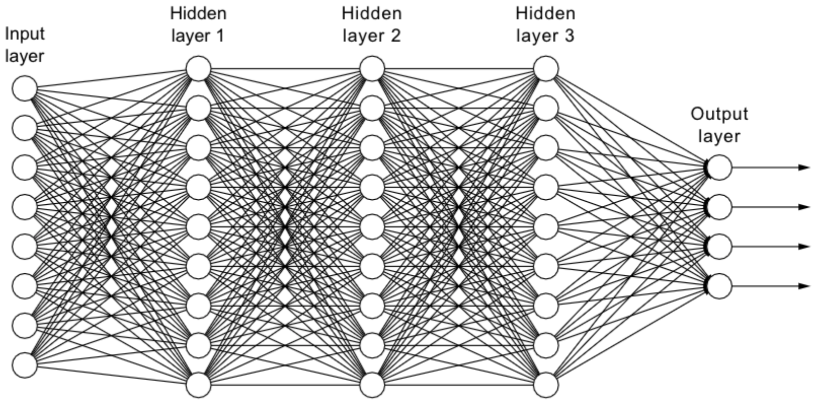
Under this basic ethical template, the issue of accountability in machine learning systems has reached a critical junction: their cost-saving potential for the public[2] and private sector[3] increasingly conflicts with their opacity, either exposing governments and companies to prosecution and sanctions, or else demanding a more circumspect approach that could slow the evolution of AI to the same crawl that led to two AI winters over the last fifty years.

The Challenge of Deconstructing AI Systems

In response to this tension, the growing field of AI deconstruction is now seeking to explain the reasoning behind the decisions and outcomes of machine learning systems, and is a rapidly growing sector with a strong commercial and political remit.

However, the labyrinthine and non-linear ways that machine learning operates on data can be difficult to monitor in a meaningful way, and even more difficult to infer from outcomes.

The internal logic flow of a neural network is threaded through a complex web of simulated neurons that evolve and change in nature as they pass through interwoven layers.



Input data becomes abstract in the hidden layers of a neural network. Source: https://freecontent.manning.com/neural-network-architectures/

Whether it's an image being splintered into pixels, or a phrase sliced into words and character frequencies, data in a neural network is broken down into particulate instances and reconstructed systematically in the search for meaningful patterns.

No mechanical observation method can easily keep track of the input information in this labyrinth, because the data itself is dissolved into a kind of 'digital gas' in the process.

Coming to Terms in Black Box AI

AI deconstruction must account for the social and practical integrity of the data. An implicit, hidden or unconscious agenda in the acquisition and design of data, models or frameworks can have a huge effect on the conclusions that AI draws from a dataset[4].

Furthermore there are semantic disputes regarding a useful and meaningful definition of 'explainability' as it relates to AI black box systems. Often, the explanations are either so dumbed-down as to be inaccurate (or at least unhelpful), or else so technical that they are difficult to understand.

In the News: Negative Indicators for Empowering AI Systems

Empowered AI can kill, and sometimes it does. But for the most part, the well-publicized failings of machine learning serve as an inhibiting factor for uptake, or for extending the autonomy of machine systems:

- 'System limitations' in Tesla's onboard AI-based Autopilot system contributed[5] to the death of a pedestrian in 2018.
- The NSA's use of a Random Forest approach in the AI of a US military drone system is reported to have caused thousands of collateral damage deaths in military operations in Pakistan[6].
- The advent of COVID-19 has compromised many brittle AI prediction models, which were not expecting such huge changes in data, and which have required manual intervention[7].
- The American Civil Liberties Union conducted a test showing that Amazon's Rekognition FR system matched up the faces of members of the US Congress with 28 known criminals, with an unpleasant additional racial bias.
- In 2014 Brisha Borden, a black minor offender aged 18 with no criminal record, was rated higher for re-offending potential than a 41 year-old white career criminal by a Florida crime prediction AI[8], leading to further scandals around bias in machine learning in the police sector[9].
- In 2016 Microsoft's experimental machine learning chatbot was hurriedly taken offline when it turned into a genocidal racist, having been easily trained by online trolls attempting to influence its personality[10].
- A Chinese tech businesswoman was shamed as a jaywalker in Ningbo by an automated face recognition system which put her picture on the side of a bus as a cautionary example to offenders, even though the jaywalker was someone else[11].
- IBM's Watson medical AI framework offered ‘unsafe and incorrect’ cancer prognoses[12], leading to the cancellation of the project in question and, eventually, layoffs in the Watson team.

