sciendo
Vietnamese Journal of Legal Sciences, Vol. 07, No. 02, 2022, pp. 01-38 DOI: 10.2478/vjls-2022-0006

EXPLAINABILITY OF ARTIFICIAL INTELLIGENCE MODELS: TECHNICAL FOUNDATIONS AND LEGAL PRINCIPLES

JAKE VAN DER LAAN*
University of New Brunswick, Canada
Email: jake.vanderlaan@unb.ca

Abstract

The now prevalent use of Artificial Intelligence (AI) and specifically machine learning driven models to automate the making of decisions raises novel legal issues. One issue of particular importance arises when the rationale for the automated decision is not readily determinable or traceable by virtue of the complexity of the model used: How can such a decision be legally assessed and substantiated? How can any potential legal liability for a "wrong" decision be properly determined? These questions are being explored by organizations and governments around the world.

A key informant to any analysis in these cases is the extent to which the model in question is "explainable".

This paper seeks to provide (1) an introductory overview of the technical components of machine learning models in a manner consumable by someone without a computer science or mathematics background, (2) a summary of the Canadian and Vietnamese response to the explainability challenge so far, (3) an analysis of what an "explanation" is in the scientific and legal domains, and (4) a preliminary legal framework for analyzing the sufficiency of explanation of a particular model and its prediction(s).

Keywords: *artificial intelligence, AI, data protection, automated decision making system, machine learning, Vietnam*

Despite already generating billions of dollars of revenue worldwide and the expectation this will grow to 15 trillion by 2030,[1] artificial intelligence (AI) - to the extent that it is actually "intelligent" - is still an immature technology. We are currently at the first level of the three levels of AI development referenced in the literature:[2]

Narrow or Weak AI - goal oriented pattern recognition based "intelligence", designed for a singular task.

---

* BBA(University of New Brunswick, Canada (UNB)), LLB(UNB), BScCS(UNB), MCS(UNB), https://www.linkedin.com/in/jakevanderlaan/. I want to thank Dr. Matthias Artzt for reviewing early drafts of this paper, and providing valuable input.

1 PwC (2017), 'Sizing the prize: PwC's Global Artificial Intelligence Study: Exploiting the AI Revolution'. Retrieved from: https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html [accessed on 17 December 2022]..

2 Wang W. and Siau K. (2019), 'Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: A review and research agenda', In: *Journal of Database Management (JDM)* 30.1, pp. 61–79. Retrieved from: https://www.researchgate.net/profile/Keng-Siau-2/publication/333423274_Artificial_Intelligence_Machine_Learning_Automation_Robotics_Future_of_Work_and_Future_of_Humanity_A_Review_and_Research_Agenda/links/5cf48f4b92851c4dd0240f42/Artificial-Intelligence-Machine-Learning-Automation-Robotics-Future-of-Work-and-Future-of-Humanity-A-Review-and-Research-Agenda.pdf [accessed on 25 July 2022]; Bostrom N. (2021), 'Superintelligence: Paths, Dangers, Strategies', in: (2014) in Treasury Board of Canada (2021), *Directive on Automated Decision-Making*, Government of Canada, Apr. 21. Retrieved from: https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592 [accessed on 07/25/2022].

General or Strong AI - general intelligence that mimics human intelligence and/or behaviours.

Artificial Superintelligence – "self-aware" machines which surpass the capacity of human intelligence and ability.

Functional Narrow AI, or perhaps its more realistic description as "machine analytics",[3] can be best thought of as a computer system which can complete a specifically bounded *prediction* task, after having learned how to do so from a set of examples.[4] Narrow AI is currently achieved primarily through "machine learning" and for the sake of simplicity I will use that term going forward.

Machine learning systems are now being used to aid in decision making in both the private and public spheres, for such things as the assessment of immigration applications, parole entitlement assessments, loan applicant reviews and insurance claim triage. Many of these decisions have impacts on an individuals' rights, entitlements and status.

A key question in the development and use of these machine learning systems is how to gain the appropriate level of comfort that the generated predictions/decisions are *right*, i.e. consistent with societal expectations of the quality of such a prediction or decision, in terms of fairness, consistency, transparency, etc.

This is a difficult task as the potential sources for getting it wrong are many. A recent meta-analysis of over 300 machine learning papers from 17 fields revealed many contained methodological errors.[5] Others have made the general assessment that machine learning systems may be (much) less reliable than they appear.[6]

Finding ways to assess the reliability and trustworthiness of machine learning systems is thus a paramount consideration in their continued use and acceptance by society and is an active area of research and development today, with new developments and insights almost on a weekly basis.

This paper will focus on exploring this question in relation to those machine learning systems which render outputs with *individual* legal effect,

3 Parentoni L. (2022), 'What should we reasonably expect from artificial intelligence?', In: *Publication pending at time of review*. Retrieved from: https://www.researchgate.net/profile/Leonardo-Parentoni/publication/361988480_What_should_we_reasonably_expect_from_artificial_intelligence/links/62d0198e953dfc1e93ff7c45/What-should-we-reasonably-expect-from-artificial-intelligence.pdf [accessed on 25 July 2022].

4 Russell S. and Norvig P. (2021), *Artificial intelligence: a modern approach, 4th Edition*. Pearson.

5 Kapoor S. and Narayanan A. (2022), 'Leakage and the Reproducibility Crisis in ML-based Science', In: *arXiv preprint arXiv:2207.07048*. Retrieved from: https://arxiv.org/pdf/2207.07048.pdf [accessed on 25 July 2022].

6 Lapuschkin S. et al. (2019), 'Unmasking Clever Hans predictors and as- sessing what machines really learn', In: *Nature communications* 10.1, pp. 1–8. Retrieved from: https://www.nature.com/articles/s41467-019-08987-4 [accessed on 25 July 2022].

and how we might move towards practical legal analyses usable in assessing whether a particular prediction based system is, in fact, doing it right *legally*, that is, generating output that does not run afoul of any legal obligations applicable to the automated decision, primarily in the realm of compliance with the legal rights of any person(s) affected by the prediction.

Having appropriate technical expertise to enable exploration of the legal issues inherent in the development of machine learning is important.[7] A core understanding of how machine learning works is helpful in exploring the explainability of these systems.[8] In fact, absent such an understanding, there is a real risk of overstating or otherwise misreading the capabilities of machine learning.[9]

With a good factual understanding of the machine learning workflow components, it is also easier to appreciate where and how errors may creep in, and how we might try to deal with them.

Even though there are excellent and valuable efforts to summarize the machine learning workflow,[10] none actually illustrate the *practical* steps in a manner easily consumable and understandable to the non-technical legal professional. I have attempted to do this in the first part of this paper, with the aid of a few illustrations, in the hope of providing a more "hands on" overview of the general process.

With an understanding of the core machine learning workflow, the paper then explores the current regulatory landscape, what an "explanation" means in the scientific and legal context, and then concludes with a preliminary framework for working through the explainability challenge for a particular machine learning system.

1. Machine learning basics

A machine learning system is computerized functionality which is able to make a prediction with respect to a predefined question. This ability to predict is created using a mathematics/statistic driven learning process which examines a large number of examples in order to build a formula for predicting an answer. Overly simplified, machine learning can be thought of as statistics based "intuition" developed by example.

7 Scherer M. U. (2015), 'Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies', In: *Harv. JL & Tech.* 29, p. 353. Retrieved from: https://euro.ecom.cmu.edu/program/law/08-732/AI/Scherer.pdf [accessed on 25 October 2022].

8 Lehr D. and Ohm P. (2017), 'Playing with the data: what legal scholars should learn about machine learning', In: *UCDL Rev.* 51, p. 653. Retrieved from: https://lawreview.law.ucdavis.edu/issues/51/2/Symposium/51-2_Lehr_Ohm.pdf [accessed on 25 July 2022].

9 In the literature review completed for this paper, the observation was made that many legal researchers and commentators do not have a strong grasp of the technical underpinnings of machine learning and make factual misstatements in that regard.

10 Lehr D. and Ohm P. (2017), *supra* note 8.

There are three general types of machine learning: (i) *Supervised Learning* – the examples dataset used for learning is labeled with the answer to the question being modeled, (ii) *Unsupervised Learning* – the examples dataset is not labeled. The goal is the identification of patterns and structures and to cluster like data together, and (iii) *Reinforcement Learning* – the dataset is a dynamic set of features (which constitute the "state" of something) and the algorithm responds sequentially to that state to advance towards a goal. Most machine learning models used to predict outcomes based on personal data use supervised learning. I will therefore focus solely on how this type of machine learning works.

1.2. Supervised learning

At a high level, creating a "prediction machine" using supervised machine learning in- volves a standard sequence of steps: creating a *dataset* from real world data relating to the prediction question; curating specific attributes within this data into *features* (a process called *feature engineering*); and then selecting an appropriate *algorithm* which can examine the features dataset and generate a formula – a recipe – (a process known as *training*) for answering the prediction question. This formula is called the *model*. Once a model has been created, we can then expose new data to that model to generate a *prediction* with respect to that new data (a process also sometimes referred to as *inference*). See Figure 1 for a simple overview of the machine learning workflow.

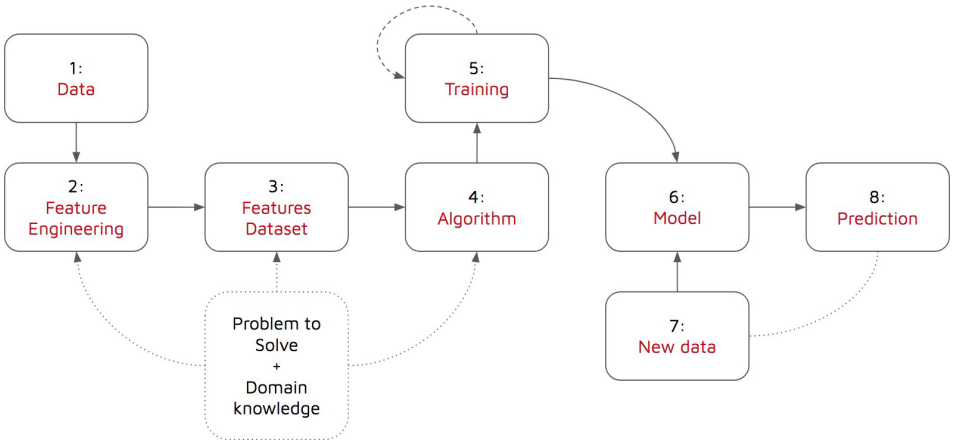


Figure 1: A high level overview of the supervised machine learning process

The predictions generated by a model can be generally classified into two broad cate- gories; *classification* and *regression*. Classification in turn is broken down into either *binary* classification, the predictive labeling of a new data sample as belonging or not belonging to a class (for example whether an email is spam or not), or *multi-class* classification, where the new data sample is predicted to belong to one of multiple classes (for example whether a fruit in

an image is a banana, apple, orange or kiwi). Regression prediction generates a predicted value in a range, for example the likely sale price for a house in a particular city.

1.3. The dataset

A dataset is a collection of similar entities with attributes, i.e., distinct pieces of information *about* those entities. For example, a dataset might contain the name, age, and financial data for a set of customers (the entities) at a bank. Datasets are often stored in a spreadsheet like format, with columns for each attribute and every row being an instance of the data entity. See Figure 2 for a simple example.

LastName	FirstName	Age	Gender	LoanRequested	CreditScore	CurrentSalary	PriorBankruptcy	LatePayments	LoanApproved
Smith	Jane	43	F	5000	560	52000	Y	N	Y
Jones	James	40	M	2500	630	43000	N	N	Y
Peters	Sally	27	F	10000	210	29000	N	Y	N
White	Michael	55	M	12500	430	76000	Y	N	N
Black	Susan	36	F	7600	510	44000	N	Y	Y
...	...	...	...		...	...	...	...	

Figure 2: A very simple example of a loan application dataset

A dataset may also be comprised of only one piece of information per entity, for example an image in a dataset of images relating to or containing a particular subject.

