The Exploratory Role of Explainable Artificial Intelligence

August 2020

Carlos Zednik
carlos.zednik@ovgu.de

Hannes Boelsen
hannes.boelsen@ovgu.de

Otto-von-Guericke-Universität Magdeburg

To be presented at the 27th Biennial Meeting
of the Philosophy of Science Association

Abstract

Models developed using machine learning (ML) are increasingly prevalent in scientific research. Because many of these models are opaque, techniques from Explainable AI (XAI) have been developed to render them transparent. But XAI is more than just the solution to the problems that opacity poses—it also plays an invaluable exploratory role. In this paper, we demonstrate that current XAI techniques can be used to (1) better understand what an ML model is a model *of*, (2) engage in causal inference over high-dimensional nonlinear systems, and (3) generate algorithmic-level hypotheses in cognitive science.

Acknowledgements

This work was supported by the German Research Foundation project ZE-1062/4-1. Carlos Zednik is thankful for the opportunity to attain insights into AI industry applications at the neurocat GmbH in early 2020.

1. Introduction

Models developed using machine learning ("ML models") are increasingly prevalent in scientific research. In neuroscience, ML-programmed classifiers are used to specify the representational contents of brain states and to predict human behavior from fMRI data (Ritchey et al. 2017). In astrophysics, classifiers trained on telescope imagery are used to determine the possible location of exoplanets (Datillo et al. 2019). In materials science, machine learning is used to discover stable materials and to predict their crystal structure (Schmidt et al. 2019).

Recent discussions have focused on the fact that many ML models are *opaque* (Humphreys 2009). Loosely speaking, a model is opaque when it is difficult to understand why it does what it does or to know how it works. Recent attempts to assess the impact of opacity generally agree that opacity prevents different stakeholders[1] from achieving goals such as intervening on the system when it breaks down, or evaluating its behavior against ethical and legal norms (Burrell 2016; Hohman et al. 2018; Zednik 2019).

In philosophy of science, the most important stakeholder is the *scientific investigator*. Scientific investigators are known to use ML models to achieve epistemic goals such as describing a phenomenon (e.g., distinguishing the fMRI signatures of fear and excitement), predicting new observations (e.g., determining the probable location of an exoplanet), and explaining observed data (e.g., identifying a causal link between smoking and lung cancer). Opacity can negatively impact scientific research by preventing investigators from using ML models to achieve some or all of these epistemic goals.

That said, little is known about the positive impact of recent attempts to overcome opacity through *Explainable Artificial Intelligence* (XAI). This nascent research program aims to develop analytic techniques with which to render opaque models *transparent* by answering questions about why they do what they do or how they work.[2] Whereas these techniques' importance for industry and governance is becoming increasingly apparent (Doran et al. 2017; Wachter et al. 2018), their utility for scientific research remains uncertain.

This paper argues that Explainable AI can play an invaluable but hitherto unrecognized role in *scientific exploration*. Recent discussions of exploration distinguish at least four distinct but not mutually exclusive aspects (for discussion see e.g., Gelfert 2016): identifying a *starting point* for future inquiry; providing a *proof-of-principle* demonstration; providing a *potential explanation* of a specific (type of) phenomenon; and assessing the *suitability of a particular target*. Whereas previous contributions have considered the exploratory role of ML models in their own right (e.g., Cichy & Kaiser 2019), little is known about the unique exploratory utility of Explainable AI.

[1] Tomsett et al. (2018) provide a helpful taxonomy of stakeholders in the *ML ecosystem*, distinguishing between creators, data-subjects, operators, executors, decision-subjects, and examiners.

[2] Although Humphreys (2009) and several others claim that some ML models are *essentially* opaque, the present discussion is agnostic with respect to this claim. That is, it only concerns models that can in fact be rendered transparent through Explainable AI, however numerous these may be.

The following discussion describes three ways in which Explainable AI facilitates scientific exploration. Section 2 shows that some XAI techniques are well-suited for determining what ML models are models *of,* and thus, for assessing a model's suitability for a particular target. Section 3 shows that other XAI techniques can be used for causal inference, and thus, for specifying starting points for future inquiry into the causes of a particular event. Finally, section 4 shows how Explainable AI can be used to generate novel hypotheses about the algorithms that are implemented in biological brains, and thus, to provide potential explanations.

Importantly, in each one of these ways, XAI techniques' exploratory contributions can be distinguished from the contributions of the ML models to which these techniques are applied. Thus, more than just being a solution to the problem that opacity poses, Explainable AI enhances the overall exploratory potential of machine learning and data-driven scientific inquiry.