Governmental and popular support for explainable AI has grown in the last five years in proportion to increased deployment, or to governments' statements of intent around incorporating AI into the public sector.

Concern has been registered over the last ten years regarding the future of big data and AI in sectors such as HR[13], public data management[14], credit scoring[15], welfare benefits access[16], health insurance[17], law enforcement[18] and vehicle systems[19], amongst many others[20].

Problems in Deconstructing Proprietary or State-Sponsored AI Algorithms

In 2017 Ed Felton, a professor of computer science and public affairs at Princeton, brought attention to the intellectual property aspect of explainable AI, wherein companies or governments may preclude reverse engineering or exposure of their algorithms, either for purposes of security, continuing exclusivity of a valuable property, or else national security[21].

Indeed, in seeking a balance between openness and security, AI research is sometimes ushered back into IP silos and away from useful public and peer scrutiny:

- One review, by the Committee on Standards in Public Life in the UK, contends that 'it may not be necessary or desirable to publish the source code for an AI system'[22].
- The UK's parliamentary Committee on Standards in Public Life observed in February 2020[23] that *'[the] continuous refinement of AI systems could also be a problem if the system is deployed in an environment where the user can alter its performance and does so maliciously'*. The committee also found that the UK government is not adequately open about the extent to which machine learning is involved in public sector decisions.

- In 2019 the research body OpenAI ironically decided not to release the source code of the text-creation algorithm GPT-2, declaring its ability to mimic human writing patterns as 'too dangerous'[24].

The impetus to withhold machine learning source code is at odds with the drive towards open standards and peer review in collaborations between the private sector and government — a further obstacle to explainable AI.

Additionally, publishing algorithms obtained through machine learning processes limits the field of AI deconstruction to a forensic, *post facto* approach, since an algorithms is only the outcome of a machine learning process, rather than the architecture of the neural network itself.

Legal Quandary: An 'Act of AI'

The ethics and legality of reverse engineering, though always contentious[25], were clearer before the age of machine learning: algorithms devised by humans were not only considered creative works that could enjoy IP protection, but lacked any concept of 'safe harbor': if the results were harmful, the creators were liable to assume responsibility and intent[26]. Under this model, due diligence was relatively straightforward.

The case for AI-generated algorithms is less clear. Proving 'intent' is problematic when the creators have an imperfect understanding of how their own algorithm was formulated. Thus, the algorithm distributors remain *responsible* for its output, without understanding exactly how the negative results transpired, or were formulated by the machine learning process — a truly toxic position.

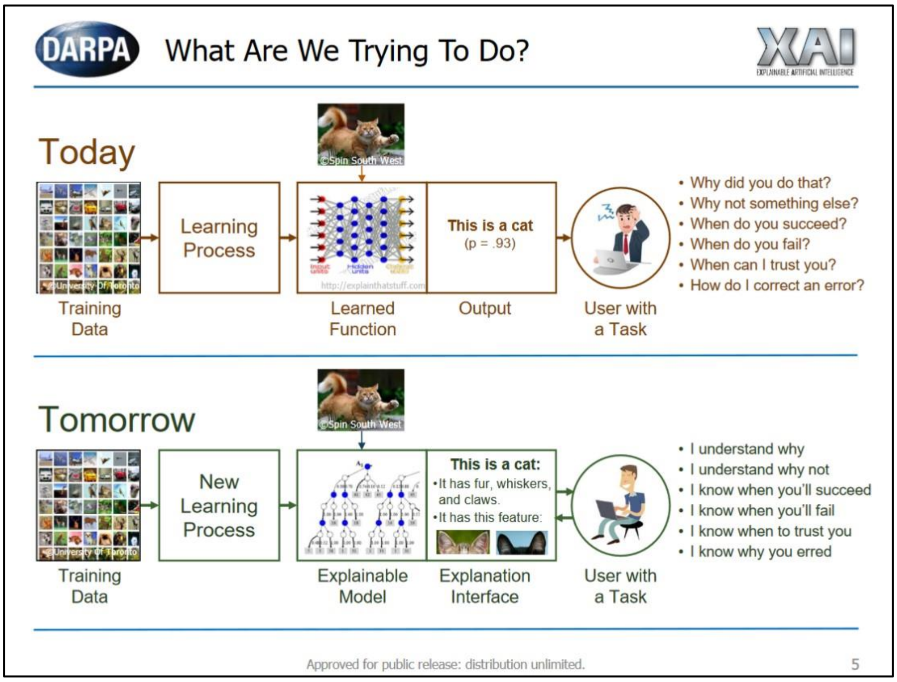
Where AI causes an unintended and negative outcome, there is a legal vacuum with regard to 'intent', since the creators presumably did not wish to cause harm. Yet 'neglect' is not applicable in the absence of any available system or process that could have prevented it.