1.3.1. Feature engineering

Once sufficient data, configurable in a dataset, is identified, the next step is to identify which specific pieces of data, or combinations of pieces of data, to use for building a prediction model. This is one of the most important steps in the model building process. Selected pieces of data are referred as *features*. The process of identifying the right features for a particular prediction problem is called *feature engineering*.[11]

The choice of features to be used, or created, should be informed by problem domain knowledge from subject matter experts, where possible. Integrating real-world knowledge into the feature selection process helps create features which are easier to understand and interpret.[12] It also ensures that features reflect "validated information about relations between entities in certain contexts".[13]

Incorporating validated domain knowledge in the feature engineering

11 Elite Data Science (2022), 'Best Practices for Feature Engineering'. Retrieved from: https://elitedatascience.com/feature-engineering-best-practices [accessed 17 December 2022].

12 Roscher R. et al. (2020), 'Explainable machine learning for scientific insights and discoveries', In: *IEEE Access* 8, pp. 42200–42216. Retrieved from: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9007737 [accessed on 25 July 2022]

13 Rueden L. V. et al. (2019), 'Informed Machine Learning–A Taxonomy and Survey of Integrating Knowledge into Learning Systems', In: *arXiv preprint arXiv:1903.12394*. Retrieved from: https://arxiv.org/pdf/1903.12394.pdf [accessed on 25 October 2022]

process avoids the model having to learn something that is already known, enhances the precision of the model, and reduces the risk of faulty correlation.[14] The purposeful inclusion of domain knowledge is sometimes referred to as *Informed Machine Learning*. A helpful history and taxonomy of approaches to incorporating domain knowledge into feature engineering is available.[15]

For some problems, such as image recognition, it is much more difficult to incorporate domain knowledge into the feature engineering process, and as a result this step may involve some form of automated abstraction or feature identification.[16] As a whole, the quest for great features is met by finding insightful ways to *meaningfully* describe the structures inherent in the domain specific data which - together - best embody the problem reality sought to be solved, which are consistent with our understanding of the prediction question, and which appeal to human cognition and understanding.[17]

1.3.2. Feature scaling, normalization and abstraction

Machine learning algorithms require numbers as input. Features in a dataset thus need to be converted to numerical values. The complete set of numerically articulated features for a particular dataset entry is known as a *feature vector*. Converting feature data into numbers can be tricky and often requires domain expertise in order to create the optimal numerical abstraction of the specific feature. For example, how does one effectively turn colour into a number? What scale of numbers should be used? There usually are a number of different strategies to achieve this. The key is finding the best one.[18] Figure 3 provides a simple example vectorization of the data from Figure 2.

LastName	FirstName	Age	Gender	LoanRequested	CreditScore	CurrentSalary	PriorBankruptcy	LatePayments	LoanApproved
id_1	id_1a	0.43	1	5000	0.66	52	1	0	1
id_2	id_1b	0.4	0	2500	0.74	43	0	0	1
id_3	id_1c	0.27	1	10000	0.25	29	0	1	0
id_4	id_1d	0.55	0	12500	0.51	76	1	0	0
id_5	id_1e	0.36	1	7600	0.60	44	0	1	1
...	...	...	...		...	...	...	...	...

Figure 3: A very simple example of a loan approval dataset, transformed for ML

14 Borghesi A., Baldo F., and Milano M. (2020), 'Improving deep learning models via constraint-based domain knowledge: a brief survey', In: *arXiv preprint arXiv:2005.10691*. Retrieved from: https://arxiv.org/pdf/2005.10691.pdf [accessed on 25 July 2022]

15 Rueden L. V. et al. (2019), *supra* note 13.

16 Dong G. and Liu H. (2018), *Feature Engineering for Machine Learning and Data Analytics*, CRC Press.

17 Zytek A. et al. (2022), 'The Need for Interpretable Features: Motivation and Taxonomy', In: *arXiv preprint arXiv:2202.11748*. Retrieved from: https://arxiv.org/pdf/2202.11748.pdf. [accessed on 25 October 2022].

18 Bhandari A. (2020), 'Feature Scaling for Machine Learning: Understanding the Difference Between Normalization vs. Standardization'. Retrieved from: https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/ [accessed on 17 December 2022].

In the case of an image dataset, the feature vectorization process may be achieved by "flattening" each image. See Figure 4 for a simple example of an image with two shades (black and white; 1 and 0) flattened into a one-dimensional numerical vector.

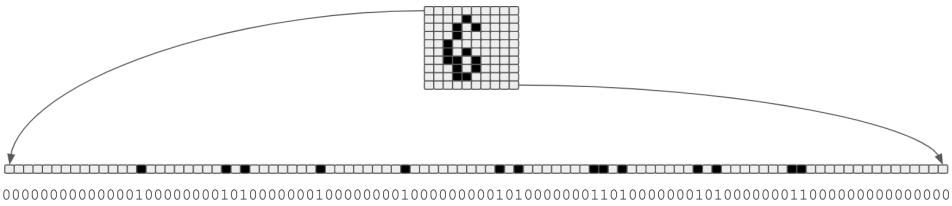


Figure 4: A two-dimensional digital image converted to a one-dimensional feature vector

1.4. The algorithm

Once the right features have been created, and subsequently properly vectorized, a modeling algorithm is selected. An algorithm is a mathematics and statistics driven tool for "learning" a formula for answering the prediction question from the sample data (the features dataset). As previously noted, this formula can then in turn be used to generate a prediction for a new instance of data (i.e., data *not* in the sample dataset).

In general terms, the more data in the training dataset, the better the algorithm will be able to learn a formula. Datasets usually contain thousands to millions of individual data entries.

The resulting formula, or model, is a mathematical function driven combination of the features in the dataset. To derive such a formula, algorithms assign a weighting variable, a *parameter*, to each feature in the feature vector for a particular date entry. In essence this parameter tracks the "amount" of that feature to use in the model being built by the algorithm. Think of a parameter as a "dial" which can be turned up or down by the algorithm as it tries to figure out how much of the feature to use as an ingredient in the model formula.

Different algorithms use different mathematical functions into which they plug the feature parameters. The type of algorithm to use for a particular modeling attempt is generally informed by the type of relationship anticipated to exist among the features within the dataset. Some of these relationships are relatively simple and can be modelled with a *linear* algorithm, whereas others are more complex, requiring a *non-linear* algorithm.

A linear algorithm can be thought of as one which uses a relatively simple combination of input feature parameters as the formula for finding correlations towards the prediction. A non-linear function uses more complex functions for combining these parameters which can represent

more complex influences by, and interactions among, the features. An algorithm may also have *hyperparameters*. These are external settings which control how the algorithm runs. Hyperparameters can be thought of as similar to configuring the performance features of a car: the number of cylinders in the engine, the type of transmission, etc. Hyperparameters may influence how efficiently an algorithm runs as well as how accurate its model will be.

Usually, an algorithm will generate a formula based on certain parameters, and then iteratively adjust the formula "dials" (the parameters for all the features) based on how accurate the predictive output currently is. This process is repeated many times until an optimal formula is found. This, in a nutshell, is the machine *learning* process.

1.5. Training: Minimizing the cost function

A cost function helps improve the current version of the model formula by calculating a "cost" for wrong outputs given the current set of "dial" configurations. Think of the cost as assigning a "bad grade" to those settings which generate a wrong output.

By adding up all the costs ("bad grades") of a particular iteration of "learning" we can estimate how well the algorithm has performed for that particular try overall.[19]

Reducing the cost function is the "learning" being done in a machine learning model and this process is commonly referred to as "training" the model.

As noted, reducing the cost function to the lowest number possible is achieved by tweaking the parameters that the algorithm assigns to the various sets of input features. This tweaking is usually done by examining data points *around* the parameter value used and determining which "direction" of changing the parameter causes the cost function for that particular parameter to decrease (the quickest). This can be achieved with relatively simple derivative calculus.[20] The cost function quantifies the *misalignment* between the model and the dataset. A lower cost function value means the model has a better fit - is more aligned - with the dataset and is thus more likely to predict better.

The actual implementation of the training process works as follows:

i. The dataset is randomly separated into two smaller sets: a *training*

19 The function used to calculate a minimum for a particular weight is actually called the "Loss Function", whereas the average of all loss function results constitutes the "cost function". For the sake of simplicity, I will simply refer to the whole process as the "cost function" here.

20 Retrieved from: https://towardsdatascience.com/understanding-the-mathematics-behind-gradient-descent-dde5dc9be06e# [accessed on 17 December 2022] for a more detailed explanation.

and a *testing* subset. In most cases the training subset comprises 80% of the dataset, and the remaining 20% is assigned to the testing subset, but different configurations are usually explored to see if they affect the accuracy of the model.

ii. Starting parameters are set, usually randomly. Any hyperparameters are also configured.

iii. The algorithm completes a preliminary calculation of the cost function for all entries of the dataset and then iteratively adjusts its parameters towards the lowest possible cost function value.

iv. Once the cost function is deemed to have been minimized as much as possible, the derived model is tested against the separate testing subset to assess its performance against new data.

v. The process terminates and a number of statistics are reported with respect to the model's accuracy.

These statistics are then used by the model developer to inform decisions about adapting the approach to modeling, which may include trying different feature combinations, different hyperparameters, as well as different algorithms, in the quest for the highest possible model accuracy. This can take a long time.

It deserves emphasizing that even though the cost function has been minimized for a particular model, this does not mean that the model will perfectly embody a mapping between the model's inputs and outputs. A model is always an approximation. It does not have the option to say "I don't know" for some of the examples in the dataset which may be difficult to incorporate into the model. As a result, there will always be errors in accuracy. It is valuable to appreciate the kinds of accuracy errors a model may exhibit.

1.6. Training Accuracy

Accuracy of machine learning models is assessed using several different metrics, all of which are calculated from the following result categories:

1. *True Positives* (TP): The number of correct positive predictions in the test subset.

2. *True Negatives* (TN): The number of correct negative predictions in the test subset.

3. *False Positives* (FP): The number of incorrect positive predictions that should have been a negative.

4. *False Negatives* (FN): The number of incorrect negative predictions that should have been a positive.

The types of metrics which may be used to evaluate the effectiveness of a particular model are briefly summarized in Chart 1 below.

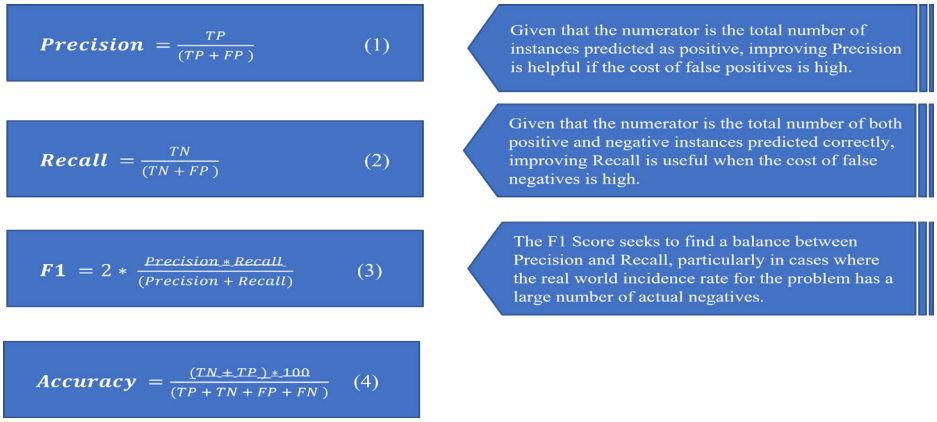


Chart 1: Metrics for model performance evaluation

It must be noted that these metrics are not necessarily a description of how *right* a particular model is, because they are completely driven by the data in the training dataset. In other words, a model may be accurate in predicting based on the training dataset but may still be "wrong" when exposed to new data.

Once a satisfactory machine learning model is developed it is often deployed into information systems where its predictions are used by other processes. At this stage certain constraints are usually present, such as the need to be able to query for a prediction quickly, which may require adaptation (simplification) of the model to accommodate such constraints.

In addition, once a model is implemented it does not grow with the environment in which it functions - it does not pick up on changes - unless it is regularly "retrained". Retraining in turn may result in now differing predictions for the same input.[21]

1.7. Deep learning

Within supervised machine learning a general distinction is made between two classes of algorithms: "traditional" statistical algorithms (such as linear regression) and more complex approaches using *neural networks*. These two approaches are sometimes referred to as "shallow learning" and "deep learning"[22] respectively.