2. Determining What a Model is a Model *Of*

In a recent commentary, Emily Sullivan (2019) examines the use of ML models in scientific research. Although she denies that opacity negatively impacts these models' scientific utility, Sullivan argues that their *link uncertainty* does. Sullivan defines link uncertainty as "a lack of scientific and empirical evidence supporting the link that connects the model to the target phenomenon" (Sullivan 2019: 1). In other words, link uncertainty arises when it is unclear what a model is a model *of*. As an illustrative example, Sullivan considers *Deep Patient*: a DNN that learns to map patients' features onto likely diseases (Miotto et al. 2016). Her point is to argue that, although the network issues reliable diagnostic predictions, the understanding that medical scientists can acquire from this model is limited. This is because it is unclear whether the model tracks genuinely causal relationships between patient features and likely diseases, or whether it is merely exploiting spurious correlations grounded in (for example) the fact that patients with certain features are tested more frequently than others.

Although Sullivan distinguishes link uncertainty from opacity, it is more appropriate to consider link uncertainty a special kind of opacity. Recall that a model is opaque when it is unclear why the model does what it does or how it works. Sullivan's discussion only concerns a lack of knowledge about how a model works. In particular, it is concerned with a lack of knowledge about a model's implementation in some particular programming language—an epistemic state that is all but guaranteed by the software-engineering practice of *encapsulation* (Mitchell 2002). This "implementation opacity" is problematic for expert creators (e.g., software developers) tasked with intervening on a model to improve its performance or to fix a bug. However, it is unproblematic for non-expert decision-subjects (e.g., medical patients) and examiners (e.g., governmental regulators), neither of which would know what to do with knowledge of a model's implementation even if they had it.

That said, stakeholders such as decision-subjects and examiners are also affected by opacity, albeit one that centers on questions about why a model does what it does, rather than on questions about how it works. Questions of this kind are answered not by specifying details of the model's implementation, but by justifying the model's behavior through *reasons* (Zerilli et al. 2018). Unlike a model's implementation details, which concern the syntactic structures specified in a computer program, reasons in this context are individuated semantically, by reference to the environmental features and regularities that the model has learned to track (Zednik 2019). Thus, the reason why Deep Patient predicts that type-2 diabetes is likely to develop in a particular patient may be that the patient is overweight (a good reason), or that she is of advanced age (a bad reason). When Sullivan writes about link uncertainty, she is referring to a particular kind of opacity: an inability to understand the reasons for an ML model's predictions.

Given this analysis, Sullivan's claim that "implementation opacity" does not negatively impact scientific research is unsurprising: Scientific investigators are more like examiners than creators. They do not generally require knowledge of how a model works. Rather, they are interested in understanding why it does what it does. For this reason, although a lack of implementation knowledge is no obstacle to scientific research, link uncertainty is.

But of course, exposing link uncertainty as a special kind of opacity is little more than a verbal clarification. Far more important is the question of whether (and if so how) this particular kind of opacity might eventually be overcome. Can Explainable AI help scientific investigators determine what a model is a model *of*? Moreover, to what extent does overcoming this kind of opacity contribute to scientific exploration?

Many XAI techniques specialize in providing semantically-individuated reasons for a particular model's outputs. Most notably, these include techniques for identifying the input elements—be they pixels in an image or values in a table—that bear a high responsibility for a particular output. For example, visualization techniques such as *Prediction Difference Analysis* (PDA, Zintgraf et al. 2017) allow investigators to understand the regularity that is being tracked by visually inspecting a heatmap. Do the highlighted pixel regions for a model of cancerous melanoma generally look like the features that are actually characteristic of cancerous melanoma, or do they look more like irrelevant (but nevertheless correlated) features such as freckles? Moreover, do the highlighted pixel regions of the model look like features that are already known to be indicators of cancerous melanoma, or do they depict hitherto unknown (but causally relevant) indicators?

Analogous non-visual techniques may be required for models trained over tabular data. For example, *Shapley Additive Explanation* (SHAP, Lundberg & Lee 2017) ranks a model's input variables by their relative importance for producing specific outputs. Do Deep Patient's predictions of type-2 diabetes depend more on (causally relevant) factors such as a patient's weight and family background, or on (spuriously correlated) factors such as age? Moreover, do the predictions depend on factors whose relevance for type-2 diabetes is already known, or do they depend on factors whose relevance has thus far gone unrecognized? Notably, because the model's input elements (e.g., pixel regions and table values) correspond to features of its environment (e.g., skin discoloration and patient features), they can be viewed as semantically-individuated reasons for the model's outputs. Insofar as techniques such as PDA and SHAP let investigators understand these reasons, they allow them to understand what an ML model is a model *of*. In this sense, these techniques can be used to combat link uncertainty.