Without definitive legislation, such events could be interpreted as 'Acts of AI', in the same sense that insurance companies define an 'act of God' as an uninsurable risk.

Thus powerful interests are both enabled and threatened by the possible outcomes of their own opaque AI systems, swept along by the economic impetus for automation and data analysis, but exposed by the lack of control mechanisms and reproducible methodologies.

DARPA's Explainable AI (XAI) Program

Though various pieces of American legislation address the issue of AI accountability to some degree[27], and though there is a notable ad hoc body of state-sponsored academic research in the field, concerns about black box AI have manifested primarily in DARPA's Explainable Artificial Intelligence (XAI) initiative[28] in the USA.



DARPA What Are We Trying To Do? XAI EXPLAINABLE ARTIFICIAL INTELLIGENCE

Today

Training Data → Learning Process → Learned Function → Output: This is a cat (p = .93) → User with a Task

- 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?

Tomorrow

Training Data → New Learning Process → Explainable Model → Explanation Interface → User with a Task

This is a cat:
- It has fur, whiskers, and claws.
- It has this feature:

- 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

Approved for public release: distribution unlimited. 5

DARPA's vision of explainable AI (XAI). SOURCE: https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

Founded in 2017, the XAI project has struggled to find effective or applicable remedies for the opacity and inscrutability of machine learning reasoning processes. Over the course of three years, its emphasis has switched from short-term architecture intervention to a collaborative academic effort to quantify, name and rationalize core concepts and possible solutions, with an emphasis on Explainable AI Planning (XAIP)[29].

The greater body of DARPA's literature centers on the forensic deconstruction of AI model decisions, with inferential analysis of AI outcomes emphasized over in-model tracking techniques.

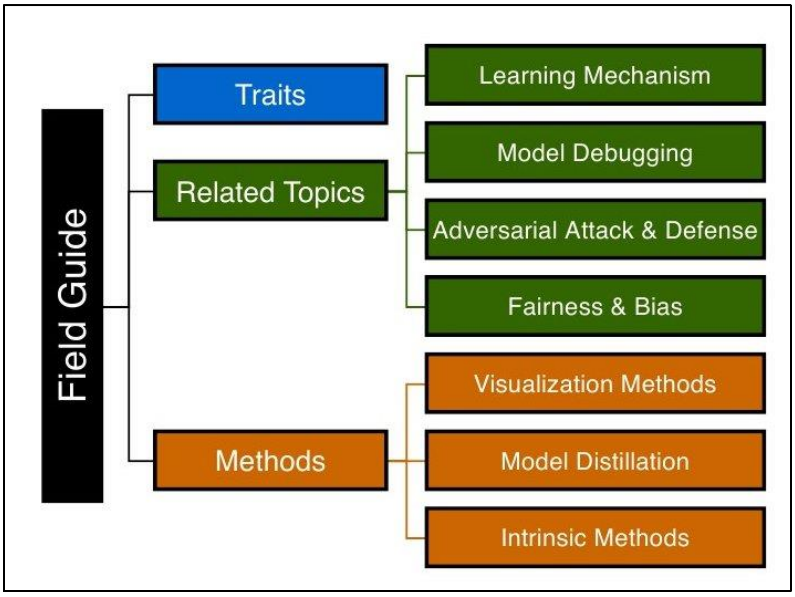
The XAI initiative proceeded from the assumption (or at least the hope) that automated reporting could be integrated into machine learning processes while keeping them performant. But a 2020 survey[30] of DARPA's curriculum by IBM and Arizona State University reflects that the project's research into automating AI explainability has evolved into a more taxing examination of potential robot/human collaboration models, with automated analysis perceived as 'computationally prohibitive'.