Deep learning models are at the root of much of the explainability challenge, because they are inherently "too complex for us to explicitly

21 Benk M. and Ferrario A. (2020), 'Explaining Interpretable Machine Learning: Theory, Methods and Applications', In: *Methods and Appli- cations (December 11, 2020)*. Retrieved from: https://www.researchgate.net/profile/Andrea-Ferrario7/publication/348678581_Explaining_Interpretable_Machine_Learning_Theory_Methods_and_Applications/links/600ab71e299bf14088b21f03/Explaining-Interpretable-Machine-Learning-Theory-Methods-and-Applications.pdf [accessed on 25 October 2022].

22 Use of the word "deep" is somewhat confusing. It does not mean that the algorithm achieves a "deep" understanding but rather that it is architecturally more complex.

understand".[23] Nonetheless, a better understanding of the various components of a deep learning system will aid in dissecting the explainability challenge it poses. To that end, a brief overview of how neural networks are put together follows.

1.8. Neural networks

The term neural network finds it root in the idea that it seeks to mimic how the human brain works, i.e., through a network of layered and connected braincells, or neurons. Leaving aside the accuracy of this analogy, a neural network can be best thought of as *a series of layers of learning functionality*, which sequentially process a feature vector towards an eventual prediction (compared to a single layer of learning functionality in a shallow learning model).

These additional layers enable the modeling of much more intricate relationships among the input features.

The functional components of each layer are a set of "neurons". A neuron is a fancier version of what is known as a "perceptron".

1.8.1. Perceptrons and neurons

A perceptron is a computational device which takes one or more binary (think yes/no) inputs, and generates a single binary (yes/no) output using certain weightings of the inputs (like the parameter "dials" discussed previously) and a *threshold value*, which flips the output between Yes and No depending on whether the sum of the weighted inputs is above or below the threshold.[24]

Think of a perceptron as a decision maker which generates a decision (an output) based on the presence or absence of differing pieces of evidence (its inputs). See Figure 5.

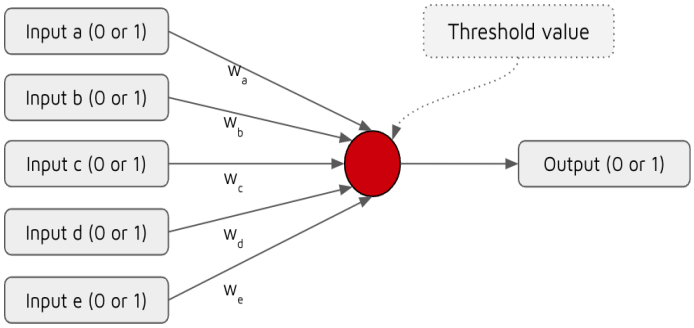


Figure 5: A simple Perceptron

23 Price II. and Nicholson W. (2017), 'Artificial intelligence in health care: applications and legal issues'. Retrieved from: https://repository.law.umich.edu/ cgi/ viewcontent.cgi?article=2932&context=articles [accessed on 27 July 2022].

24 This threshold value is effectively a measure of how easy it is to get the neuron to "fire", i.e.,generate an output a 1.

A neuron is a perceptron which takes a *fractional number* (i.e. a value between 0 and 1) as input rather than either a 0 or a 1, and which outputs a fractional number as well. A neuron can thus make a more nuanced "decision" by assigning different weights to its various inputs (rather than just use their presence or absence), as well as express its prediction with what can be thought of as "a measure of confidence".

A neural network is created by layering together different sets of neurons, with each layer's neurons taking inputs from each feature from the input features vector. Each neuron's output is in turn fed to each of the next layer's neurons, and so on, until the final output layer (representing the prediction options) is reached. See Figure 6 for a simple neural network example. The layering of sets of neurons enables longer and thus more complex computational pathways, which are able to take into account much more complex interactions *between* features.[25]

A neural network can have feature vectors comprised of thousands of features (for example one for each of the pixels in an image) and many layers of tens to hundreds of neurons between the input and output layers. Once trained, the neural network can be thought of as a complex "circuit board", custom designed to generate a prediction for a specific problem. The final "wiring" of this circuit board is achieved using an optimization approach not dissimilar to that discussed previously.

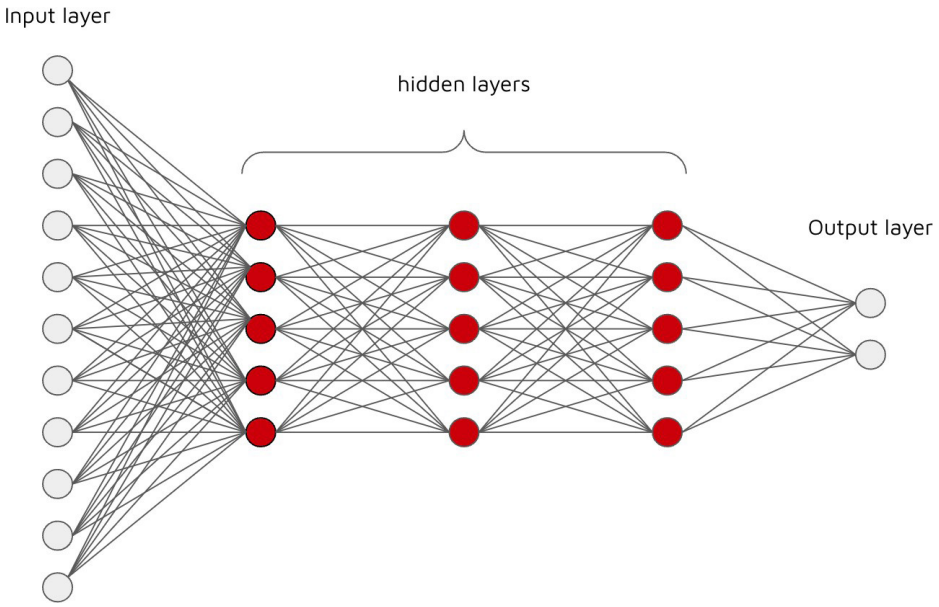


Figure 6: A simple neural network with 11 inputs, 2 outputs, and 3 hidden layers

25 Russell S. and Norvig P. (2021), *supra* note 4.

1.8.2. Optimizing a Neural network

As noted, each neuron keeps track of weights (the parameters) for each of its inputs, as well as a few other "dials" which attenuate its output based on the weights' values.[26] Based on these, it generates an output value, which it passes along to the next layer.When a neural network is trained, each of the neurons, other than those in the input layer, are set to have various random weights. Each of the neurons then calculate their output value and propagate this forward to the output layer. The first time this happens, the outputs will of course be all wrong, but that can now be improved upon by using a cost function (as explained above), which gives a "bad grade" to every wrong output in the output layer and a "good grade" to a correct one (based on the pre-labeled correct output for the particular instance of the training dataset). Averaging the cost functions for all the training instances again provides a sense of how well the overall neural network performed during a particular iteration.

The neural network layers now "feed back" the cost function results to the previous layer. Think of these as "suggestions" on how to adjust the weights of that neuron to get a better performance for the next iteration. These suggestions and resulting adjustment of weights, result in turn in a set of suggestions for the previous layer, which makes adjustments, and so on. This process is called *back propagation* and is completed all the way back to the first layer of processing - the layer right after the input layer. The algorithm then runs again, measures the error, backpropagates and updates, until the error values for the global cost function are minimized.

As the neural network adapts its neurons' weights over many iterations, weights within the various layers stabilize and create what is effectively a circuit - a set of pathways - for interpreting the input layer towards a prediction at the output layer. What these pathways actually look like and how they translate inputs into predictions is opaque. This has given rise to what is known as the "black box problem", models that generate what appear to be correct predictions, but the rationale for which is not readily determinable.

Without an accessible "explanation", the fear is that such models, or the persons who created them, may get it "wrong"[27] without the user, or the person affected, being able to tell.[28] These concerns are justified, as there exist

26 A "bias" value as well as an "activation" function. For the purpose of our simplified explanation,I have glossed over these aspects as I don't believe they contribute a great deal to understandingthe core functionality. The key idea is that a neuron has a number of dials which are adjusted towards an optimal state.

27 By "wrong" I mean that the prediction is inconsistent with societal norms for the decision in question, for example because it is based on incorrect facts, incorrect logic, or is derived as a result of some discriminatory or other form of unacceptable bias.

28 Rudin C. (2019), 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', In: *Nature Machine Intelligence* 1.5, pp. 206–215. Retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9122117/ [accessed on 27 July 2022].

many different sources for errors in prediction. These potential areas of error are generally referred to as model "biases".

1.9. Sources of bias

There are many different biases which may cause "wrong" predictions.[29] These include:

- *Selection bias or sample bias*: The training data does not represent the whole problem population accurately. This may be because the dataset over represents a particular demographic or fails to include it altogether (for example, a face recognition model trained on only a Caucasian dataset). This bias may also occur when a model trained from one population is used in a different one (for example, using a car guidance system trained in a city, in a rural area). A model will generally only work if new data comes from a similar *distribution* as that reflected in the training set.

- *Prejudice bias*: Stereotypes influence the data, either directly or indirectly, even if the dataset has been sampled perfectly. For example, a model seeking to distinguish between men and women in images, may associate kitchen attributes with women, simply because there are more pictures of women in kitchens in the dataset.

- *Historical or temporal bias*: The age of a dataset causes misalignment with current realities, skewing predictions along patterns which may no longer exist. These types of changes can also cause model degradation such as Data Drift and Concept Drift.[30]

- *Interaction bias*: A model which learns through interaction with humans may quickly learn the prejudices of those humans. See for example the experience with Microsoft's chatbot "Tay", which exhibited racist behaviour very quickly.[31]

- *Latent bias*: A model identifies an incorrect correlation based on latent features in a dataset, for example associating identification of doctors as male, because a dataset of images of doctors contains mostly images of male doctors.

- *Measurement bias*: The dataset contains faulty measurements or incorrectly entered or labeled data. This bias can also arise from seemingly innocent or

29 Gaon A. and Stedman I. (2018), 'A call to action: Moving forward with the governance of artificial intelligence in Canada', In: *Alta. L. Rev.* 56. Retrieved from: http://albertalawreview.com/index.php/ALR/article/download/2547/2514 [accessed on 27 July 2022]; Masis S. (2021), *Interpretable Machine Learning with Python: Learn to build interpretable high-performance models with hands-on real-world examples*, Packt Publishing Ltd; Mehrabi N. et al. (2021), 'A survey on bias and fairness in machine learning', In: ACM Computing Surveys (CSUR) 54.6, pp. 1–35. Retrieved from: https://arxiv.org/pdf/1908.09635.pdf [accessed on 27 July 2022].

30 Retrieved from: https://towardsdatascience.com/data-drift-part-1-types-of-data-drift-16b3eb175006 [accessed on 17 December 2022] for a more detailed explanation of these issues.

31 Vincent J. (2016), 'Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day', Tayandyou (Twitter). Retrieved from: https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist [accessed on 17 December 2022].

thought to be meaningless differences between the training dataset and new data, such as the use of a different camera from the one used in the training set for an image recognition model.

- *Omitted variable bias or exclusion bias*: Features or entity attributes that are thought to not contribute to the model, are missing or removed during the feature engineering step.

- *Observer bias*: Labels placed on example data are influenced by the subjective perspective of those doing the labeling. For example, humans may have different subjective, societal or cultural notions of human expressions when labeling faces as "angry", "surprised" or "sad".

- *Inductive bias*: Biases arising from the limitations or constraints of a particular algorithm used. For example, using an algorithm which is not capable of capturing feature interaction for a problem which does have feature interaction.

- *Aggregation bias*: False conclusions are drawn about subgroups within a dataset as a result of generalizations observed from the entire population. For example, ethnic subgroups within the diabetes population may have markedly different morbidity from the average. There are a number of examples from different domains where the average does not match any of the population's subgroups, resulting in what is known as Simpson's Paradox[32].

- *Longitudinal data bias*: Failure to synchronize time-based modeling data over time. For example, a study of Reddit comments seemed to suggest that while average comment length decreased over time, when parsed by account age they actually increased over time[33].

- *Evaluation bias*: Use of inappropriate or disproportionate benchmarks for evaluation of a model. For example, relying on Precision when the cost of false negatives is high.