Notably, Sullivan herself mentions some of these techniques in passing. Nevertheless, she stops short of recognizing their full significance for scientific exploration. In particular, although Sullivan argues that heatmaps are useful for "determining the suitability of the model" (Sullivan 2019: 25) because they can allow investigators to determine which regularity it has learned to track, she does not recognize that these techniques can also be used to identify such regularities in the first place. Indeed, ML models are renowned for their ability to uncover subtle and unintuitive regularities that would be difficult to uncover otherwise. By using techniques such as PDA and SHAP to better understand what ML models are models *of*—that is, to identify the regularities they have learned to track—scientific investigators can discover previously unknown regularities in the environment.

3. Enabling Causal Inference

The examples of link uncertainty mentioned by Sullivan are ones in which it is unclear whether the model has learned to track causal relationships as opposed to spurious correlations. But although XAI techniques such as PDA and SHAP allow investigators to determine which particular regularity is being tracked, they do not help determine whether any particular regularity is in fact a causal regularity. Put differently, these techniques do not enable causal inference.

Other XAI techniques can be used for exactly this purpose. Consider techniques that provide what Wachter et al. (2018) call *counterfactual explanations*. Counterfactual explanations specify possible worlds in which variations in a model's input yield non-actual (and possibly, desirable) outputs. A recent software tool for providing counterfactual explanations is the *Counterfactory*.[3] Given a model and input, this tool generates counterfactuals of arbitrary closeness (distance to actual input values) and complexity (number of input variables) to produce a desired but non-actual output. Thus for example, given a bank's credit-scoring model, the Counterfactory might generate counterfactuals for achieving an improved credit score: increasing income, decreasing monthly expenses, or some combination of both.

XAI techniques for counterfactual explanation can be used for causal inference, that is, for inferring the cause(s) of a particular effect. To understand how, it is worth briefly reviewing the close connection between counterfactual reasoning and causal inference. Consider an actual scenario in which event *C* (e.g. the striking of a match) precedes event *E* (e.g. the match catching fire), over an arbitrary number of background conditions *B* (e.g. the surrounding temperature being 19°C, there being oxygen in the air, etc.). Assuming that all *B* remain constant, one can infer that *C* is causally relevant for *E* if and only if a counterfactual change in *C* co-occurs with a change in *E*.

Causal inference can serve the purposes of many different stakeholders. Decision-subjects can assume a degree of control over model-driven decisions if they can

[3] Proprietary technology currently being developed by the neurocat GmbH: https://www.neurocat.ai/ (retrieved August 18th, 2020).

infer the changes to make so as to effect a different model output (e.g., whether they need to earn more to improve their credit score). Examiners can assess a model's compliance with ethical or legal norms if they can determine the causal relevance of certain key variables (e.g., whether credit scoring causally depends on gender or ethnicity). More relevant in the present context, scientific investigators can engage in causal inference to determine whether the regularity being tracked by a model is in fact a causal regularity. If a software tool can generate counterfactuals in which a change in *E* is predicted from a change in *C,* investigators might infer (assuming all *B* remain equal) that the learned relationship between *C* and *E* is genuinely causal as opposed to merely correlative.

Of course, the differences between the industrial and scientific contexts are significant. In industry, what matters is (typically) the model itself. In such contexts, XAI techniques for counterfactual explanation are perfect guides to causal inference: If the Counterfactory generates a counterfactual in which a higher income yields an improved credit score, then a higher income will actually yield an improved credit score. In science, by contrast, what matters is (typically) the domain that the model is a model *of*. Accordingly, in these contexts, XAI techniques for counterfactual explanation are imperfect guides to causal inference: If the Counterfactory generates a counterfactual in which losing weight yields a reduced probability of type 2-diabetes, then it is still possible that losing weight does not actually reduce the probability of type-2 diabetes. Because scientific models can be false, the causal inferences grounded on these models are insecure.