General Academic Research Into Explainable AI

In April 2020 researchers in the US and the Netherlands released a comprehensive overview of current initiatives in explainable deep learning. The report reveals how metaphysical the field of AI deconstruction currently is, and that the forward direction centers more around human-focused planning strategies rather than 'static analysis' techniques applied against an existing neural network.

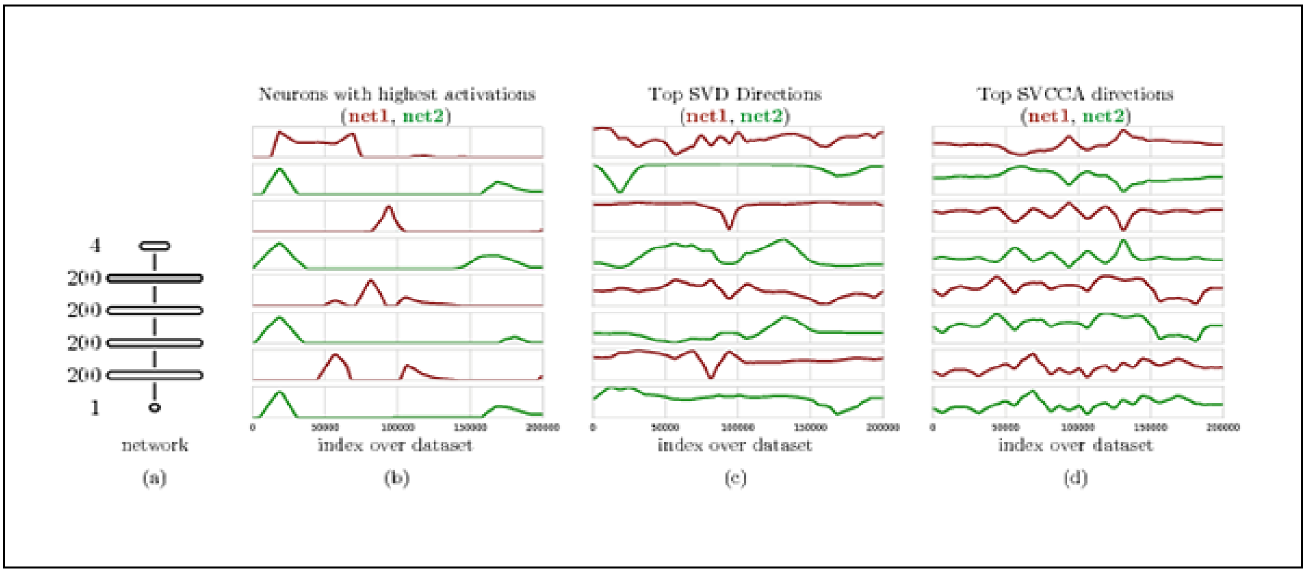
The researchers break down the challenge into three core areas that cover a great deal of interrelated research:

- ***Traits,*** wherein the objectives of AI explanations are examined, and terms for explainability are defined.
- ***Related Topics***, where XAI is stress-tested against analogous research fields.
- ***Methods***, which examines and questions recent academic assumptions in the search for foundational principles for XAI.