- *Overfitting bias*: A model learning irrelevant details and noise as meaningful aspects of the training data to the extent that it negatively impacts its performance.[34]

2. Technical explainability strategies

Against the background of the above range of potential bias risks, the general opacity of neural network derived models, as well as numerous

32 Suresh H. and Guttag J. (2021), 'A framework for understanding sources of harm throughout the machine learning life cycle', In: *Equity and access in algorithms, mechanisms, and optimization*, pp. 1–9. Retrieved from: https://dl.acm.org/doi/fullHtml/10.1145/3465416.3483305 [accessed on 27 July 2022].

33 Barbosa S. et al. (2016), 'Averaging gone wrong: Using time-aware analyses to better understand behavior', In: *Proceedings of the 25th International Conference on World Wide Web.*, pp. 829–841. Retrieved from: https://dl.acm.org/doi/pdf/10.1145/2872427.2883083 [accessed on 27 July 2022].

34 The potential for overfitting can usually be reduced with effective model validation with separate data.

examples of models with material errors,[35] concern about being able to effectively assess the "rightness" of a particular model has grown.

In response, there is currently much academic interest in developing technical "explainability" methodologies for analyzing machine learning models. The technical strategies for producing such insights into a model can are classified as either being *global* – creating insight into how the model behaves as a whole – or *local* – focusing on providing greater clarity around a specific prediction.

2.1. Global methods

Global methods usually provide statements about which features, or combination of features, "drive" the predictions generated by the model. Some examples:[36]

2.1.1. PDP - Partial dependence plot

This relatively simple strategy examines the independent effect of each feature on the outputs, by changing its value while keeping other feature values the same, and repeating this for all features.

It thus provides a metric for describing how each feature affects the output *on average*, i.e., its average importance in all predictions. PDP is only reliable where features do not interact, i.e., have effects on each other.

2.1.2. ALE - Accumulated local effects

This variant of PDP is used where features are dependent or correlated (influence each other). This method thus helps to interpret the *relative importance* of features and their combined effects, on the output.

2.1.3. H-Statistic - Feature interaction

This metric calculates the extent to which variation of the prediction depends on the interaction of a feature with other features. It thus helps to *quantify the interaction* of feature values and the extent to which the prediction is the result of such joint effects.

It must be noted that an H-statistic tells us the extent of interaction, but not what the actual interaction is. H-statistic analysis is also not meaningful in image classification.

2.1.4. Permutation feature importance

This approach measures the importance of a feature by calculating the increase in the model's prediction error after changing its value. A feature is "important" if changing its value increases the error of the model's output.

---

35 Kapoor S. and Narayanan A. (2022), *supra* note 5; Northcutt C. G., Athalye A., and Mueller J. (2021), 'Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks', In: *arXiv preprint arXiv:2103.14749*. Retrieved from: https://arxiv.org/pdf/2103. 14749.pdf [accessed on 27 July 2022].

36 Molnar C. (2020), *Interpretable machine learning*. Retrieved from: https://christophm.github.io/interpretable-ml-book/index.html [accessed on 27 July 2022].

2.1.5. Surrogate models

A surrogate model is a new machine learning model usually trained on the original dataset as inputs, and the predictions of the model to be explained as outputs (its labels).

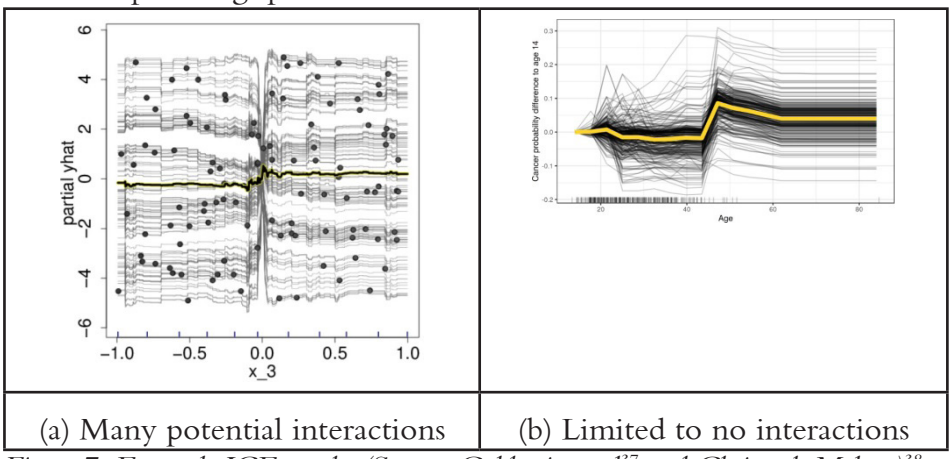
A surrogate model makes statements about the model to be explained, but *not* about the "correctness" of that model in interpreting the dataset. There is lots of academic debate around the utility of surrogate models.

2.2. Local methods

Local methods aid in explaining a *particular prediction* generated by a model, usually by testing how robust that prediction is. This is achieved in a few different ways:

2.2.1. Individual conditional expectation

An Individual Conditional Expectation (or ICE) plot, is created by generating a PDP (see above) for all instances in the dataset. It thus visualizes the dependence of the prediction on a feature for each instance separately. An ICE plot provides some insight into how "interactive" a particular feature is. See figure 7 for example ICE graphs of an interactive versus a non-interactive feature.



(a) Many potential interactions | (b) Limited to no interactions

Figure 7: Example ICE graphs (Source: Goldstain et al[37] *and Christoph Molnar)*[38]

2.2.2. Local surrogate models (LIME)

The Local Interpretable Model-agnostic Explanation (LIME) based approach was first presented in 2016,[39] and is thus a newer methodology. A LIME implementation seeks to create a surrogate model by using

37 Goldstein A. et al. (2015), 'Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation', In: *journal of Computational and Graphical Statistics* 24.1, pp. 44–65. Retrieved from: https://arxiv.org/ pdf/1309.6392.pdf [accessed on 27 July 2022].

38 Molnar C. (2020), *supra* note 36.

39 Ribeiro M. T., Singh S., and Guestrin C. (2016), 'Why should I trust you? - Explaining the predictions of any classifier', In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144. Retrieved from: https://arxiv.org/pdf/1602.04938.pdf [accessed on 27 July 2022].

multiple variations of the feature values of the original dataset (an idea called *perturbation*) and the resulting predictions from the original model. This new larger dataset is now used to train what is thought to be a more interpretable model.

Even though LIME is a promising approach, many problems need to be solved before it can be safely applied.[40] For example, the LIME approach assumes linear behavior of the machine learning model locally, but there is no (current) theory as to why this should work.[41]

There is also the observation that a really effective LIME implementation begs the question whether the dataset could have been properly modeled with a simpler model to begin with.

2.2.3. Scoped Rules (Anchors)

Anchor identification is another recently developed approach which involves the deriving of certain "if-then" type rules for the model, specifically where certain feature values determine the output, irrespective of the values of other features.

The process for generating these rules also involves "perturbing" feature values and looking for their effect among similar (neighbouring) instances within the dataset[42]. An example anchor rule from[43] is provided in Figure 8.

IF Country = United-States **AND** Capital Loss = Low
AND Race = White **AND** Relationship = Husband
AND Married **AND** 28 < Age ≤ 37
AND Sex = Male **AND** High School grad
AND Occupation = Blue-Collar
THEN PREDICT Salary > $50K

Figure 8: An example anchor rule for a loan application model (Source: Marco Tulio Ribeiro et al)[44]

This approach requires significant hyperparameter configuration in order to generate good results and, like LIME, is still a work in progress.

2.2.4. Counterfactual Explanations

Counterfactual Explanation is another "what-if" feature tweaking approach. The general idea is to explore how much specific features need to be changed in order to achieve a different prediction. Put another way, a counterfactual of a prediction captures the smallest change to feature values which change the

40 Molnar C. (2020), *supra* note 36.
41 *Ibid*.
42 Ribeiro M. T., Singh S., and Guestrin C. (2018), 'Anchors: High- precision model-agnostic explanations', In: *Proceedings of the AAAI con- ference on artificial intelligence*, Vol. 32. Retrieved from: https://ojs.aaai.org/index.php/AAAI/article/view/11491/11350 [accessed on 27 July 2022].
43 *Ibid*.
44 *Ibid*.

prediction output in a particular way. A good example is how much a particular person's income or other attributes need to change, in order to flip a previously denied loan application to approved. Large required changes mitigate towards more robustness of a particular prediction.

Counterfactuals are intuitive and can help provide real insight into how a model predicts, but it is important to create them properly. Specifically, counterfactuals should be realistic and involve as few features as possible, curated using a diverse process.[45] However, counterfactuals (and some other interpretability methodologies) may suffer from what is known as the "Rashomon[46] Effect" - contradictory, but plausible, explanations of the same outcome, from different (feature) perspectives.

2.2.5. Shapley Values

A Shapley value seeks to determine how much each feature contributes to a specific prediction. Shapley values are a common game theory driven strategy for distributing gains and costs among a series of actors, and are used in numerous other domains.[47] In machine learning, a Shapley value can be thought of as *usefulness/importance distribution* among the feature set of a particular prediction. An example is provided in Figure 9.[48]

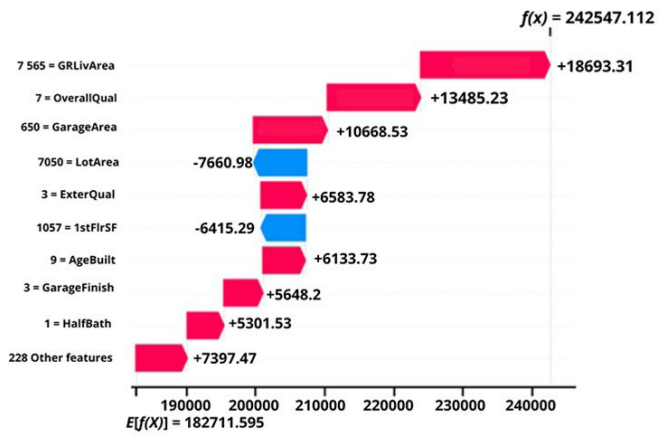


Figure 9: An example Shapley value distribution for the predicted sale price of a specific house (Source: Soufiane Fadel[49])

45 Dandl S. et al. (2020), 'Multiobjective counterfactual explanations', In: *International Conference on Parallel Problem Solving from Nature*. Springer, pp. 448–469.

46 Rashomon is a 1950 Japanese film where the murder of a Samurai is described by different people, all with different "versions" of the truth, many of which contradict each other.

47 Roth A. E. (1988), 'Introduction to the Shapley value', In: *The Shapley value*, pp. 1–27. Retrieved from: http://library.fa.ru/files/Roth2.pdf#page=9 [accessed on 27 July 2022].

48 Retrieved from: https://www.statcan.gc.ca/en/data-science/network/explainable-learning [accessed on 17 December 2022].

49 Fadel S. (2022), *Explainable Machine Learning, Game Theory, and Shapley Values: A Technical Review*. Statistics Canada, Oct. 7. Retrieved from: https://www.statcan.gc.ca/en/data- science/network/explainable- learning [accessed on 10 July 2022].

Shapley values are easily understood and unlike many of the other approaches set out above, offer a complete interpretation of which features "drive" a particular prediction. A downside to Shapley value calculation is that it is computationally demanding, particularly for large feature sets. As a result, in most applications the Shapley value can only be approximated. Nonetheless, there appears to be a growing consensus that using Shapley value-based approaches are a key insight and hold promise.[50]

2.2.6. SHAP

SHapley Additive exPlanations, or SHAP, were developed in 2017,[51] and is a newer computation method for Shapley values which adds global interpretation methods based on combinations of Shapley values across the full dataset.

2.3. User tools

In addition to the above technical tools, there are also some other approaches to pro- viding insight into machine learning models worth mentioning.

2.3.1. Model "fact sheets"

One group of researchers has proposed and implemented a fact sheet-based approach to providing general information about a particular machine learning model, not unlike the "Nutritional Facts" label usually found on food.[52]

2.3.2. Interactive model exploration

One paper suggests creating tools which allow users to "play" with the features in a particular model, along the lines of what many of the above local methods seek to do.[53] This would then permit a specifically affected user to gain specific insight into "their" prediction.