That said, the insecurity of XAI-driven causal inference does not render it useless for scientific research. On the contrary, it can serve an invaluable exploratory purpose. In particular, XAI techniques for counterfactual explanation can be used to refine extant causal hypotheses as well as to generate new ones. Consider the hypothesis that excessive weight is causally relevant for type 2-diabetes. This is a well-confirmed hypothesis, despite the fact that many overweight people never actually become diabetic (Wu et al. 2014). Nevertheless, it may be desirable to subsume the exceptions under a more-refined hypothesis. Indeed, applying the Counterfactory to Deep Patient might suggest suitable refinements. For example, counterfactuals generated for a desired outcome of less-probable diabetes might combine weight-loss with an additional factor, such as an absence of sleep apnea. Motivated by these counterfactuals, scientists might conduct further experiments, and if necessary, refine the original hypothesis so that excessive weight is only deemed causally relevant when it co-occurs with sleep apnea. In this (admittedly hypothetical) scenario, XAI-driven causal inference identified a starting point for scientific inquiry: generating new hypotheses, devising potential explanations, and inspiring new experiments.

Notably, XAI-driven causal inferences can perform this exploratory function in almost any scientific domain in which ML models have been developed for predictive purposes. In synthetic biology, for example, investigators may deploy such inferences to identify and test genetic modifications that are likely to yield desirable phenotypic traits (Ma et al. 2018). Analogously, in chemistry they might use XAI techniques for counterfactual explanation to discover new compounds with desirable (e.g., pharmaceutical) properties (Zhavoronkov 2018). Given the increasingly important role that machine learning plays in many different scientific domains, the exploratory promise of XAI-driven causal inference is tantalizing.

Before moving on, it is worth dwelling briefly on the kinds of domains for which XAI-driven causal inference might be particularly useful. Software tools such as the Counterfactory are remarkably efficient even for high-dimensional nonlinear DNNs, can be applied to any model-type and a wide variety of use-cases, and can generate counterfactuals even for intrinsically high-dimensional data-types such as naturalistic images. Given that ML models are capable of tracking high-dimensional and nonlinear regularities in complex systems such as the brain or the climate, such tools (assuming the relevant model is approximately true) might facilitate causal inference even for systems of such high levels of complexity. If true, this would be a significant achievement indeed: high-dimensionality and nonlinearity are among the biggest obstacles for traditional causal inference methods, which tend to work well only when the variables are few and the relationships are linear (Bühlmann 2013). Insofar as ML models can be trained to replicate the behavior of ever larger and more complex systems, and insofar as XAI techniques can be used to counterfactually explain the behavior of these models, Explainable AI is poised to significantly extend the limits of causal inference.

4. Generating Algorithmic-Level Hypotheses

Techniques from Explainable AI can perform at least one more exploratory role: generating *algorithmic-level hypotheses* that serve as potential explanations. The notion of an algorithmic-level hypothesis requires elaboration. Some physical systems—most notably biological brains—are computational systems insofar as they perform computational tasks in their surrounding environments (Shagrir 2006). Although these systems can be described at a physical level of analysis, by specifying the spatiotemporal structures and processes that underlie their behavior, it is often more insightful to describe them at an *algorithmic* level of analysis, by specifying the algorithms they execute in the service of the task (Marr 1982). Indeed, cognitive science is to a large extent in the business of formulating testable hypotheses about the structure, efficiency, and representational content of algorithms that biological organisms use to accomplish cognitive tasks such as perception, categorization, memory-formation, and language-learning. Notably, although many such hypotheses have been articulated and evaluated in the past, there is no general agreement about the way in which new algorithmic-level hypotheses should be developed in the future. To a certain extent, cognitive modeling remains an inscrutable “dark art”.

Explainable AI may help transform this “dark art” into a semi-autonomous exploratory process. Specifically, XAI techniques can facilitate the specification of algorithms to test as possible explanatory hypotheses. Indeed, given that many ML models are trained to perform tasks that closely resemble the ones that are performed by biological cognizers, and given that these models are often trained on naturalistic datasets that mirror the real-world environments in which those cognizers develop and learn, it is at least not wholly unreasonable to assume that ML models might implement algorithms that bear at least some similarity to the algorithms that are implemented in biological brains (see also Zednik 2018). Insofar as XAI techniques allow cognitive scientists to understand and describe the algorithms that are learned by a particular model, they can also be used to articulate new and hitherto unconsidered hypotheses about the algorithms that are learned by biological brains.

