Put it to the Test

Getting Serious About Explanation in XAI

Florian J. Boge*
& Axel Mosig†

*Institute for Philosophy & Political Science, TU Dortmund
†Bioinformatics Group, Ruhr University Bochum

UDNN IfPP PRODI LAMARR Institute for Machine Learning and Artificial Intelligence
Emmy Noether-Programm Deutsche Forschungsgemeinschaft DFG

Two Views of XAI

Rudin (2019)

'explanation' [...] refers to an understanding of how a model works, as opposed to an explanation of how the world works. [...]

Rudin (2019)

'explanation' [...] refers to an understanding of how a model works, as opposed to an explanation of how the world works. [...] a function that is too complicated for any human to comprehend

Rudin (2019)

‘explanation’ [...] refers to an understanding of how a model works, as opposed to an explanation of how the world works. [...] a function that is too complicated for any human to comprehend

Lipton (2018)

An interpretation may prove informative even without shedding light on a model’s inner workings. [...]

Rudin (2019)

'explanation' [...] refers to an understanding of how a model works, as opposed to an explanation of how the world works. [...] a function that is too complicated for any human to comprehend

Lipton (2018)

An interpretation may prove informative even without shedding light on a model's inner workings. [...] The real goal might be to explore the underlying structure of the data [...].

Rudin (2019)

'explanation' […] refers to an understanding of how a model works, as opposed to an explanation of how the world works. […] a function that is too complicated for any human to comprehend

Lipton (2018)

An interpretation may prove informative even without shedding light on a model's inner workings. […] The real goal might be to explore the underlying structure of the data […].

- narrow construal of XAI: efforts for explaining a model / its outputs

Rudin (2019)

'explanation' [...] refers to an understanding of how a model works, as opposed to an explanation of how the world works. [...] a function that is too complicated for any human to comprehend

Lipton (2018)

An interpretation may prove informative even without shedding light on a model's inner workings. [...] The real goal might be to explore the underlying structure of the data [...].

- narrow construal of XAI: efforts for explaining a model / its outputs
- broad construal of XAI: efforts for explaining things to do with a model / its outputs

Minds and Machines (2022) 32:43–75
https://doi.org/10.1007/s11023-021-09569-4

SI: MACHINE LEARNING: PREDICTION WITHOUT EXPLANATION

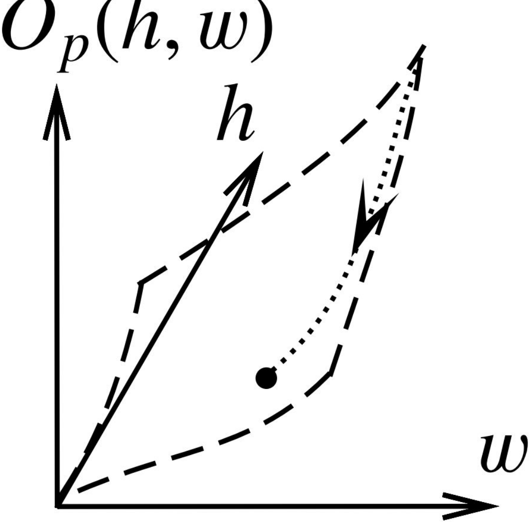
Two Dimensions of Opacity and the Deep Learning Predicament

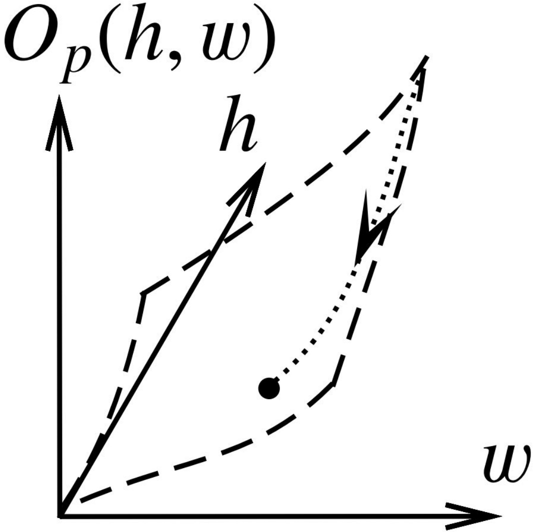
Florian J. Boge[1]

Received: 1 December 2020 / Accepted: 1 August 2021 / Published online: 3 September 2021

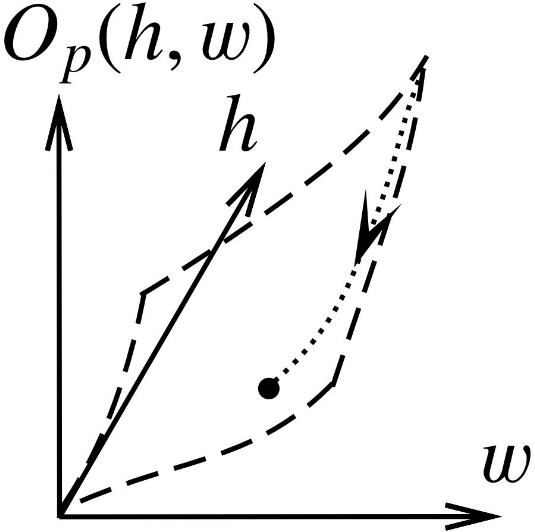
Abstract
Deep neural networks (DNNs) have become increasingly successful in applications from biology to cosmology to social science. Trained DNNs, moreover, correspond to models that ideally allow the prediction of new phenomena. Building in part on the literature on 'eXplainable AI' (XAI), I here argue that these models are instrumental in a sense that makes them non-explanatory, and that their automated generation is opaque in a unique way. This combination implies the possibility of an unprecedented gap between discovery and explanation: When unsupervised models are successfully used in exploratory contexts, scientists face a whole new challenge in forming the concepts required for understanding underlying mechanisms.

Keywords Machine learning · Opacity · Models · Explanation · Scientific understanding · Exploratory experimentation


$O_p(h, w)$
h
w



- understanding how a model works



- understanding how a model works
- understanding what a model learns

Opening the black box of Deep Neural Networks via Information

Ravid Schwartz-Ziv RAVID.ZIV@MAIL.HUJI.AC.IL
Edmond and Lilly Safra Center for Brain Sciences
The Hebrew University of Jerusalem
Jerusalem, 91904, Israel

Naftali Tishby* TISHBY@CS.HUJI.AC.IL
School of Engineering and Computer Science
and Edmond and Lilly Safra Center for Brain Sciences
The Hebrew University of Jerusalem
Jerusalem, 91904, Israel

Editor: ICRI-CI

Abstract

Despite their great success, there is still no comprehensive theoretical understanding of learning with Deep Neural Networks (DNNs) or their inner organization. Previous work [Tishby and Zaslavsky (2015)] proposed to analyze DNNs in the *Information Plane*; i.e., the plane of the Mutual Information values that each layer preserves on the input and output variables. They suggested that the goal of the network is to optimize the Information Bottleneck (IB) tradeoff between compression and prediction, successively, for each layer.