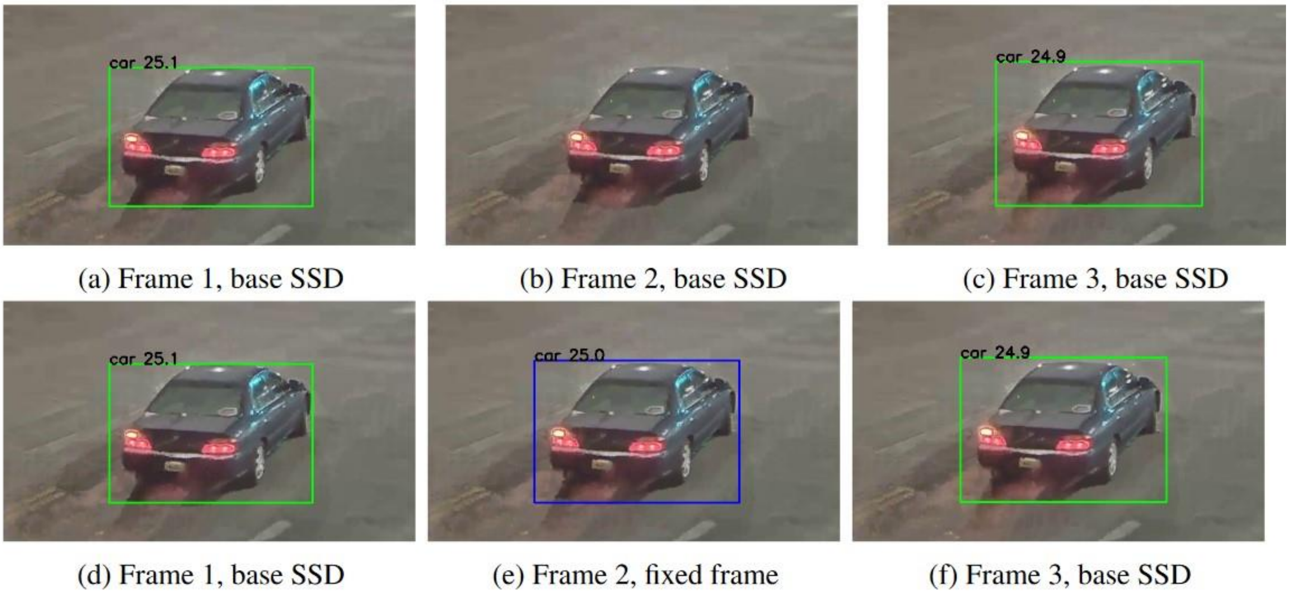
A roadmap for developing explanation systems for AI. Source: https://arxiv.org/pdf/2004.14545.pdf

One approach noted in the research is Singular Vector Canonical Correlation Analysis (SVCCA)[31], wherein individual neurons in a DNN are mapped into a vector set that can reveal layer activations at runtime.



Under Singular Vector Canonical Correlation Analysis (SVCCA), the most notably activated neurons represent a map of the transformations that tagged data are undergoing through the neural network. Source: https://www.semanticscholar.org/paper/SVCCA%3A-Singular-Vector-Canonical-Correlation-for-Raghu-Gilmer/4f03aa472ff9a8ff4e1deda52cc4f504b130d115

Though the literature on real-time model reporting is less extensive, research out of the Stanford Dawn Project proposes 'Model Assertions'[32] as a run-time method of monitoring and intervening in the live processes of a deep neural network. However, the technique's success with image processing pipelines may be difficult to replicate in more abstract realms such as language processing or sentiment analysis.