At this point, it may be necessary to clarify why XAI should be necessary at all, within the context of understanding the algorithms that are learned by ML-programmed models. Although human programmers typically decide on a model's *learning* algorithm, they have limited influence on the structure and function of what might be called the *learned* algorithm. For example, although they might train a DNN using some variant of the backpropagation algorithm, they do not determine the values that this algorithm (when applied to a particular learning environment) eventually assigns to individual network parameters (e.g., connection weights). Since it is these parameters that govern the model's output for any particular input, they implement a learned algorithm for computing a particular function. But what exactly this algorithm is, and how it might be characterized in a concise, understandable (and potentially modifiable) way, is obscured by the fact that the number of network parameters is high and their interdependencies are nonlinear.

Notably, whereas the XAI techniques considered in previous sections serve to answer questions about why an ML model does what it does by specifying reasons, the techniques to be considered here answer questions about how such a model works by uncovering algorithms. One way of uncovering algorithms is by using any one of a diverse family of *surrogate modeling* techniques. These techniques specify (relatively) simple algorithms to replicate (to an arbitrary degree of precision) an opaque model's overt behavior and internal processing. In particular, *rule-extraction* methods (e.g., Zilke et al. 2016) produce rule lists that approximate the input-output behavior of any high-dimensional DNN. Similarly *tree-extraction* methods (e.g., Wu et al. 2018) produce decision-trees that replicate the internal decision-structure of complex and (even recurrent) neural networks.

Intriguingly, these surrogate models bear a structural resemblance to classic "symbolic" models that were used widely in cognitive science throughout the 1960s, 70s and 80s. Because some of these models remain in use today, it is not unreasonable to suppose that surrogate models for explaining the behavior of trained ML models might be advanced as candidate hypotheses for explaining the behavior of biological cognizers. That said, many areas of cognitive science have by now moved on to "subsymbolic" methods that more closely resemble the methods commonly used by neuroscientists. Indeed, some of these methods may even serve double-duty, simultaneously explaining the behavior of biological brains and of artificial neural networks.

Consider, for example, *representational similarity analysis* (RSA, Kriegeskorte & Kievit 2013; Kriegeskorte et al. 2008). RSA is an integrative technique for data-analysis that lets neuroscientists relate multi-channel brain-activity data to each other, to behavioral data, to data produced by conceptual and computational models, and to stimulus descriptions by comparing (representational) dissimilarity matrices (RDMs). Cichy et al. (2016) have recently deployed this technique to compare temporal and spatial brain representations with representations in a deep feed-forward neural network trained for object categorization. That is, they aim to use RSA to identify a DNN's learned representations for object-recognition, and to determine whether these representations bear a structural similarity to the brain's representations in an analogous task.

How exactly is this aim achieved? First, for each signal space (DNN, fMRI, and MEG) Cichy et al. estimate the representational activity patterns associated with 118 experimental stimuli (images of natural objects over real-world backgrounds). Second, for each signal space of every pair of experimental stimuli, they compute the activity pattern dissimilarity. This yields 118-by-118 RDMs (each one of which contains the dissimilarity values for all experimental stimuli-pairs) for every DNN layer, every fMRI region-of-interest or searchlight, and every millisecond in the MEG signal. Third, DNN RDMs are directly compared to fMRI or MEG RDMs by calculating the Spearman rank correlation coefficients between them, yielding a relatively easy measure of brain-DNN representational similarity. In this way, RSA permits a specification of the representations that are used by both the DNN and the brain, and a subsequent comparison of these representations at the level of RDMs.

Indeed, the comparison reveals that "the DNN captured the stages of human visual processing in both time and space from early visual areas towards the dorsal and ventral streams" (ibid.: 1). Moreover, a close analysis of the representational structures in the DNN supports a series of specific empirical predictions:

> "Our results demonstrate the explanatory and discovery power of the brain-DNN comparison approach to understand the spatio-temporal neural dynamics underlying object recognition. They provide novel evidence for a role of parietal cortex in visual object categorization, and give rise to the idea that the organization of the visual cortex may be influenced by processing constraints imposed by visual categorization the same way that DNN representations were influenced by object categorization tasks." (ibid.: 9)

Overall, although (or perhaps because) RSA was originally developed by neuroscientists to investigate representations in the brain, this technique may not only be used to explain the behavior of trained neural networks, but also to generate and test algorithmic-level hypotheses about biological brains. Notably, in this particular case, the generated hypothesis seems likely to be confirmed, suggesting that XAI may not only facilitate exploration, but also explanation.