PHYSICAL REVIEW D **97,** 056009 (2018)

What is the machine learning?

Spencer Chang, Timothy Cohen, and Bryan Ostdiek
Institute of Theoretical Science, University of Oregon, Eugene, Oregon 97403, USA

(Received 19 October 2017; published 13 March 2018)

Applications of machine learning tools to problems of physical interest are often criticized for producing sensitivity at the expense of transparency. To address this concern, we explore a data planing procedure for identifying combinations of variables—aided by physical intuition—that can discriminate signal from background. Weights are introduced to smooth away the features in a given variable(s). New networks are then trained on this modified data. Observed decreases in sensitivity diagnose the variable's discriminating power. Planing also allows the investigation of the linear versus nonlinear nature of the boundaries between signal and background. We demonstrate the efficacy of this approach using a toy example, followed by an application to an idealized heavy resonance scenario at the Large Hadron Collider. By unpacking the information being utilized by these algorithms, this method puts in context what it means for a machine to learn.

DOI: 10.1103/PhysRevD.97.056009

I. INTRODUCTION

A common argument against using machine learning for physical applications is that they function as a black box: send in some data and out comes a number. While this kind of nonparametric estimation can be extremely useful, a physicist often wants to understand what aspect of the input of human-friendly variables that best characterize the data. While we are not inverting the network to find its functional form, we are providing a scheme for understanding classifiers.

For context, we acknowledge related studies within the growing machine learning for particle physics literature.

Sullivan (2022)

it is not the complexity or black box nature of a model that limits how much understanding the model provides.

Sullivan (2022)

it is not the complexity or black box nature of a model that limits how much understanding the model provides.

Sullivan (2022)

it is a lack of scientific and empirical evidence supporting the link that connects a model to the target phenomenon

Sullivan (2022)

it is not the complexity or black box nature of a model that limits how much understanding the model provides.

Sullivan (2022)

it is a lack of scientific and empirical evidence supporting the link that connects a model to the target phenomenon

- link may be severed due to unknown phenomena and "what-opacity"

Räz and Beisbart (2022)

researchers do not fully understand which features the DNN picks up on

Räz and Beisbart (2022)

researchers do not fully understand which features the DNN picks up on

Räz and Beisbart (2022)

understanding how this works means understanding how the model as such behaves in general [...] and not how the model relates to a particular [...] target.

Räz and Beisbart (2022)

researchers do not fully understand which features the DNN picks up on

Räz and Beisbart (2022)

understanding how this works means understanding how the model as such behaves in general [...] and not how the model relates to a particular [...] target.

- how- and what-opacity can come apart

Räz and Beisbart (2022)

researchers do not fully understand which features the DNN picks up on

Räz and Beisbart (2022)

understanding how this works means understanding how the model as such behaves in general [...] and not how the model relates to a particular [...] target.

- how- and what-opacity can come apart
- what-opacity does concern links to the particular target (what's being found out about it?)

What follows is most naturally understood in terms of

What follows is most naturally understood in terms of

- methods that fall under XAI in the broad, but not necessarily the narrow sense

What follows is most naturally understood in terms of

- methods that fall under XAI in the broad, but not necessarily the narrow sense
- what-opacity

What follows is most naturally understood in terms of

- methods that fall under XAI in the broad, but not necessarily the narrow sense
- what-opacity
- explanations of what the ML system finds in the data, in order to succeed

What follows is most naturally understood in terms of

- methods that fall under XAI in the broad, but not necessarily the narrow sense
- what-opacity
- explanations of what the ML system finds in the data, in order to succeed
- how to use this for fostering scientific progress

What is an “Explanation” in XAI?

‘Explanation’ in the philosophy of science:

‘Explanation’ in the philosophy of science:

- deductive-nomological (Hempel and Oppenheim, 1948)

‘Explanation’ in the philosophy of science:

- deductive-nomological (Hempel and Oppenheim, 1948)
- statistical relevance (Salmon, 1970)

‘Explanation’ in the philosophy of science:

- deductive-nomological (Hempel and Oppenheim, 1948)
- statistical relevance (Salmon, 1970)
- causal-mechanical (Dowe, 2000; Salmon, 1984)

‘Explanation’ in the philosophy of science:

- deductive-nomological (Hempel and Oppenheim, 1948)
- statistical relevance (Salmon, 1970)
- causal-mechanical (Dowe, 2000; Salmon, 1984)
- unificationist (Friedman, 1974)

‘Explanation’ in the philosophy of science:

- deductive-nomological (Hempel and Oppenheim, 1948)
- statistical relevance (Salmon, 1970)
- causal-mechanical (Dowe, 2000; Salmon, 1984)
- unificationist (Friedman, 1974)
- pragmatic (Van Fraassen, 1980)

‘Explanation’ in the philosophy of science:

- deductive-nomological (Hempel and Oppenheim, 1948)
- statistical relevance (Salmon, 1970)
- causal-mechanical (Dowe, 2000; Salmon, 1984)
- unificationist (Friedman, 1974)
- pragmatic (Van Fraassen, 1980)
- minimal model (Batterman and Rice, 2014)

‘Explanation’ in the philosophy of science:

- deductive-nomological (Hempel and Oppenheim, 1948)
- statistical relevance (Salmon, 1970)
- causal-mechanical (Dowe, 2000; Salmon, 1984)
- unificationist (Friedman, 1974)
- pragmatic (Van Fraassen, 1980)
- minimal model (Batterman and Rice, 2014)
- causal(-graph theoretic) (Pearl, 2009; Spirtes et al., 2000)

‘Explanation’ in the philosophy of science:

- deductive-nomological (Hempel and Oppenheim, 1948)
- statistical relevance (Salmon, 1970)
- causal-mechanical (Dowe, 2000; Salmon, 1984)
- unificationist (Friedman, 1974)
- pragmatic (Van Fraassen, 1980)
- minimal model (Batterman and Rice, 2014)
- causal(-graph theoretic) (Pearl, 2009; Spirtes et al., 2000)
- mathematical (Baker, 2005)

‘Explanation’ in the philosophy of science:

- deductive-nomological (Hempel and Oppenheim, 1948)
- statistical relevance (Salmon, 1970)
- causal-mechanical (Dowe, 2000; Salmon, 1984)
- unificationist (Friedman, 1974)
- pragmatic (Van Fraassen, 1980)
- minimal model (Batterman and Rice, 2014)
- causal(-graph theoretic) (Pearl, 2009; Spirtes et al., 2000)
- mathematical (Baker, 2005)
- functional (Cummins, 1975)

‘Explanation’ in the philosophy of science:

- deductive-nomological (Hempel and Oppenheim, 1948)
- statistical relevance (Salmon, 1970)
- causal-mechanical (Dowe, 2000; Salmon, 1984)
- unificationist (Friedman, 1974)
- pragmatic (Van Fraassen, 1980)
- minimal model (Batterman and Rice, 2014)
- causal(-graph theoretic) (Pearl, 2009; Spirtes et al., 2000)
- mathematical (Baker, 2005)
- functional (Cummins, 1975)
- simulacrum (Cartwright and McMullin, 1984)

‘Explanation’ in the philosophy of science:

- deductive-nomological (Hempel and Oppenheim, 1948)
- statistical relevance (Salmon, 1970)
- causal-mechanical (Dowe, 2000; Salmon, 1984)
- unificationist (Friedman, 1974)
- pragmatic (Van Fraassen, 1980)
- minimal model (Batterman and Rice, 2014)
- causal(-graph theoretic) (Pearl, 2009; Spirtes et al., 2000)
- mathematical (Baker, 2005)
- functional (Cummins, 1975)
- simulacrum (Cartwright and McMullin, 1984)
- how possibly (Dray, 1957)

‘Explanation’ in the philosophy of science:

- deductive-nomological (Hempel and Oppenheim, 1948)
- statistical relevance (Salmon, 1970)
- causal-mechanical (Dowe, 2000; Salmon, 1984)
- unificationist (Friedman, 1974)
- pragmatic (Van Fraassen, 1980)
- minimal model (Batterman and Rice, 2014)
- causal(-graph theoretic) (Pearl, 2009; Spirtes et al., 2000)
- mathematical (Baker, 2005)
- functional (Cummins, 1975)
- simulacrum (Cartwright and McMullin, 1984)
- how possibly (Dray, 1957)
- ...

(Methods for) 'explanation' in XAI:

(Methods for) ‘explanation’ in XAI:

- saliency maps (Simonyan et al., 2013)

(Methods for) ‘explanation’ in XAI:

- saliency maps (Simonyan et al., 2013)
- layer-wise relevance-propagation (Bach et al., 2015)

(Methods for) ‘explanation’ in XAI:

- saliency maps (Simonyan et al., 2013)
- layer-wise relevance-propagation (Bach et al., 2015)
- deep lift (Li et al., 2021)

(Methods for) ‘explanation’ in XAI:

- saliency maps (Simonyan et al., 2013)
- layer-wise relevance-propagation (Bach et al., 2015)
- deep lift (Li et al., 2021)
- integrated gradients (Sundararajan et al., 2017)

(Methods for) ‘explanation’ in XAI:

- saliency maps (Simonyan et al., 2013)
- layer-wise relevance-propagation (Bach et al., 2015)
- deep lift (Li et al., 2021)
- integrated gradients (Sundararajan et al., 2017)
- network dissection (Bau et al., 2017, 2018)

(Methods for) ‘explanation’ in XAI:

- saliency maps (Simonyan et al., 2013)
- layer-wise relevance-propagation (Bach et al., 2015)
- deep lift (Li et al., 2021)
- integrated gradients (Sundararajan et al., 2017)
- network dissection (Bau et al., 2017, 2018)
- information bottleneck (Schwartz-Ziv and Tishby, 2017)

(Methods for) 'explanation' in XAI:

- saliency maps (Simonyan et al., 2013)
- layer-wise relevance-propagation (Bach et al., 2015)
- deep lift (Li et al., 2021)
- integrated gradients (Sundararajan et al., 2017)
- network dissection (Bau et al., 2017, 2018)
- information bottleneck (Schwartz-Ziv and Tishby, 2017)
- counterfactual explanations (Wachter et al., 2017)

(Methods for) ‘explanation’ in XAI:

- saliency maps (Simonyan et al., 2013)
- layer-wise relevance-propagation (Bach et al., 2015)
- deep lift (Li et al., 2021)
- integrated gradients (Sundararajan et al., 2017)
- network dissection (Bau et al., 2017, 2018)
- information bottleneck (Schwartz-Ziv and Tishby, 2017)
- counterfactual explanations (Wachter et al., 2017)
- LIME (Ribeiro et al., 2016)

(Methods for) 'explanation' in XAI:

- saliency maps (Simonyan et al., 2013)
- layer-wise relevance-propagation (Bach et al., 2015)
- deep lift (Li et al., 2021)
- integrated gradients (Sundararajan et al., 2017)
- network dissection (Bau et al., 2017, 2018)
- information bottleneck (Schwartz-Ziv and Tishby, 2017)
- counterfactual explanations (Wachter et al., 2017)
- LIME (Ribeiro et al., 2016)
- concept-attribution vectors (Kim et al., 2018)

(Methods for) 'explanation' in XAI:

- saliency maps (Simonyan et al., 2013)
- layer-wise relevance-propagation (Bach et al., 2015)
- deep lift (Li et al., 2021)
- integrated gradients (Sundararajan et al., 2017)
- network dissection (Bau et al., 2017, 2018)
- information bottleneck (Schwartz-Ziv and Tishby, 2017)
- counterfactual explanations (Wachter et al., 2017)
- LIME (Ribeiro et al., 2016)
- concept-attribution vectors (Kim et al., 2018)
- deep dream (Mordvintsev et al., 2015)

(Methods for) ‘explanation’ in XAI:

- saliency maps (Simonyan et al., 2013)
- layer-wise relevance-propagation (Bach et al., 2015)
- deep lift (Li et al., 2021)
- integrated gradients (Sundararajan et al., 2017)
- network dissection (Bau et al., 2017, 2018)
- information bottleneck (Schwartz-Ziv and Tishby, 2017)
- counterfactual explanations (Wachter et al., 2017)
- LIME (Ribeiro et al., 2016)
- concept-attribution vectors (Kim et al., 2018)
- deep dream (Mordvintsev et al., 2015)
- data-planing (Chang et al., 2018)

(Methods for) ‘explanation’ in XAI:

- saliency maps (Simonyan et al., 2013)
- layer-wise relevance-propagation (Bach et al., 2015)
- deep lift (Li et al., 2021)
- integrated gradients (Sundararajan et al., 2017)
- network dissection (Bau et al., 2017, 2018)
- information bottleneck (Schwartz-Ziv and Tishby, 2017)
- counterfactual explanations (Wachter et al., 2017)
- LIME (Ribeiro et al., 2016)
- concept-attribution vectors (Kim et al., 2018)
- deep dream (Mordvintsev et al., 2015)
- data-planing (Chang et al., 2018)
- ...