2.4. Model "Correctness"

Most of the above technical strategies are applied to an already rendered model and do not necessarily provide any assessment of the overall "correctness"

50 Retrieved from: https://www.statcan.gc.ca/en/data-science/network/explainable-learning. Retrieved from: https://papers.nips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Reviews.html [accessed on 17 December 2022].

51 Lundberg S. M. and Lee S.-I. (2017), 'A Unified Approach to Interpreting Model Predictions', In: *Advances in Neural Information Process- ing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc.. Retrieved from: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf [accessed on 27 July 2022].

52 Baracaldo N. et al. (2022), 'Towards an Accountable and Reproducible Federated Learning: A FactSheets Approach', In: *arXiv preprint arXiv:2202.12443*. Retrieved from: https://arxiv.org/pdf/2202.12443.pdf [accessed on 17 December 2022]; Arnold M. et al. (2019), 'FactSheets: Increasing trust in AI services through supplier's declarations of conformity', In: *IBM Journal of Research and Development* 63.4/5. Retrieved from: https://aifs360.mybluemix.net/introduction [accessed on 17 December 2022]; Retrieved from: https://aifs360.mybluemix.net/introduction [accessed on 17 December 2022]. See specifically the example overview of a mortgage loan application review predictor model at https://aifs360.mybluemix.net/examples/hmda [accessed on 17 December 2022].

53 Edwards L. and Michael V. (2017), 'Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for', *Duke L. & Tech. Rev.* 16, p. 18. Retrieved from: https://discovery.ucl.ac.uk/id/eprint/1574817/1/Veale_slavetothealgorithm_published.pdf [accessed on 27 July 2022].

of a model. Most seek to demonstrate a certain degree of consistency in the predictions made. As such, they only go so far in building comfort that a particular model is generating the "right" predictions. This would seem to require a broader and more holistic analysis of the model building process, in addition to these technical analyses.

With the technical overview complete, let us now turn to the legal dimensions of "explainability".

3. Regulating explainability

3.1. Canada

Machine learning systems are being explored, created and/or used in Canada in a number of domains with varying maturity. These include the surveillance of investment ad- visors and capital markets,[54] modeling outcomes in tax law,[55] prediction of legal appeal outcomes,[56] triaging temporary visa applications,[57] criminal recidivism risk assessment,[58] and health insurance fraud detection,[59] among others.

The Government of Canada is examining automated decision-making systems focused legislation as well. Recently, it has proposed the Bill C-27,[60] which currently at the First Reading phase in the Canadian House of Commons. Bill C-27 is designed to update Canada's federal private sector privacy law, the Personal Information Protection and Electronic Documents Act (PIPEDA), to create a new tribunal, and to propose new rules for artificial intelligence (AI) systems. Bill C-27 is actually a re-working of Bill C-11, the Digital Charter Implementation Act, that was introduced

54 Lokanan M. E. and Sharma K. (2022), 'Fraud prediction using ma- chine learning: The case of investment advisors in Canada', In: *Machine Learning with Applications* 8. p. 100-269. Retrieved from: https://www.sciencedirect.com/science/article/pii/S2666827022000111/pdfft?isDTMRedir=true& download=true [accessed on 27 July 2022].

55 Alarie B., Niblett A., and Yoon A. H. (2016), 'Using machine learning to predict outcomes in tax law', In: *Can. Bus. LJ 58,* p. 231.

56 Almuslim I. and Inkpen D., 'Legal Judgment Prediction for Cana- dian Appeal Cases', In: *2022 7th International Conference on Data Sci- ence and Machine Learning Applications (CDMA)*. IEEE. 2022, pp. 163– 168.

57 Nalbandian L. (2021), 'Using Machine-Learning to Triage Canada's Tempo- rary Resident Visa Applications', In: *Ryerson Centre for Immigration and Settlement (RCIS) and the CERC in Migration and Integration*. Retrieved from: https://www.torontomu.ca/content/dam/centre-for-immigration-and-settlementtmcis/publications/workingpapers/2021_9_Nalbandian_Lucia_Using_Machine_Learning_to_Triage_Canadas_Temporary_Resident_Visa_Applications.pdf [accessed on 27 July 2022]

58 Ghasemi M. et al. (2021), 'The Application of Machine Learning to a General Risk–Need Assessment Instrument in the Prediction of Criminal Recidi- vism', In: *Criminal Justice and Behavior* 48.4, pp. 518–538. Retrieved from: https://journals.sagepub.com/doi/pdf/10.1177/0093854820969753 [accessed on 27 July 2022].

59 Gill J. K. (2020), 'Health insurance fraud detection', In:. https://era.library.ualberta.ca/items/e68678e1-1021-4e4c-8fa2-54455deb9fd0 [accessed on 27 July 2022]; Goldstein A. et al. (2015), *supra* note 37.

60 *The Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts*. Retrieved from: https://www.parl.ca/DocumentViewer/en/44-1/bill/C- 27/first-reading [accessed on 17 December 2022].

in November 2020, but died on the order paper with the announcement of the federal election. Notably, a significant portion of Bill C-11 has been transported over to Bill C-27.

Bill C-27, among numerous other data protection focused provisions, will require certain disclosures by those seeking to use an "automated decision system". The relevant provisions are as follows:

[…] *automated decision system* means any technology that assists or replaces the judgment of human decision-makers through the use of a rules-based system, regression analysis, predictive analytics, machine learning, deep learning, a neural network or other technique.

[...]

high-impact system means an artificial intelligence system that meets the criteria for a high-impact system that are established in regulations.

[…]

Publication of description — making system available for use 11 (1) A person who makes available for use a high-impact system must, in the time and manner that may be prescribed by regulation, publish on a publicly available website a plain-language description of the system that includes an explanation of

(a) how the system is intended to be used;

(b) the types of content that it is intended to generate and the decisions, recommendations or predictions that it is intended to make;

(c) the mitigation measures established under section 8 in respect of it; and

(d) any other information that may be prescribed by regulation.

Publication of description — managing operation of system (2) A person who manages the operation of a high-impact system must, in the time and manner that may be prescribed by regulation, publish on a publicly available website a plain-language description of the system that includes an explanation of

(a) how the system is used;

(b) the types of content that it generates and the decisions, recommendations or predictions that it makes;

(c) the mitigation measures established under section 8 in respect of it; and

(d) any other information that may be prescribed by regulation.

[...]

Policies and practices

62(1) An organization must make readily available, in plain language, information that explains the organization's policies and practices put in place to fulfill its obligations under this Act.

[...]

(2) In fulfilling its obligation under subsection (1), an organization must

make the following information available:

[...]

(c) a general account of the organization's use of any automated decision system to make predictions, recommendations or decisions about individuals that could have a significant impact on them;

[...]

Automated decision system

63 (3) If the organization has used an automated decision system to make a prediction, recommendation or decision about the individual that could have a significant impact on them, the organization must, on request by the individual, provide them with an explanation of the prediction, recommendation or decision.

Explanation

(4) The explanation must indicate the type of personal information that was used to make the prediction, recommendation or decision, the source of the information and the reasons or principal factors that led to the prediction, recommendation or decision. [Emphasis Added]

What constitutes an "explanation" is not further defined in the proposed Act. The Canadian government has provided a "Directive on Automated Decision Making",[61] but this document also lacks any details on the meaning of "explanation". The associated "Algorithmic Impact Assessment Tool",[62] although very useful, does not provide further clarity on the issue either.

The question of what constitutes an explanation of a machine learning system has not (yet) been considered by Canadian Courts. The Supreme Court of Canada has, however, recognized the use of a faulty automated model as justification for judicial relief from the resulting decision.[63]

The Canadian legal community has generated a number of interesting commentaries tangential to the issue of explainability:

Ann Cavoukian, the former Information and Privacy Commissioner for Canada, contin- ually advocates for transparency and accountability of algorithms as part of her seven core principles of AI Ethics by Design.[64] She proposed a governance model for AI and examines the various dimensions of the government's role in ensuring the responsible use of AI technologies, which is an excellent illustration of some of the dynamics of building machine

61 Treasury Board of Canada (2021), *Directive on Automated Decision-Making*, Gov- ernment of Canada, Apr. 21. Retrieved from: https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592 [accessed on 07/25/2022].

62 Retrieved from: https://www.canada.ca/en/government/system/digital-government/digital- government-innovations/responsible-use-ai/algorithmic-impact-assessment.html [accessed on 17 December 2022].

63 *Ewert* v. *Canada*, 2018 SCC 30.

64 Retrieved from: https://www.ryerson.ca/content/dam/pbdce/papers/AI_Ethics_by_Design.docx [accessed on 17 December 2022].

learning models of judicial decision making.[65] This particular effort sought to model the length of notice period awarded in legal decisions regarding wrongful dismissal. The model was unable to generate satisfactory results, it appears primarily due to a small dataset size, low feature count, and perhaps poorly attenuated or incomplete feature engineering. It may also be that judicial decision making is hard to model because of the role of a certain amount of subjectivity in these assessments, contributed to in some degree by the variance in predispositions, life experiences, and perspectives among judges. There appears to be no Canadian legal commentary which more deeply explores how to legally define the concept of an "explanation" in the context of machine learning based systems.

3.2. Vietnam

Vietnam has experienced significant industrialisation, modernisation, and international integration over the last four decades.[66] Digital transformation is expected to play a key role in advancing economic development of Vietnam and the government has been a big driver for digital innovation.[67] Much like many other countries, the prospect of increasing use of artificial intelligence to aid in decision making is motivating academic examination of the proper parameters of a legal framework for the use of artificial intelligence.[68] The Vietnamese government has demonstrated strong interests toward research and development of a proper AI ecosystem.[69]

The Vietnamese policy makers see AI technology as a core driver for the acceleration of the economy. In 2021, the Prime Minister issued the National Strategy on Research, Development and Application of AI to 2030,[70] which aims to promote research, development and application of AI technology in contributing to socio-economic development of the country, and gradually turning Vietnam into a centre for AI in the region and in the world.[71] In Vietnam, AI has been recently applied in various fields such as healthcare,

65 Dahan S. et al. (2020), 'Predicting Employment Notice Period with Machine Learning: Promises and Limitations', In: *McGill Law Journal/Revue de droit de McGill* 65.4, pp. 711–753.

66 Bui T. H. and Nguyen V. P. (2022), 'The impact of artificial intelligence and digital economy on vietnam's legal system', In: *International Journal for the Semiotics of Law-Revue internationale de S´emiotique ju- ridique*, pp. 1–21.

67 Cameron A., Pham T., and Atherton J. (2018), 'Vietnam Today–first report of the Vietnam's Future Digital Economy Project', In: *Canberra: CSIRO.*

68 Chao P.-J. et al. (2021), 'Knowledge of and competence in artificial intelligence: Perspectives of Vietnamese digital-native students', In: *IEEE Access* 9, pp. 75751–75760; Tran D. M. et al. (2022), 'Digital Health Policy and Programs for Hospital Care in Vietnam: Scoping Review', In: *Journal of medical Internet re- search* 24.2; Ablameyko M. et al. (2022), 'Legal aspects of e-commerce cooperation be- tween Eurasian economic union and Vietnam', In: *Journal of Science and Technology – Binh Duong University* 5.2.

69 Retrieved from: https://asia.nikkei.com/Economy/Vietnam-reveals-AI-strategy-as-new-leaders- enter-office2 [accessed on 17 December 2022].

70 Decision No. 127/QD-TTg of the Prime Minister, dated January 26, 2021, on the National Strategy on Research, Development and Application of AI to 2030.

71 *Ibid*.

education, agriculture, transportation, e-commerce... the AI platforms gradually bring new achievements.

The direction that the AI National Strategy takes is building data and computing infrastructure for AI research, development and application. The government also expressly acknowledged the importance of development of the AI ecosystem, such as: (i) Human resource development; (ii) Organizational infrastructure; (iii) Promoting research and application development; and (iv) Promoting the construction of incubation centers and attract investment for the development of AI businesses.