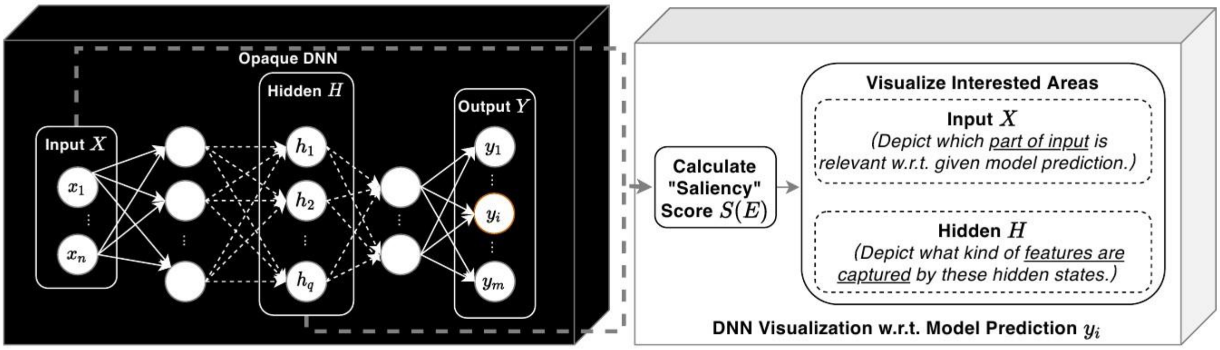


(a) Frame 1, base SSD (b) Frame 2, base SSD (c) Frame 3, base SSD

(d) Frame 1, base SSD (e) Frame 2, fixed frame (f) Frame 3, base SSD

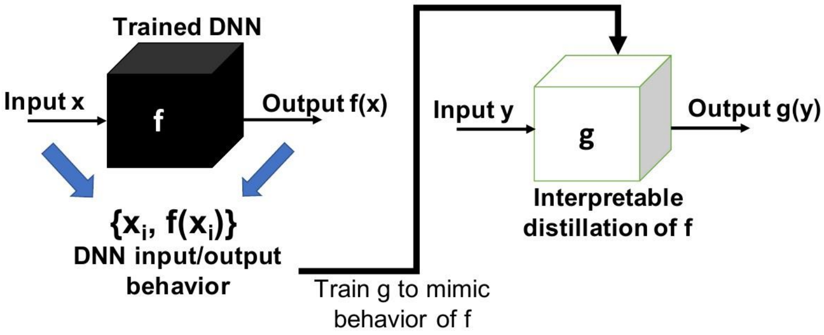
When AI image recognition loses track of an object, threatening the integrity of the output, runtime monitoring can intervene on a live basis. SOURCE: https://cs.stanford.edu/~matei/papers/2018/mlsys_model_assertions.pdf

Visualization methods can also help map model features that cause high stimulation in a neural network, and are facilitated either by back-propagation or perturbation-based visualization[33,34,35].


Opaque DNN
Input X
x_1
x_n
Hidden H
h_1
h_2
h_q
Output Y
y_1
y_i
y_m
Calculate "Saliency" Score $S(E)$
Visualize Interested Areas
Input X
(Depict which part of input is relevant w.r.t. given model prediction.)
Hidden H
(Depict what kind of features are captured by these hidden states.)
DNN Visualization w.r.t. Model Prediction y_i

When a specific input can be seen to activate features that are normally hidden in a neural network, some degree of logic-mapping becomes possible. Source: https://arxiv.org/pdf/2004.14545.pdf

Model Distillation is a post-facto technique wherein the output of a DNN is used as input data for a secondary, explicatory DNN.


Trained DNN
Input x
f
Output f(x)
{x_i, f(x_i)}
DNN input/output behavior
Train g to mimic behavior of f
Input y
g
Output g(y)
Interpretable distillation of f

With model distillation, the secondary DNN has access to both the input and output data of the primary network, and can correlate one to the other. Source: https://arxiv.org/pdf/2004.14545.pdf

A number of 'intrinsic' methods suggest possible ways that deep learning models can self-report on the logic of their process pipelines:

- **Attention Mechanisms** can assign tracking capabilities to specific inputs or else force the DNN to ask questions before it continues processing data, and current methods include *Single-Modal Weighting* and *Multi-Modal Interaction*.

- **Joint Training** proposes several methods of 'chaperoning' the data with functions designed to account for what is happening to it in a neural network. Current methods include *Text Explanation*, where a DNN is augmented with an explanation generation component; *Explanation Association*, where data is associated with human-interpretable concepts and objects, in the hope that the evolution of this secondary layer makes sense of the primary layer; and *Model Prototype*, designed for classification models, where case-based reasoning forms an association between input data and prototypes observations in the dataset.

Conclusion

In one famous example of sci-fi satire, the fictional supercomputer Deep Thought took 7.5 million years to compute the answer to 'the meaning of life', which turned out, unhelpfully, to be '42'[36].