(Methods for) ‘explanation’ in XAI:

- saliency maps (Simonyan et al., 2013)
- layer-wise relevance-propagation (Bach et al., 2015)
- deep lift (Li et al., 2021)
- integrated gradients (Sundararajan et al., 2017)
- network dissection (Bau et al., 2017, 2018)
- information bottleneck (Schwartz-Ziv and Tishby, 2017)
- counterfactual explanations (Wachter et al., 2017)
- LIME (Ribeiro et al., 2016)
- concept-attribution vectors (Kim et al., 2018)
- deep dream (Mordvintsev et al., 2015)
- data-planing (Chang et al., 2018)
- . . .

?

‘Explanation’ in the philosophy of science:

- deductive-nomological (Hempel and Oppenheim, 1948)
- statistical relevance (Salmon, 1970)
- causal-mechanical (Dowe, 2000; Salmon, 1984)
- unificationist (Friedman, 1974)
- pragmatic (Van Fraassen, 1980)
- minimal model (Batterman and Rice, 2014)
- causal(-graph theoretic) (Pearl, 2009; Spirtes et al., 2000)
- mathematical (Baker, 2005)
- functional (Cummins, 1975)
- simulacrum (Cartwright and McMullin, 1984)
- how possibly (Dray, 1957)
- . . .

Páez (2019)

explanations in the present stage of AI are incommensurable with the types of explanations discussed in the philosophy of science.

Páez (2019)

explanations in the present stage of AI are incommensurable with the types of explanations discussed in the philosophy of science.

Krishnan (2020)

There is a substantial literature within philosophy of science concerning the nature of explanation [...] largely orthogonal to the concerns of those seeking explicability or interpretability of ML algorithms.

Erasmus et al. (2021)

A DN explanation of how [a CNN] assesses an input image involves listing the weights attached to each and every node and the informational routes

Erasmus et al. (2021)

A DN explanation of how [a CNN] assesses an input image involves listing the weights attached to each and every node and the informational routes

Erasmus et al. (2021)

explaining [...] how the weights of all relevant nodes and edges produced the output value, along with the law that an output is assigned to the most probable class [...] which includes the set of input values assigned to [image] I, and the output value c.

Erasmus et al. (2021)

A DN explanation of how [a CNN] assesses an input image involves listing the weights attached to each and every node and the informational routes

Erasmus et al. (2021)

explaining [...] how the weights of all relevant nodes and edges produced the output value, along with the law that an output is assigned to the most probable class [...] which includes the set of input values assigned to [image] *I*, and the output value *c*.

- very sketchy

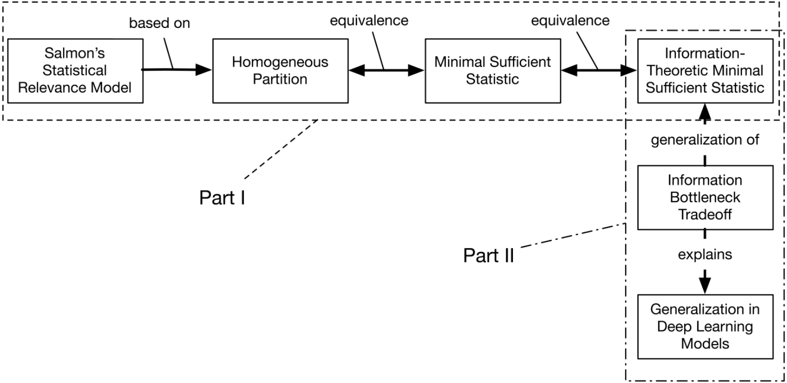
Erasmus et al. (2021)

A DN explanation of how [a CNN] assesses an input image involves listing the weights attached to each and every node and the informational routes

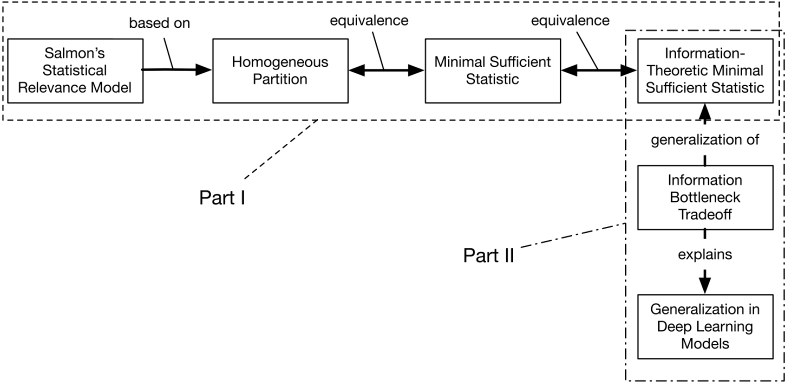
Erasmus et al. (2021)

explaining [...] how the weights of all relevant nodes and edges produced the output value, along with the law that an output is assigned to the most probable class [...] which includes the set of input values assigned to [image] I, and the output value c.

- very sketchy
- not aligned with actual XAI methods

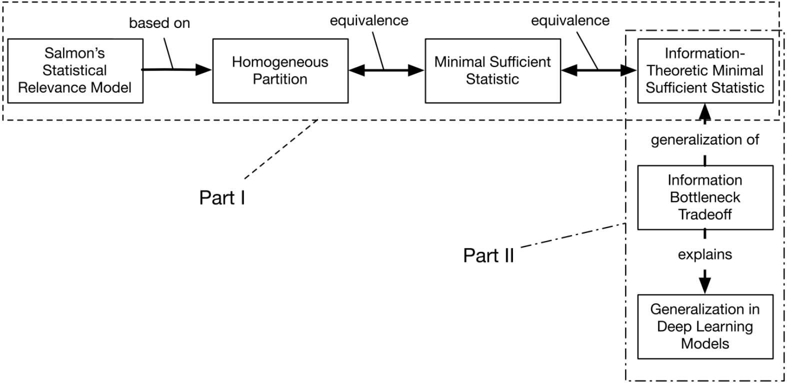

based on
equivalence
equivalence
Salmon's Statistical Relevance Model
Homogeneous Partition
Minimal Sufficient Statistic
Information-Theoretic Minimal Sufficient Statistic
generalization of
Information Bottleneck Tradeoff
explains
Generalization in Deep Learning Models
Part I
Part II

(Räz, 2022)