The Vietnamese Ministry of Justice (MOJ) has been mandated to develop a legal framework on AI. Particularly, it must research and design additional rules to regulate legal liabilities of AI-related subjects[72]. One of the central issues raised by the MOJ in the process of preparing to develop a proposal on regulations related to legal liabilities of entities related to AI is how and to which entity does a user of an AI system explain the AI decision-making process. This will have important implications for the determination of legal liability for the parties involved (if errors in the AI system causing damages to the client). Most Vietnamese experts recognize the importance of the "explainability" element in controlling the safety of AI systems.[73]

Machine learning research on explainability lacks a consensus of what an explanation actually is. Issues of explainability have become problematic due to the prominent view that the accuracy of an AI system trades off against its explainability. These issues often arise in discussions of medical AI systems, where reliance on artificial decision making in medical contexts could have serious consequences. The law therefore needs a clear and concise provision to ensure that "explainability" is appropriate, thereby helping to determine liability.

4. What is an "explanation" in the context of AI regulation

The question of what constitutes an "explanation" has been the subject of many different and sometimes subjective analyses. These include articulating "explainability" as:

- A combination of "intelligibility", "faithfulness", and "stability".[74]
- "The ability to extract knowledge about the relationships within the

72 *Ibid*.

73 MOJ & HCMUL, Minutes of the Conference 'Legal Responsibility in Artificial Intelligence Application: International Practices and Experiences for Vietnam', Ministry of Justice and Ho Chi Minh City University of Law, HCMC 11/12/ 2022.

74 Melis D. A. and Jaakkola T. (2018), 'Towards robust interpretabil- ity with self-explaining neural networks', In: *Advances in neural informa- tion processing systems* 31. Retrieved from: https://proceedings.neurips.cc/paper/2018/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf [accessed on 27 July 2022].

(input) data, which have been learned by a particular model".[75]

- "The extent to which a human can directly understand how a model operates and the causes of its decisions".[76]

- "The quest to imbue humans with a high level of understanding of how a model works and reaches decisions without trying to account for all the minutia of its calculations".[77]

Some boil the analysis down to defining the *how* question as "interpretability", and the *why* question as "explainability".[78] The existing literature makes a good deal of reference to various indicia or desiderata of interpretability and explainability, referencing concepts such as "fairness", "unbiasedness", "privacy", "reliability", "robustness", "causality", "usability", and "trust", or any combination of these, to substantiate a working definition, without any form of solidifying consensus emerging.[79]

Some argue that the inability to congeal a common approach is due to the fact that explainability related concepts are very subjective, often domain specific, and thus hard to universally formalize.[80] Further, depending on the context, different types of explanations may be required. For example, one might want to personally know the primary reason why a loan was declined by the bank, but a judge reviewing the matter might want to understand the role of all factors.[81] In the same vein, a patient might not be helped by a full causal explanation of a diagnosis but rather by a trustworthy account of understandable reasons expressed in clear and simple language.[82]

Among this sea of options, a recent paper takes a "back to core principles" approach to the concepts of "explanation", "interpretation"

75 Murdoch W. J. et al. (2019), 'Definitions, methods, and applications in in- terpretable machine learning', In: *Proceedings of the National Academy of Sciences* 116.44, pp. 22071–22080. Retrieved from: https://www.pnas.org/doi/pdf/10.1073/pnas.1900654116 [accessed on 27 July 2022].

76 Gilpin L. H et al. (2018), 'Explaining explanations: An overview of inter- pretability of machine learning', In: *IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pp. 80–89. Retrieved from: https://arxiv.org/pdf/1806.00069.pdf [accessed on 27 July 2022].

77 Goldstein A. et al. (2015), *supra* note 37.

78 Bhattacharya A. (2022), *Applied Machine Learning Explainability Tech- niques: Make ML models explainable and trustworthy for practical appli- cations using LIME, SHAP, and more*. Packt Publishing.

79 Finale D.-V. and Kim B. (2017), 'Towards a rigorous science of inter- pretable machine learning', In: *arXiv preprint arXiv:1702.08608*. Retrieved from: https://arxiv.org/pdf/1702.08608.pdf [accessed on 27 July 2022]; Marcinkevics R. and Vogt J. E. (2020), 'Interpretability and ex- plainability: A machine learning zoo mini-tour', In: *arXiv preprint arXiv:2012.01805*. Retrieved from: https://arxiv.org/pdf/2012.01805.pdf [accessed on 27 July 2022].

80 Graziani M. et al. (2022), 'A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences', In: *Artificial Intelligence Review*, pp. 132. Retrieved from: https://link.springer.com/content/pdf/10.1007/s10462-022-10256-8.pdf [accessed on 27 July 2022].

81 Carvalho D. V, Pereira E. M., and Cardoso J. S. (2019), 'Machine learning interpretability: A survey on methods and metrics', In: *Electronics* 8.8. p. 832. Retrieved from: https://pdfs.semanticscholar.org/232d/e9d0c1947ec757bb6644f27d203e488b1aaf.pdf [accessed on 27 July 2022].

82 Graziani M. et al. (2022), *supra* note *80*.

and "understanding", which I find compelling and persuasive, and which appears (to me) to be a solid foundation for interpreting these terms in any domain.[83] The core ideas, useful in exploring solutions to articulating the criteria for a proper "explanation" of a machine learning model, are summarized below.

4.1. Philosophical foundations

Philosophers have explored what constitutes an explanation, well before the era of com- puters, in order to capture the building blocks of rational evaluation of our world and inform sound decision making.[84]

The overall philosophical consensus, built over the last century, is that an explanation may be formulated using four generally accepted scientific explanation methodologies:

(i) *Deductive Nomological* (DN): A deduction (a "general to specific" inference) based on some law of nature of law-like proposition. For example, a ball falls because of Newton's Law of Universal Gravitation. This approach is useful for explaining *deterministic* phenomena.

(ii) *Inductive Statistical* (IS): An induction (a "specific to general" inference) of an individual event from a statistical law and empirical information about the event. For example, smoking causes cancer because many people who smoke get cancer. This approach is useful for explaining more "noisy" phenomena driven by probabilities.

(iii) *Causal Mechanical* (CM): Showing how the thing being explained fits into the *causal* structure of the world, by way of a known *causal process* or *causal interaction*, or both. For example, we can explain waves on a shoreline by the wake of a boat going by. This approach is useful when explaining interactions between physical structures in a spatio-temporal way.

(iv) *New Mechanist* (NM): Showing how some phenomenon arises from a *collection of entities and activities*. A successful explanation involves identifying the entities and activities that bring about a phenomenon with regularity and without gaps, and without missing entities or activities. For example, a particular chemical reaction occurring when combining certain compounds in certain states. This approach is useful when explaining more complex (usually biological or chemical) phenomena.

4.2. Scientific explanation

A scientific explanation, first and foremost, is thus a description of the *underlying theory and structure* in a manner consistent with our knowledge about the real world:[85]

83 Erasmus A., Brunet T. DP, and Fisher E. (2021), 'What is interpretability?', In: *Philosophy & Technology* 34.4, pp. 833–862. Retrieved from: https://link.springer.com/article/10.1007/s13347-020-00435-2 [accessed on 27 July 2022].

84 *Ibid*.

85 *Ibid*.

"What makes something an explanation is not some set of pragmatic conditions such as whether it produces understanding, or whether it is a "good" explanation because it satisfies some set of explanatory virtues, but rather whether it *accurately maps onto the world* via any one [or more] of the explanation processes."

Any purported explanation lacking such a mapping and not sufficiently rooted in one or more of these methods for rationally analyzing our world, is thus not a proper explanation, but rather an unproven conjecture.

This is an important nuance, particularly in machine learning where novel correlations between certain outputs and inputs may occur which may not readily fit one or more of the above methodologies of constructing an explanation. This problem is exacerbated when models are constructed simply by trying all kinds of features, but without any underlying feature engineering theory and/or understanding of the problem domain. A proper problem formulation *before* commencing to build a model is imperative.[86]

An explanation may be highly technical and difficult for the average person to understand. Transforming an explanation into understanding requires "interpretation".

4.3. Interpretation and understanding

Interpretation is something one does to an explanation to make it more understandable,[87] and may be performed among several different "vectors":

4.3.1. Total or Partial Interpretability

Total interpretability is an interpretation which "translates" a dimension of the problem formulation for every instance in the model. For example, in the case of an image processing model, being able to describe a particular image feature, material to classification, more precisely, such as, for example, "white tissue in this area" is relevant to a cancer diagnosis.

Partial interpretability is a modified interpretation which presents the explanation in a different way, for instance by way of a more succinct or simpler deduction.

4.3.2. Global or Local Interpretability

Global interpretability offers an interpretation which applies equally to all instances of the model.

Local interpretability involves an interpretation which applies to a subset of instances, and which is subject to identified constraints. For example, a model which finds the absolute value of a number[88] is interpreted differently depending on whether the number is positive or negative.

86 Lipton P. (2009), 'Understanding Without Explanation', In: *Scientific under- standing: Philosophical perspectives*.

87 Erasmus A., Brunet T. DP, and Fisher E. (2021), *supra* note 83.

88 The magnitude of a real number without regard to its sign. For example, the absolute values of -3 and 3 are both 3.

4.3.3. Interpretability by Approximation or Isomorphism

Approximate or Isomorphic Interpretability involves providing another, more understandable and *similar* explanation. The extent to which this type of interpretation is a proper proxy for the model sought to be explained is a key issue. Approximations and isomorphisms are by default not identical to the original model and thus subject to errors in interpretation.

4.4. Summary

The above interpretation vectors are easily recognizable in the various methodologies set out above.

Effecting "understanding" thus involves first ensuring that a phenomenon is con- stituted as an accurate mapping to the real world, and then seeking to interpret ("translate") the details of that mapping using various interpretability vectors and methodologies, to make it (more) understandable. See Figure 10.

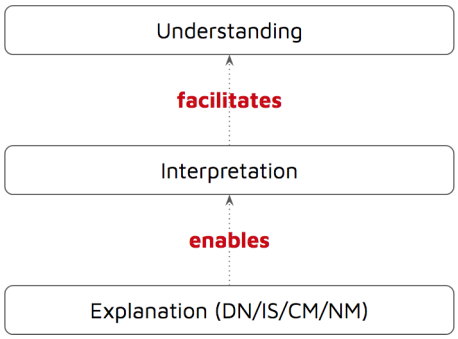


Figure 10: The process of achieving understanding by interpreting explanations

It is suggested that the *quality* of an interpretation may be assessed along three dimensions: *accuracy* (how precisely the interpretation describes a model's processes in "building" a prediction), *simplicity* (the understand- ability of the interpretation to a given stakeholder) and *relevance* (the extent to which the elements of the interpretation are in fact responsible for the model's specific output). [89] The ability to understand an explanation of course varies between individuals and is sensitive to the subjective features of each individual. Some people simply "get it" easier and faster - with less interpretation - than others.

4.5. Parallels with legal concepts

A number of parallels exist between machine learning process and explanation on the one hand, and various processes and procedures in the legal realm on the other.

89 Watson D. S. and Floridi L. (2021), 'The Explanation Game: a Formal Framework for Interpretable Machine Learning', In: *Ethics, Governance, and Policies in Artificial Intelligence*. Springer, pp. 185–219. Retrieved from: https://link.springer.com/content/pdf/10.1007/s11229-020-02629-9.pdf [accessed on 27 July 2022].

4.5.1. Reasons

It is trite to say that the process of "explanation" is foundational to the proper application of the law. It underlies the general requirement to provide reasons for judicial decisions.

As noted by the Supreme Court of Canada in *Canada (Minister of Citizenship and Immigration) v. Vavilov*[90] reasons are "the means by which the decision maker communicates the rationale for its decision" (para 84). A reasonable decision, in turn, is one that "is based on an internally coherent and rational chain of analysis" and in which the relevant reasons are "justified in relation to the facts and law that constrain the decision maker" (para 85).

Well-articulated reasons engender understanding, acceptance and trust in the proper functioning of the law, and therewith the structure of society.

4.5.2. Legal decision making

The scientific explanation process also has obvious parallels to the legal concept of *cau- sation* and the process of legal *inference*, both instrumental in legal decision making.

Administrative and judicial decision makers generally follow a predictable path, not unlike that of building a machine learning model:

- Identify the domain of the law and the legal question to be answered (define the prediction to be made).

- Identify and gather relevant evidence (select the right features).