According to the latest literature, life is imitating art, as the locus of research into XAI turns towards asking better questions, defining clearer terms and conceiving new AI frameworks whose reasoning can be made more explicable than that of current machine learning approaches.

In the balance, progress in interpretive, post facto XAI methods and principle development remains far ahead of in-process explanation mechanisms for deep learning.

[1] *https://www.theglobeandmail.com/news/world/if-a-robot-kills-someone-who-is-to-blame/article23996250/*

[2] *https://www2.deloitte.com/content/dam/insights/us/articles/3834_How-much-time-and-money-can-AI-save-government/DUP_How-much-time-and-money-can-AI-save-government.pdf*

[3] *https://www.forbes.com/sites/joemckendrick/2017/01/24/artificial-intelligence-doesnt-just-cut-costs-it-expands-business-brainpower/*

[4] *https://www.ibm.com/design/ai/fundamentals/*

[5] *https://news.sky.com/story/tesla-criticised-for-lack-of-system-safeguards-after-autopilot-crash-11943347*

[6] *https://www.theguardian.com/science/the-lay-scientist/2016/feb/18/has-a-rampaging-ai-algorithm-really-killed-thousands-in-pakistan*

[7] *https://www.technologyreview.com/2020/05/11/1001563/covid-pandemic-broken-ai-machine-learning-amazon-retail-fraud-humans-in-the-loop/*

[8] *https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing*

[9] *https://www.newscientist.com/article/2166207-discriminating-algorithms-5-times-ai-showed-prejudice/*

[10] *https://www.businessinsider.com/microsoft-deletes-racist-genocidal-tweets-from-ai-chatbot-tay-2016-3*

[11] *https://www.scmp.com/tech/innovation/article/2174564/facial-recognition-catches-chinas-air-con-queen-dong-mingzhu*

[12] *https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/*

[13] *https://www.theatlantic.com/technology/archive/2013/11/your-job-their-data-the-most-important-untold-story-about-the-future/281733/*

[14] *https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf*

[15] *https://www.brookings.edu/research/credit-denial-in-the-age-of-ai/*

[16] *https://www.theguardian.com/technology/2020/feb/05/welfare-surveillance-system-violates-human-rights-dutch-court-rules*

[17] *https://www.lexalytics.com/lexablog/ai-healthcare-data-privacy-ethics-issues*

[18] *https://www.brookings.edu/research/5-questions-policymakers-should-ask-about-facial-recognition-law-enforcement-and-algorithmic-bias/*

[19] *https://www.forbes.com/sites/lanceeliot/2019/12/18/latest-ai-that-learns-on-the-fly-is-raising-serious-concerns-including-for-self-driving-cars/*

[20] *https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/*

[21] *https://freedom-to-tinker.com/2017/05/31/what-does-it-mean-to-ask-for-an-explainable-algorithm/*

[22] *https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/868284/Web_Version_AI_and_Public_Standards.PDF*

[23] *https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/868284/Web_Version_AI_and_Public_Standards.PDF*

[24] *https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction*

[25] *https://peillaw.com/the-legalities-of-reverse-engineering/*

[26] *https://qz.com/1427621/companies-are-on-the-hook-if-their-hiring-algorithms-are-biased/*

[27] *https://www.loc.gov/law/help/artificial-intelligence/americas.php*

[28] *https://www.darpa.mil/attachments/XAIProgramUpdate.pdf*

[29] *http://icaps18.icaps-conference.org/xaip/*

[30] *https://arxiv.org/pdf/2002.11697v1.pdf*

[31] *https://papers.nips.cc/paper/7188-svcca-singular-vector-canonical-correlation-analysis-for-deep-learning-dynamics-and-interpretability*

[32] *https://cs.stanford.edu/~matei/papers/2018/mlsys_model_assertions.pdf*

[33] *https://arxiv.org/abs/1506.06579*

[34] *https://distill.pub/2017/feature-visualization/*

[35] *https://distill.pub/2018/building-blocks/*

[36] *https://www.bbc.co.uk/programmes/profiles/wqGHb88RDCJ2j8hXGwgBYn/deep-thought*