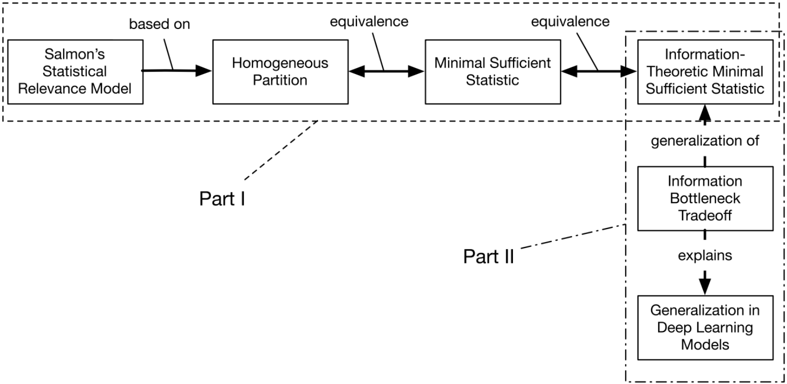
(Räz, 2022)

- rigorous



(Räz, 2022)

- rigorous
- aligned with XAI



(Räz, 2022)

- rigorous
- aligned with XAI
- just one example

Buijsman (2022)

counterfactuals are presented but without overarching generalizations [...] doesn't truly explain the functioning of an algorithm

Buijsman (2022)

counterfactuals are presented but without overarching generalizations [...] doesn't truly explain the functioning of an algorithm

Baron (2023)

basic causal certification [...] a guarantee that the information provided to users is always genuine causal information.

Buijsman (2022)

counterfactuals are presented but without overarching generalizations [...] doesn’t truly explain the functioning of an algorithm

Baron (2023)

basic causal certification [...] a guarantee that the information provided to users is always genuine causal information.

proof of concept

So What?

Douglas (2009)

A scientific explanation will be expected to produce new, generally successful predictions. An explanation that is not in fact used to generate predictions, or whose predictions quickly and obviously fail, would be scientifically suspect. An example of an explanation that fails to meet these criteria is any “just-so” story.

Douglas (2009)

A scientific explanation will be expected to produce new, generally successful predictions. An explanation that is not in fact used to generate predictions, or whose predictions quickly and obviously fail, would be scientifically suspect. An example of an explanation that fails to meet these criteria is any "just-so" story.

- many evolutionary stories are irrefutable (Gould and Lewontin, 1979)

Douglas (2009)

A scientific explanation will be expected to produce new, generally successful predictions. An explanation that is not in fact used to generate predictions, or whose predictions quickly and obviously fail, would be scientifically suspect. An example of an explanation that fails to meet these criteria is any "just-so" story.

- many evolutionary stories are irrefutable (Gould and Lewontin, 1979)
- some optimality models can be shown to make testable predictions (Orzack and Sober, 1994)

Douglas (2009)

A scientific explanation will be expected to produce new, generally successful predictions. An explanation that is not in fact used to generate predictions, or whose predictions quickly and obviously fail, would be scientifically suspect. An example of an explanation that fails to meet these criteria is any "just-so" story.

- many evolutionary stories are irrefutable (Gould and Lewontin, 1979)
- some optimality models can be shown to make testable predictions (Orzack and Sober, 1994)
- scientific explanations stick their neck out

Douglas (2009)

A scientific explanation will be expected to produce new, generally successful predictions. An explanation that is not in fact used to generate predictions, or whose predictions quickly and obviously fail, would be scientifically suspect. An example of an explanation that fails to meet these criteria is any “just-so” story.

- many evolutionary stories are irrefutable (Gould and Lewontin, 1979)
- some optimality models can be shown to make testable predictions (Orzack and Sober, 1994)
- scientific explanations stick their neck out
- why apply less rigorous standards in XAI?

- taking the X in XAI seriously enables applying rigorous standards

- taking the X in XAI seriously enables applying rigorous standards
- especially: testability

- taking the X in XAI seriously enables applying rigorous standards
- especially: testability
- makes sense if we want to trust ML models

- taking the X in XAI seriously enables applying rigorous standards
- especially: testability
- makes sense if we want to trust ML models
- … and if we want the explanation to relate to reality, not just the model (broad construal / what-opacity)

Can we Really Test Hypotheses?

Popper (1959)

A theory is to be called [...] ‘falsifiable’ if it divides the class of all [conceivable singular statements of fact] unambiguously into [...] those [...] with which it is inconsistent [...] and [...] those [...] which it does not contradict

Popper (1959)

A theory is to be called […] ‘falsifiable’ if it divides the class of all [conceivable singular statements of fact] unambiguously into […] those […] with which it is inconsistent […] and […] those […] which it does not contradict

Popper (1959)

We shall take [a theory] as falsified only if […] a low-level empirical hypothesis which describes […] a reproducible effect which refutes the theory […] is proposed and corroborated

Popper (1959)

A theory is to be called [...] ‘falsifiable’ if it divides the class of all [conceivable singular statements of fact] unambiguously into [...] those [...] with which it is inconsistent [...] and [...] those [...] which it does not contradict

Popper (1959)

We shall take [a theory] as falsified only if [...] a low-level empirical hypothesis which describes [...] a reproducible effect which refutes the theory [...] is proposed and corroborated

- insufficient consideration of holism (Duhem, 1914; Quine, 1951)

Popper (1959)

A theory is to be called […] ‘falsifiable’ if it divides the class of all [conceivable singular statements of fact] unambiguously into […] those […] with which it is inconsistent […] and […] those […] which it does not contradict

Popper (1959)

We shall take [a theory] as falsified only if […] a low-level empirical hypothesis which describes […] a reproducible effect which refutes the theory […] is proposed and corroborated

- insufficient consideration of holism (Duhem, 1914; Quine, 1951)
- theory-ladenness of the falsifying hypothesis?

Lakatos (1970)

continuity evolves from a genuine research programme [which] consists of methodological rules

Lakatos (1970)

continuity evolves from a genuine research programme [which] consists of methodological rules

Lakatos (1970)

The negative heuristic specifies the 'hard core' of the programme [...]; the positive heuristic consists of a partially articulated set of suggestions or hints on [...] how to modify, sophisticate, the 'refutable' protective belt.

Lakatos (1970)

continuity evolves from a genuine research programme [which] consists of methodological rules

Lakatos (1970)

The negative heuristic specifies the 'hard core' of the programme [...]; the positive heuristic consists of a partially articulated set of suggestions or hints on [...] how to modify, sophisticate, the 'refutable' protective belt.

- what about statistical hypotheses

Lakatos (1970)

continuity evolves from a genuine research programme [which] consists of methodological rules

Lakatos (1970)

The negative heuristic specifies the 'hard core' of the programme [...]; the positive heuristic consists of a partially articulated set of suggestions or hints on [...] how to modify, sophisticate, the 'refutable' protective belt.