- Exclude certain evidence which may be prejudicial (correct for bias).

- Weigh the evidence and match it against the current state of the law (use an algorithm to pursue alignment with the training dataset).

- Render a decision based on the best alignment of the evidence with the law and subject to a required standard of proof (generate a prediction given a certain threshold function).

- Provide reasons (enable understanding of the prediction).

The law is rich with detailed analyses in each of these dimensions, and the legal tools for completing each of them should thus be a valuable source of insights into how to deal with the explainability problem.

Among these, how we deal with expert evidence appears to be particularly apt.

4.5.3. Expert evidence

In Canadian and many other legal systems, expert evidence is used to "translate" factual observations into usually scientific conclusions by way of the learned ability of the expert witness to "interpret" the explanation of a particular phenomenon.

90 2019 SCC 65, [2019] 4 S.C.R. 653.

As noted in the *Science Manual for Canadian Judges*,[91] effective expert evidence is *understandable* (see also *R. v. J.-L.J.*):[92]

The principles and tools of science are increasingly invoked in legal disputes. In such cases, the trier of fact need not become a scientist nor resolve scientific debates, but he or she must be capable of developing an informed understanding of the science in question.

An expert's opinion must be firmly rooted in science ("mapped to the real world") and the underlying theory or technique tested and subjected to peer review and publication.[93]

The extent to which the science is subject to known or potential rates of error or is the subject of standards are additional considerations.[94]

Novel scientific theories, in particular, require "special scrutiny" with respect to verification and testing.[95] The *Science Manual for Canadian Judges* summarizes the point succinctly:[96]

Science advances by testing and retesting scientific hypotheses. As such, the number of tests of the hypothesis – that is, the absolute amount of evidence one way or the other – matters. Hypotheses that have been subjected to many independent tests, and come through with flying colours, are more likely to be true than those subjected to few tests. Similarly, hypotheses subjected to many independent tests and found to consistently fail are more likely to be false.

In addition, an expert's methodology and impartiality are material considerations. See *R. v. Abbey* [97] at paragraph 87:

Reliability concerns reach not only the subject matter of the evidence, but also the methodology used by the proposed expert in arriving at his or her opinion, the expert's expertise and the extent to which the expert is shown to be impartial and objective.

4.5.4. "Classification" experts

The *Mohan* and *R. v. J.-L.J.* decisions were two instances among numerous criminal cases in Canada where parties sought to bring expert evidence from psychologists and other human behaviour researchers, in order to establish that the defendant's profile was either inconsistent or consistent with the profile of persons guilty of the particular crime. Another notable example is *R. v. Malbœuf*,

91 Published by the National Judicial Institute in 2013, and last updated in 2018. Retrieved from: https://www.nji-inm.ca/index.cfm/publications/science-manual-for-canadian-judges/?langSwitch=en [accessed on 17 December 2022].

92 [2000] 2 S.C.R. 600. Retrieved from: https://www.canlii.org/en/ca/scc/doc/2000/2000scc51/2000scc51 [accessed on 17 December 2022].

93 *R. v. J.-L.J.*, supra, at para. 33.

94 *Ibid*.

95 *R. v. J.-L.J.*, supra. See also *R. v. Mohan* [1994] 2 SCR 9.

96 *Ibid*, Chapter 2.

97 2009ONCA624.Retrievedfrom:https://www.canlii.org/en/on/onca/doc/2009/2009onca624/2009onca624.html [accessed on 17 December 2022].

[1997] O.J. No. 1398 (QL) (C.A.), leave to appeal refused, [1998] 3 S.C.R. vii, a case where the Crown successfully introduced expert evidence that the defendant "demonstrated distinctive characteristics that would place him in the category of persons who would commit this type of crime" (para. 5).

The necessity for careful scrutiny of any underlying "model" is self-evident.

4.6. Summary

It is not a great leap to suggest that a machine learning system is an "artificial expert" "testifying" to a particular "prediction".

Applying the above constraints for expert evidence as well as the previously cited scientific explanation principles, it thus seems logical to articulate that effective "explainability" of a machine learning model used to make predictions which impact an individual should be driven by:

1. The soundness of the scientific hypothesis of the model and the extent to which it is capable of being made understandable.

2. The extent to which the hypothesis has been tested and verified.

3. The methodology used in developing the model.

4. The impartiality and objectivity - the absence of influential bias - of the model.

6. Some Steps Toward a Framework

Armed with the knowledge of how machine learning systems are constructed, the types of errors and biases they are susceptible to, the currently available tools for testing and interpreting a model, as well as the functional parallels between the processes of machine learning, legal decision making and the handling of expert evidence, it is possible to construct an initial framework to aid in answering the question of whether a particular model and its predictions are sufficiently "explainable".

This author would like to to propose a preliminary framework of important questions, the full assessment of which should enable the gaining of a level of comfort (or not), and thus aid in the determination of the amount of weight to ascribe to a particular model's prediction.

The role of the model: A model which does not materially contribute to an administrative or judicial decision may not require a detailed analysis. The identification of a model prediction's role may benefit from asking the following questions:

- What role does the prediction play in the decision? How much weight is attributed compared to other criteria?
- Are predictions advisory or integrated into an automated system without review?
- What are the risks/effects of a wrong decision?

If this assessment indicates a prediction is sufficiently material, then we can proceed further. In these questions, the model hypothesis - the basis upon which an answer to the question is being predicted is referred to as the "phenomenon".

	Questions to be addressed
The Data Team	- What are the qualifications of the team who built the model? - How diverse was the team? - What subject matter experts were used to gain an understanding of the phenomenon domain?
The Model Hypothesis	- Is there a summary of the domain knowledge used, or an existing body of knowledge with respect to the phenomenon? - Is the hypothesis or theory of the model explainable in a manner consistent with one or more of the scientific methodologies? - Is the statistical analysis of the data consistent with best practices?[98]
The Data	- What is the source of the data and who gathered it? - Does the data contain enough attributes to enabling modeling of the phenomenon? Are any important attributes missing? - How old is the data? If the data is older, has the data demographic changed since the data was gathered? - Is the data population articulated and defined? - What steps were taken to clean the data? - What was the approach for dealing with missing data for specific instances? - Was synthetic data used? - Given an understanding of the data population, are there known or potential biases in the data? If so, what was done to address such biases? - How was the data labeled? Were any appropriate standards in place? - How were the training and test datasets assembled?
The Features	- Are the features consistent with our knowledge about the phenomenon? - Does the combination of features capture the known "drivers" of the phenomenon? - Did any feature abstractions or combinations alter the "meaning" of any features? - Are there known or suspected causal relationships or interactions among the features? If so, how are they addressed in the engineering of the used features? - Does the choice of features enable any known "rules" of the phenomenon? - Are there any "risky" features (features which may enable a certain bias, prejudice or discriminatory effect)?

The *Model building methodology*: Accurate and proper evaluation of the Model Building Methodology of an AI system is very important. For the

98 See the *Science Manual for Canadian Judges*, page 76, section 5 et seq.

assessment to be complete and effective, it is suggested that the stakeholders will have to systematically answer the following questions:

	Questions to be addressed
Choice of Algorithm	- Why was this particular algorithm chosen? - Has the algorithm been successfully used in problems similar to the phenomenon? - Does the choice of algorithm align with the behaviour or nature of the phenomenon? - Were other algorithms tried and discarded, and if so, why?
Training Accuracy	- Do the chosen metrics for measuring accuracy align with the risk profile of the phenomenon? - How well did the model perform against the training dataset? - How well did the model perform against the test dataset? - Was an analysis of inaccurate predictions performed to inform model improvement?
Deployment	- Was the model architecture adjusted in order to be able to deploy it? if so, how? - Was the deployment tested for consistency with the trained model? - Has the model been re-trained since deployment?

Testing and Verification:

	Questions to be addressed
Global Consistency	- Are the model's predictions consistent with known statistical and empirical data about the phenomenon? - Have any global interpretability methods been implemented? - Are patterns in prediction logical, consistent and repeatable? - Are there any independent evaluations, audits or tests of the model?
Local Prediction Reliability	- Have local interpretability methodologies been employed? - Are local patterns logical, consistent and repeatable? - Are there any local "outliers" similar to the specific prediction, which are cause for concern?

Conclusion

Developing a legal framework for AI is a complex process as it requires legislators to have comprehensive and accurate understanding of the technical aspects of AI systems.

As noted, one approach, where an AI system is assessed as "high-impact," requires persons responsible for AI systems to:

- develop a risk mitigation plan;
- monitor those risk mitigation measures;
- to the extent the system is being used or being made available for

use, publish a plain-language description on a website describing (i) how the system is or is intended to be used, (ii) the types of content it is or is intended to generate and the decisions, recommendations or predictions it makes or is intended to make, (iii) the mitigation measures in place, and (iv) any other information as prescribed by regulation;

- to the extent the use of the system results or is likely to result in "material harm," notify, as soon as feasible, the state competent authorities.

The law will likely require persons/entities responsible for AI systems to assess these systems' potential to cause harm or produce biased outputs, develop mitigation plans to reduce or eliminate these risks, and publicly disclose when high-impact systems are being used, among other obligations. It should be noted that there are still conflicting views on the necessary content of the AI regulations in most countries.

The analysis in this paper could provide an accessible explanation of the process of machine learning as well as a rudimentary framework, rooted in scientific and legal principles, for the analysis of the explanation sufficiency of machine learning based prediction systems.

Further study and analysis are obviously required, particularly given the fact that machine learning is evolving and expanding at a rapid pace, with new and novel methodologies being developed both to build models, as well as how to go about "explaining" them. ●

References

[1] Ablameyko M. et al. (2022), 'Legal aspects of e-commerce cooperation between Eurasian economic union and Vietnam', In: *Journal of Science and Technology – Binh Duong University* 5.2

[2] Alarie B., Niblett A., and Yoon A. H. (2016), 'Using machine learning to predict outcomes in tax law', In: *Can. Bus. LJ 58,* p. 231

[3] Almuslim I. and Inkpen D., 'Legal Judgment Prediction for Canadian Appeal Cases', In: *2022 7th International Conference on Data Science and Machine Learning Applications (CDMA)*. IEEE. 2022, pp. 163– 168

[4] Arnold M. et al. (2019), 'FactSheets: Increasing trust in AI services through supplier's declarations of conformity', In: *IBM Journal of Research and Development* 63.4/5. Retrieved from: https://aifs360.mybluemix.net/introduction [accessed on 17 December 2022]

[5] Baracaldo N. et al. (2022), 'Towards an Accountable and Reproducible Federated Learning: A FactSheets Approach', In: *arXiv preprint arXiv:2202.12443*. Retrieved from: https://arxiv.org/pdf/2202.12443.pdf [accessed on 17 December 2022]

[6] Barbosa S. et al. (2016), 'Averaging gone wrong: Using time-aware analyses to better understand behavior', In: *Proceedings of the 25th International Conference on World Wide Web.*, pp. 829–841. Retrieved from: https://dl.acm.org/doi/pdf/10.1145/2872427.2883083 [accessed on 27 July 2022]

[7] Benk M. and Ferrario A. (2020), 'Explaining Interpretable Machine Learning: Theory, Methods and Applications', In: *Methods and Applications (December 11, 2020)*. Retrieved from: https://www.researchgate.net/profile/Andrea-Ferrario7/publication/348678581_Explaining_Interpretable_Machine_Learning_Theory_Methods_and_Applications/links/600ab71e299bf14088b21f03/Explaining-Interpretable-Machine-Learning-Theory-Methods-and-Applications.pdf [accessed on 25 October 2022]

[8] Bhandari A. (2020), 'Feature Scaling for Machine Learning: Understanding the Difference Between Normalization vs. Standardization'. Retrieved from: https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/ [accessed on 17 December 2022]

[9] Bhattacharya A. (2022), *Applied Machine Learning Explainability Techniques: Make ML models explainable and trustworthy for practical appli- cations using LIME, SHAP, and more*. Packt Publishing

[10] Borghesi A., Baldo F., and Milano M. (2020), 'Improving deep learning models via constraint-based domain knowledge: a brief survey', In: *arXiv preprint arXiv:2005.10691*. Retrieved from: https://arxiv.org/pdf/2005.10691.pdf [accessed on 25 July 2022]