5. Conclusion

Models developed using Machine Learning are assuming an increasingly prominent place in scientific research. Many recent discussions recognize the problem that opacity poses to the use of such models, and some of these discussions have begun to reflect on the possibility of solving this problem through the use of Explainable AI. However, Explainable AI appears to be more than just a solution to a problem. This paper has sought to show that XAI techniques can serve an invaluable exploratory role in their own right, over and above the ML models to which these techniques are applied.

In particular, tools such as PDA and SHAP have been shown to answer questions about why a model does what it does. Thus, they allow scientific investigators to better understand what a model is a model *of*, and to assess its suitability for a particular target. Moreover, XAI techniques for counterfactual explanation have been shown to enable causal inference—perhaps even over domains that are at once high-dimensional and nonlinear. In this way, these techniques reveal new starting points for scientific inquiry: new hypotheses to test, and new experiments to conduct. Finally, surrogate modeling techniques and analytic techniques such as RSA can be used to better understand the algorithms and representations that are learned by models to accomplish particular tasks. Insofar as there is reason to believe that these algorithms might also be implemented in biological brains, they can be advanced as potential explanations in cognitive science. For all of these reasons and more, Explainable AI is a promising new tool for scientific exploration.

6. References

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 205395171562251.

Bühlmann, P. (2013). Causal statistical inference in high dimensions. *Mathematical Methods in Operations Research, 77*(3), 357–370.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*, 27755.

Cichy, R. M. & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences, 23(4), 305–317.*

Dattilo, A. et al. (2019). Identifying exoplanets with deep learning II: Two new super-earths uncovered by a neural network in K2 data. *arXiv*, 1903.10507.

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. *arXiv*, 1710.00794.

Gelfert, A. (2016). *How to do science with models. A philosophical primer.* Springer: Dordrecht.

Hohman, F. M., Kahng, M., Pienta, R., & Chau, D. H. (2018). Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics.*

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese*, *169*(3), 615–626.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2*(4), 1–28.

Kriegeskorte, N. & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences, 17*(8), 401–412.

Lundberg, S. M. & Lee, S. (2017). A unified approach to interpreting model predictions. *arXiv*, 1705.07874v2.

Ma, W., Qiu, Z., Song, J., Li, J., Cheng, Q., Zhai, J., & Ma, C. (2018). A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta, 248*(5), 1307–1318.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge, MA: MIT Press.

Miotto, R., Li, L., Kidd, B. A. and Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, *6*(1), 1–10.

Mitchell, J. C. (2002). *Concepts in programming languages*. Cambridge: Cambridge University Press.

Ritchie, J. B., Kaplan, D.M. & Klein, C. (2019). Decoding the brain: neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *British Journal for the Philosophy of Science* 70(2): 581-607.

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.

Schmidt, J., Marques, M. R. G., Botti, S. *et al.* (2019). Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials, 5*, 83.

Shagrir, O. (2006). Why we view the brain as a computer. *Synthese*, *153*(3): 393–416.

Sullivan, E. (2019). Understanding from machine learning models. *The British Journal for the Philosophy of Science*, axz035.

Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *arXiv*, 1806.07552.

Wachter, S., Mittelstadt, B. & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, *31*(2).

Wu, Y., Ding, Y., Tanaka, Y. & Zhang, W. (2014). Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention. *International Journal of Medical Sciences* 11(11): 1185-1200.

Wu, M., Hughes, M. C., Parbhoo, S., Zazzi, M., Roth, V., & Doshi-Velez, F. (2018). Beyond sparsity: Tree regularization of deep models for interpretability. *arXiv*, 1711.06178v1.

Zednik, C. (2018). Will machine learning yield machine intelligence? In V. Müller (ed.) *Philosophy and Theory of Artificial Intelligence 2017. PT-AI 2017. Studies in*

Applied Philosophy, Epistemology and Rational Ethics, 44. Springer: Cham

Zednik, C. (2019). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*.

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2018). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology*, *32*(4), 661–683.

Zhavoronkov, A. (2018). Artificial intelligence for drug discovery, biomarker development, and generation of novel chemistry. *Molecular Pharmaceutics,15*(10), 4311-4313.

Zilke, J. R., Mencia, E. L., & Janssen, F. (2016). DeepRED – Rule extraction from deep neural networks. In T. Calders, M. Ceci, D. Malerba (Eds.): *Discovery Science 19th International Conference* (pp. 457–473).

Zintgraf, L. M., Cohen, T. S., Adel, T,. & Welling, M. (2017). Visualizing deep neural network decisions: Prediction difference analysis. *The fifth International Conference on Learning Representations* (ICLR).