- what about statistical hypotheses
- measurement (almost?) inevitably introduces probabilities...

Gillies (2000)

falsifying rule for probability statements [...] if the value obtained for X is in the tails of the distribution, this should be regarded as falsifying H

Gillies (2000)

falsifying rule for probability statements [...] if the value obtained for X is in the tails of the distribution, this should be regarded as falsifying H

Gillies (2000)

broad agreement between the proposed falsifying rule and the practice of statistical testing

Gillies (2000)

falsifying rule for probability statements [...] if the value obtained for X is in the tails of the distribution, this should be regarded as falsifying H

Gillies (2000)

broad agreement between the proposed falsifying rule and the practice of statistical testing

- non-reproducible effects in HEP: $3\sigma \mapsto 5\sigma$

Gillies (2000)

falsifying rule for probability statements [...] if the value obtained for X is in the tails of the distribution, this should be regarded as falsifying H

Gillies (2000)

broad agreement between the proposed falsifying rule and the practice of statistical testing

- non-reproducible effects in HEP: $3\sigma \mapsto 5\sigma$
- in psychology: re-assessment of replication

Gillies (2000)

falsifying rule for probability statements [...] if the value obtained for X is in the tails of the distribution, this should be regarded as falsifying H

Gillies (2000)

broad agreement between the proposed falsifying rule and the practice of statistical testing

- non-reproducible effects in HEP: $3\sigma \mapsto 5\sigma$
- in psychology: re-assessment of replication
- what's the overarching standard?

Popper (1959)

We must clearly distinguish between falsifiability and falsification. [...] falsifiability [...] as a criterion for the empirical character of a system of statements.

Popper (1959)

We must clearly distinguish between falsifiability and falsification. [...] falsifiability [...] as a criterion for the empirical character of a system of statements.

Genin (2022)

a variety of different methodologies of falsification [...] give rise to exactly the same collection of falsifiable hypotheses.

Popper (1959)

We must clearly distinguish between falsifiability and falsification. [...] falsifiability [...] as a criterion for the empirical character of a system of statements.

Genin (2022)

a variety of different methodologies of falsification [...] give rise to exactly the same collection of falsifiable hypotheses.

Genin (2022)

statistically falsifiable propositions [...] are exactly the closed sets in the weak topology

Popper (1959)

We must clearly distinguish between falsifiability and falsification. [...] falsifiability [...] as a criterion for the empirical character of a system of statements.

Genin (2022)

a variety of different methodologies of falsification [...] give rise to exactly the same collection of falsifiable hypotheses.

Genin (2022)

statistically falsifiable propositions [...] are exactly the closed sets in the weak topology

- value attributions in measurements $m = \lambda \pm \delta$ correspond to open sets, $m \in]\lambda - \delta, \lambda + \delta[$

Popper (1959)

We must clearly distinguish between falsifiability and falsification. […] falsifiability […] as a criterion for the empirical character of a system of statements.

Genin (2022)

a variety of different methodologies of falsification […] give rise to exactly the same collection of falsifiable hypotheses.

Genin (2022)

statistically falsifiable propositions […] are exactly the closed sets in the weak topology

- value attributions in measurements $m = \lambda \pm \delta$ correspond to open sets, $m \in]\lambda - \delta, \lambda + \delta[$
- lots of scientific claims aren't falsifiable

testability lost?

testability lost?

- unlike ‘god exists’, open interval can be turned into closed one

testability lost?

- unlike ‘god exists’, open interval can be turned into closed one
- varying standards of de facto falsification may be reasonable:

testability lost?

- unlike ‘god exists’, open interval can be turned into closed one
- varying standards of de facto falsification may be reasonable:
  - large enough amounts of data make effects more likely $\rightsquigarrow$ higher standards required (HEP)

testability lost?

- unlike ‘god exists’, open interval can be turned into closed one
- varying standards of de facto falsification may be reasonable:
    - large enough amounts of data make effects more likely $\rightsquigarrow$ higher standards required (HEP)
    - framing effects etc. introduce different subtleties $\rightsquigarrow$ careful consideration of reproduction indicated (psychology)

testability lost?

- unlike 'god exists', open interval can be turned into closed one
- varying standards of de facto falsification may be reasonable:
  - large enough amounts of data make effects more likely $\rightsquigarrow$ higher standards required (HEP)
  - framing effects etc. introduce different subtleties $\rightsquigarrow$ careful consideration of reproduction indicated (psychology)
  - in general: external values should influence our willingness to reject hypotheses (Douglas, 2000)

testability lost?

testability lost?

- testability as an incremental process of (dis-)confirmation (e.g. Sprenger and Hartmann, 2019)

testability lost?

- testability as an incremental process of (dis-)confirmation (e.g. Sprenger and Hartmann, 2019)
- can happen precisely for the reason that one hypothesis explains the data better than another (Schupbach, 2016)

The FXAI framework

Medical Image Analysis 82 (2022) 102594

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

A framework for falsifiable explanations of machine learning models with an application in computational pathology

David Schuhmacher [a,b,1], Stephanie Schörner [a,c,1], Claus Küpper [a,c], Frederik Großerueschkamp [a,c], Carlo Sternemann [a,d], Celine Lugnier [a,e], Anna-Lena Kraeft [a,e], Hendrik Jütte [a,d], Andrea Tannapfel [a,d], Anke Reinacher-Schick [a,e], Klaus Gerwert [a,c], Axel Mosig [a,b,*]

[a] Ruhr-University Bochum, Center for Protein Diagnostics, Bochum, 44801, Germany
[b] Ruhr-University Bochum, Faculty of Biology and Biotechnology, Bioinformatics Group, 44801 Bochum, Germany
[c] Ruhr-University Bochum, Faculty of Biology and Biotechnology, Department of Biophysics, 44801 Bochum, Germany
[d] Institute of Pathology, Ruhr-University Bochum, 44789 Bochum, Germany
[e] Department of Hematology, Oncology and Palliative Care, Ruhr-University Bochum, St. Josef-Hospital, 44791 Bochum, Germany

A R T I C L E I N F O

MSC:
62M45
68T27

A B S T R A C T

In recent years, deep learning has been the key driver of breakthrough developments in computational pathology and other image based approaches that support medical diagnosis and treatment. The underlying neural networks as inherent black boxes lack transparency and are often accompanied by approaches to