[11] Bui T. H. and Nguyen V. P. (2022), 'The impact of artificial intelligence and digital economy on vietnam's legal system', In: *International Journal for the Semiotics of Law-Revue internationale de S´emiotique ju- ridique*, pp. 1–21

[12] Cameron A., Pham T., and Atherton J. (2018), 'Vietnam Today–first report of the Vietnam's Future Digital Economy Project', In: *Canberra: CSIRO*

[13] Carvalho D. V, Pereira E. M., and Cardoso J. S. (2019), 'Machine learning interpretability: A survey on methods and metrics', In: *Electronics* 8.8. p. 832. Retrieved from: https://pdfs.semanticscholar.org/232d/e9d0c1947ec757bb6644f27d203e488b1aaf.pdf [accessed on 27 July 2022]

[14] Chao P.-J. et al. (2021), 'Knowledge of and competence in artificial intelligence: Perspectives of Vietnamese digital-native students', In: *IEEE Access* 9, pp. 75751–75760

[15] Dandl S. et al. (2020), 'Multiobjective counterfactual explanations', In: *International Conference on Parallel Problem Solving from Nature*. Springer, pp. 448–469

[16] Dahan S. et al. (2020), 'Predicting Employment Notice Period with Machine Learning: Promises and Limitations', In: *McGill Law Journal/Revue de droit de McGill* 65.4, pp. 711–753

[17] Dong G. and Liu H. (2018), *Feature Engineering for Machine Learning and Data Analytics*, CRC Press

[18] Edwards L. and Michael V. (2017), 'Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for', *Duke L. & Tech. Rev*. 16, p. 18. Retrieved from: https://discovery.ucl.ac.uk/id/eprint/1574817/1/Veale_slavetothealgorithm_published.pdf [accessed on 27 July 2022]

[19] Elite Data Science (2022), 'Best Practices for Feature Engineering'. Retrieved from: https://elitedatascience.com/feature-engineering-best-practices [accessed 17 December 2022]

[20] Erasmus A., Brunet T. DP, and Fisher E. (2021), 'What is interpretability?', In: *Philosophy & Technology* 34.4, pp. 833–862. Retrieved from: https://link.springer.com/article/10.1007/s13347-020-00435-2 [accessed on 27 July 2022]

[21] Fadel S. (2022), *Explainable Machine Learning, Game Theory, and Shapley Values: A Technical Review*. Statistics Canada, Oct. 7. Retrieved from: https://www.statcan.gc.ca/en/data-science/network/explainable- learning [accessed on 10 July 2022]

[22] Finale D.-V. and Kim B. (2017), 'Towards a rigorous science of inter- pretable machine learning', In: *arXiv preprint arXiv:1702.08608*. Retrieved from: https://arxiv.org/pdf/1702.08608.pdf [accessed on 27 July 2022]

[23] Gaon A. and Stedman I. (2018), 'A call to action: Moving forward with the governance of artificial intelligence in Canada', In: *Alta. L. Rev*. 56. Retrieved from: http://albertalawreview.com/index.php/ALR/article/download/2547/2514 [accessed on 27 July 2022]

[24] Ghasemi M. et al. (2021), 'The Application of Machine Learning to a General Risk–Need Assessment Instrument in the Prediction of Criminal Recidi- vism', In: *Criminal Justice and Behavior* 48.4, pp. 518–538. Retrieved from: https://journals.sagepub.com/doi/pdf/10.1177/0093854820969753 [accessed on 27 July 2022]

[25] Gill J. K. (2020), 'Health insurance fraud detection', In:. https://era.library.ualberta.ca/items/e68678e1-1021-4e4c-8fa2-54455deb9fd0 [accessed on 27 July 2022]

[26] Gilpin L. H et al. (2018), 'Explaining explanations: An overview of inter- pretability of machine learning', In: *IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pp. 80–89. Retrieved from: https://arxiv.org/pdf/1806.00069.pdf [accessed on 27 July 2022]

[27] Goldstein A. et al. (2015), 'Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation', In: *journal of Computational and Graphical Statistics* 24.1, pp. 44–65. Retrieved from: https://arxiv.org/ pdf/1309.6392.pdf [accessed on 27 July 2022]

[28] Graziani M. et al. (2022), 'A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences', In: *Artificial Intelligence Review*, pp. 132. Retrieved from: https://link.springer.com/content/pdf/10.1007/s10462-022-10256-8.pdf [accessed on 27 July 2022]

[29] Kapoor S. and Narayanan A. (2022), 'Leakage and the Reproducibility Crisis in ML-based Science', In: *arXiv preprint arXiv:2207.07048*. Retrieved from: https://arxiv.org/pdf/2207.07048.pdf [accessed on 25 July 2022]

[30] Lehr D. and Ohm P. (2017), 'Playing with the data: what legal scholars should learn about machine learning', In: *UCDL Rev*. 51, p. 653. Retrieved from: https://lawreview.law.ucdavis.edu/issues/51/2/Symposium/51-2_Lehr_Ohm.pdf [accessed on 25 July 2022]

[31] Lipton P. (2009), 'Understanding Without Explanation', In: *Scientific understanding: Philosophical perspectives*

[32] Lokanan M. E. and Sharma K. (2022), 'Fraud prediction using machine learning: The case of investment advisors in Canada', In: *Machine Learning with Applications* 8. p. 100-269. Retrieved from: https://www.sciencedirect.com/science/article/pii/S2666827022000111/pdfft?isDTMRedir=true& download=true [accessed on 27 July 2022]

[33] Lundberg S. M. and Lee S.-I. (2017), 'A Unified Approach to Interpreting Model Predictions', In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc.. Retrieved from: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf. [accessed on 27 July 2022]

[34] Marcinkevics R. and Vogt J. E. (2020), 'Interpretability and explainability: A machine learning zoo mini-tour', In: *arXiv preprint arXiv:2012.01805*. Retrieved from: https://arxiv.org/pdf/2012.01805.pdf [accessed on 27 July 2022]

[35] Masis S. (2021), *Interpretable Machine Learning with Python: Learn to build interpretable high-performance models with hands-on real-world examples*, Packt Publishing Ltd

[36] Mehrabi N. et al. (2021), 'A survey on bias and fairness in machine learning', In: ACM Computing Surveys (CSUR) 54.6, pp. 1–35. Retrieved from: https://arxiv.org/pdf/1908.09635.pdf [accessed on 27 July 2022]

[37] Melis D. A. and Jaakkola T. (2018), 'Towards robust interpretabil- ity with self-explaining neural networks', In: *Advances in neural informa- tion processing systems* 31. Retrieved from: https://proceedings.neurips.cc/paper/2018/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf [accessed on 27 July 2022]

[38] MOJ & HCMUL, Minutes of the Conference 'Legal Responsibility in Artificial Intelligence Application: International Practices and Experiences for Vietnam', Ministry of Justice and Ho Chi Minh City University of Law, HCMC 11/12/ 2022

[39] Molnar C. (2020), *Interpretable machine learning*. Retrieved from: https://christophm.github.io/interpretable-ml-book/index.html [accessed on 27 July 2022]

[40] Murdoch W. J. et al. (2019), 'Definitions, methods, and applications in interpretable machine learning', In: *Proceedings of the National Academy of Sciences* 116.44, pp. 22071–22080. Retrieved from: https://www.pnas.org/doi/pdf/10.1073/pnas.1900654116 [accessed on 27 July 2022]

[41] Nalbandian L. (2021), 'Using Machine-Learning to Triage Canada's Temporary Resident Visa Applications', In: *Ryerson Centre for Immigration and Settlement (RCIS) and the CERC in Migration and Integration*. Retrieved from: https://www.torontomu.ca/content/dam/centre-for-immigration-and-settlementtmcis/publications/workingpapers/2021_9_Nalbandian_Lucia_Using_Machine_Learning_to_Triage_Canadas_Temporary_Resident_Visa_Applications.pdf [accessed on 27 July 2022]

[42] Northcutt C. G., Athalye A., and Mueller J. (2021), 'Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks', In: *arXiv preprint arXiv:2103.14749*. Retrieved from: https://arxiv.org/pdf/2103. 14749.pdf [accessed on 27 July 2022]

[43] Lapuschkin S. et al. (2019), 'Unmasking Clever Hans predictors and assessing what machines really learn', In: *Nature communications* 10.1, pp. 1–8. Retrieved from: https://www.nature.com/articles/s41467-019-08987-4 [accessed on 25 July 2022]

[44] Parentoni L. (2022), 'What should we reasonably expect from artificial intelligence?', In: *Publication pending at time of review*. Retrieved from: https://www.researchgate.net/profile/Leonardo-Parentoni/publication/361988480_What_should_we_reasonably_expect_from_artificial_intelligence/links/62d0198e953dfc1e93ff7c45/What-should-we-reasonably-expect-from-artificial-intelligence.pdf [accessed on 25 July 2022]

[45] Price II. and Nicholson W. (2017), 'Artificial intelligence in health care: applications and legal issues'. Retrieved from: https://repository.law.umich.edu/ cgi/viewcontent.cgi?article=2932&context=articles [accessed on 27 July 2022]

[46] PwC (2017), 'Sizing the prize: PwC's Global Artificial Intelligence Study: Exploiting the AI Revolution'. Retrieved from: https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html [accessed on 17 December 2022]

[47] Ribeiro M. T., Singh S., and Guestrin C. (2018), 'Anchors: High- precision model-agnostic explanations', In: *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32. Retrieved from: https://ojs.aaai.org/index.php/AAAI/article/view/11491/11350 [accessed on 27 July 2022]

[48] Ribeiro M. T., Singh S., and Guestrin C. (2016), 'Why should I trust you? - Explaining the predictions of any classifier', In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144. Retrieved from: https://arxiv.org/pdf/1602.04938.pdf [accessed on 27 July 2022]

[49] Roscher R. et al. (2020), 'Explainable machine learning for scientific insights and discoveries', In: *IEEE Access* 8, pp. 42200–42216. Retrieved from: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9007737 [accessed on 25 July 2022]

[50] Roth A. E. (1988), 'Introduction to the Shapley value', In: *The Shapley value*, pp. 1–27. Retrieved from: http://library.fa.ru/files/Roth2.pdf#page=9 [accessed on 27 July 2022]

[51] Rudin C. (2019), 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', In: *Nature Machine Intelligence* 1.5, pp. 206–215. Retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9122117/ [accessed on 27 July 2022]

[52] Rueden L. V. et al. (2019), 'Informed Machine Learning–A Taxonomy and Survey of Integrating Knowledge into Learning Systems', In: *arXiv preprint arXiv:1903.12394*. Retrieved from: https://arxiv.org/pdf/1903.12394.pdf [accessed on 25 October 2022]

[53] Russell S. and Norvig P. (2021), *Artificial intelligence: a modern approach, 4th Edition*. Pearson

[54] Scherer M. U. (2015), 'Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies', In: *Harv. JL & Tech*. 29, p. 353. Retrieved from: https://euro.ecom.cmu.edu/program/law/08-732/AI/Scherer.pdf [accessed on 25 October 2022]

[55] Suresh H. and Guttag J. (2021), 'A framework for understanding sources of harm throughout the machine learning life cycle', In: *Equity and access in algorithms, mechanisms, and optimization*, pp. 1–9. Retrieved from: https://dl.acm.org/doi/fullHtml/10.1145/3465416.3483305 [accessed on 27 July 2022]

[56] Tran D. M. et al. (2022), 'Digital Health Policy and Programs for Hospital Care in Vietnam: Scoping Review', In: *Journal of medical Internet re- search* 24.2

[57] Treasury Board of Canada (2021), *Directive on Automated Decision-Making*, Government of Canada, Apr. 21. Retrieved from: https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592 [accessed on 07/25/2022]

[58] Vincent J. (2016), 'Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day', Tayandyou (Twitter). Retrieved from: https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist [accessed on 17 December 2022]

[59] Watson D. S. and Floridi L. (2021), 'The Explanation Game: a Formal Framework for Interpretable Machine Learning', In: *Ethics, Governance, and Policies in Artificial Intelligence*. Springer, pp. 185–219. Retrieved from: https://link.springer.com/content/pdf/10.1007/s11229-020-02629-9.pdf [accessed on 27 July 2022]