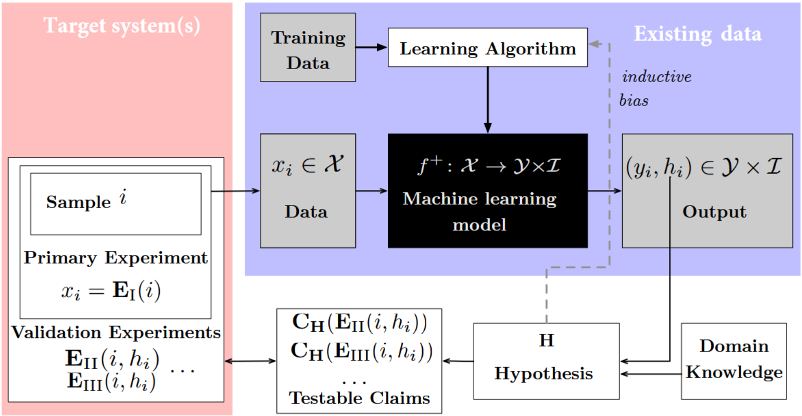


Image credit: Axel Mosig @ BioInf / RUB

Interpretation first

Interpretation first

Erasmus et al. (2021)

interpretation is something that one does to an explanation to make it more understandable.

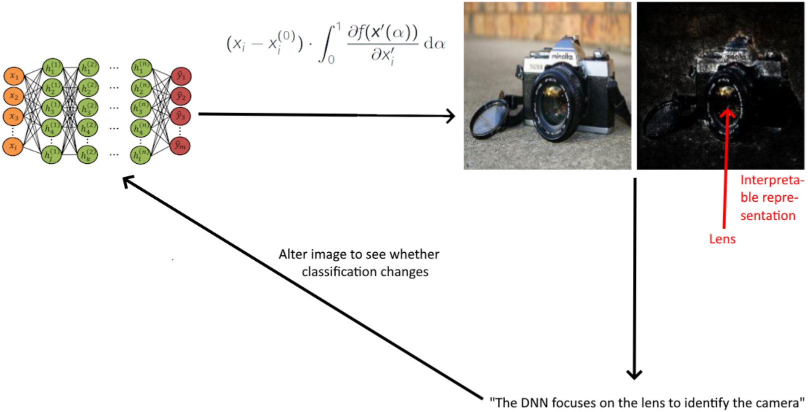
Interpretation first

Erasmus et al. (2021)

interpretation is something that one does to an explanation to make it more understandable.

Ribeiro et al. (2016)

explanations [...] must be interpretable, i.e., provide qualitative understanding between the input variables and the response

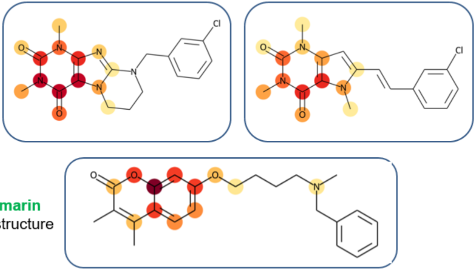

$(x_i - x_i^{(0)}) \cdot \int_0^1 \frac{\partial f(x'(\alpha))}{\partial x_i'} \, \mathrm{d}\alpha$
Interpreta-
ble repre-
sentation
Lens
Alter image to see whether
classification changes
"The DNN focuses on the lens to identify the camera"

Case study

Atom-based mapping of most important **features** on DT-CPDs

TP1
Caffeine
substruct.

TP2
Coumarin
substructure



Color code: number of matching features per atom 1 2 3 4 5 6 7

Image credit: Jürgen Bajorath @ Lamarr / U Bonn

Atom-based mapping of most important **features** on DT-CPDs

TP1
Caffeine
substruct.

TP2
Coumarin
substructure

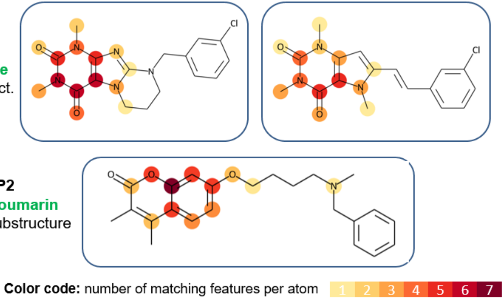


Image credit: Jürgen Bajorath @ Lamarr / U Bonn

- absence / presence of certain features relevant for single / dual target prediction

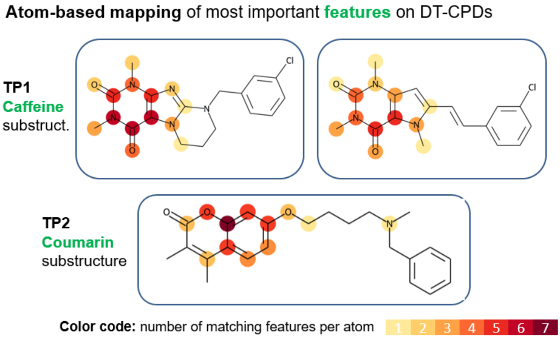


Image credit: Jürgen Bajorath @ Lamarr / U Bonn

- absence/ presence of certain features relevant for single / dual target prediction
- ‘coherent substructures’ as interpretable representations

Atom-based mapping of most important **features** on DT-CPDs

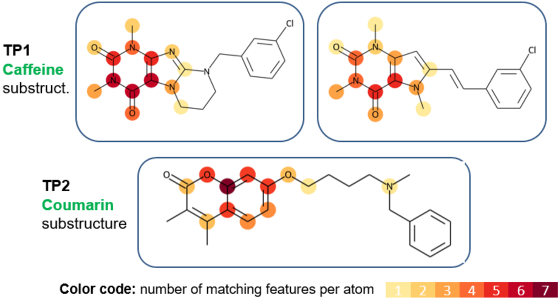


Image credit: Jürgen Bajorath @ Lamarr / U Bonn

- absence / presence of certain features relevant for single / dual target prediction
- 'coherent substructures' as interpretable representations
- explanatory hypothesis: caffeine / coumarin causally responsible for dual-target behavior

Atom-based mapping of most important **features** on DT-CPDs

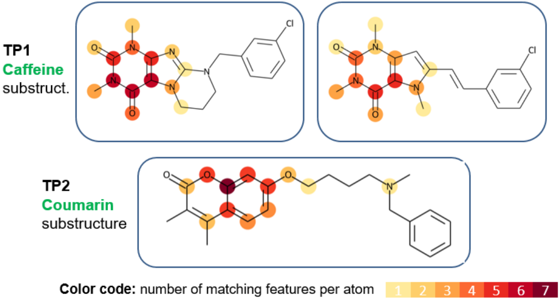


Image credit: Jürgen Bajorath @ Lamarr / U Bonn

- absence/ presence of certain features relevant for single / dual target prediction
- 'coherent substructures' as interpretable representations
- explanatory hypothesis: caffeine / coumarin causally responsible for dual-target behavior
- requires experimental validation / further testing

Atom-based mapping of most important **features** on DT-CPDs

TP1
Caffeine
substruct.

TP2
Coumarin
substructure

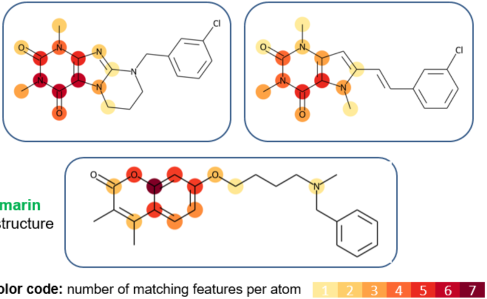


Image credit: Jürgen Bajorath @ Lamarr / U Bonn

- absence / presence of certain features relevant for single / dual target prediction
- 'coherent substructures' as interpretable representations
- explanatory hypothesis: caffeine / coumarin causally responsible for dual-target behavior
- requires experimental validation / further testing
- confirmation through literature search

Conclusion

- XAI can serve the purpose of understanding AI and the world

- XAI can serve the purpose of understanding AI and the world
- if we want to make headway, we should treat ‘AI explanations’ with scientific rigor

- XAI can serve the purpose of understanding AI and the world
- if we want to make headway, we should treat ‘AI explanations’ with scientific rigor
- for that, they should be probed for predictivity and empirically tested

- XAI can serve the purpose of understanding AI and the world
- if we want to make headway, we should treat ‘AI explanations’ with scientific rigor
- for that, they should be probed for predictivity and empirically tested
- as a matter of fact, this has lead to progress in actual research

Thank You!

References

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Baker, A. (2005). Are there genuine mathematical explanations of physical phenomena? *Mind*, 114(454):223–238.

Baron, S. (2023). Explainable ai and causal understanding: Counterfactual approaches considered. *Minds and Machines*, 33(2):347–377.

Batterman, R. W. and Rice, C. C. (2014). Minimal model explanations. *Philosophy of Science*, 81(3):349–376.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. *arXiv preprint arXiv:1704.05796*.

Bau, D., Zhu, J.-Y., Strobelt, H., Zhou, B., Tenenbaum, J. B., Freeman, W. T., and Torralba, A. (2018). Gan dissection: Visualizing and understanding generative adversarial networks. *arXiv preprint arXiv:1811.10597*.

Buijsman, S. (2022). Defining explanation and explanatory depth in xai. *Minds and Machines*, 32(3):563–584.

Cartwright, N. and McMullin, E. (1984). How the laws of physics lie.

Chang, S., Cohen, T., and Ostdiek, B. (2018). What is the machine learning? *Physical Review D*, 97(5):6.

Cummins, R. E. (1975). Functional analysis. *Journal of Philosophy*, 72(November):741–64.

Douglas, H. (2000). Inductive risk and values in science. *Philosophy of Science*, 67(4):559–579.

Douglas, H. E. (2009). Reintroducing prediction to explanation. *Philosophy of Science*, 76(4):444–463.

Dowe, P. (2000). *Physical Causation*. Cambridge: Cambridge University Press.

Dray, W. H. (1957). *Laws and Explanation in History*. Greenwood Press.

Duhem, P. (1954[1914]). *The Aim and Structure of Physical Theory*. Princeton University Press, second edition. translated by Philip P. Wiener.

Erasmus, A., Brunet, T. D., and Fisher, E. (2021). What is interpretability? *Philosophy & Technology*, 34(4):833–862.

Friedman, M. (1974). Explanation and scientific understanding. *The Journal of Philosophy*, 71(1):5–19.

Genin, K. (2022). On falsifiable statistical hypotheses. *Philosophies*, 7(2).

Gillies, D. (2000). *Philosophical Theories of Probability*. Routledge.

Gould, S. J. and Lewontin, R. C. (1979). The spandrels of san marco and the panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 205(1161):581–598.

Hempel, C. G. and Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, 15(2):135–175.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., and sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, pages 2668–2677. PMLR. https://proceedings.mlr.press/v80/kim18d.html.

Krishnan, M. (2020). Against interpretability: a critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, 33(3):487–502.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In Lakatos, I. and Musgrave, A., editors, *Criticism and the Growth of Knowledge*, page pp. 91–196. Cambridge: Cambridge University Press.

Li, J., Zhang, C., Zhou, J. T., Fu, H., Xia, S., and Hu, Q. (2021). Deep-lift: Deep label-specific feature learning for image annotation. *IEEE transactions on Cybernetics*, 52(8):7732–7741.

Lipton, Z. (2018). The mythos of model interpretability. *Queue*, 16:31–57. https://doi.org/10.1145/3236386.3241340.

Mordvintsev, A., Olah, C., and Tyka, M. (2015). Inceptionism: Going deeper into neural networks. *Google research blog*, 20(14):5.

Orzack, S. H. and Sober, E. (1994). Optimality models and the test of adaptationism. *The American Naturalist*, 143(3):361–380.

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (xai). *Minds and Machines*, 29(3):441–459.

Pearl, J. (2009). *Causality*. Cambridge university press.

Popper, K. (1959). *The Logic of Scientific Discovery*. London: Routledge.

Quine, W. (1951). Main trends in recent philosophy: Two dogmas of empiricism. *The Philosophical Review*, 60(1):20–43.

Räz, T. (2022). Understanding deep learning with statistical relevance. *Philosophy of Science*, 89(1):20–41.

Räz, T. and Beisbart, C. (2022). The importance of understanding deep learning. *Erkenntnis*. https://doi.org/10.1007/s10670-022-00605-y.

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*, pages 1135–1144.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Salmon, W. C. (1970). Statistical explanation. In Colodny, R., editor, *The Nature and Function of Scientific Theories*, pages 173–231. University of Pittsburgh Press.

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.

Schupbach, J. N. (2016). Robustness analysis as explanatory reasoning. *The British Journal for the Philosophy of Science*, 69(1):275–300.

Schwartz-Ziv, R. and Tishby, N. (2017). Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, prediction, and search*. MIT press.

Sprenger, J. and Hartmann, S. (2019). *Bayesian Philosophy of Science*. OUP Oxford.

Sullivan, E. (2022). Understanding from machine learning models. *The British Journal for the Philosophy of Science*, 73(1):109–133.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328. PMLR. https://proceedings.mlr.press/v70/sundararajan17a.html.

Van Fraassen, B. (1980). *The Scientific Image*. Clarendon Library of Logic and Philosophy. Clarendon Press.

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841.