Is explainable AI a race against model complexity?

Advait Sarkar[a,b]

[a]Microsoft Research, Cambridge, United Kingdom
[b]University of Cambridge, United Kingdom

Abstract

Explaining the behaviour of intelligent systems will get increasingly and perhaps intractably challenging as models grow in size and complexity. We may not be able to expect an explanation for every prediction made by a brain-scale model, nor can we expect explanations to remain objective or apolitical. Our functionalist understanding of these models is of less advantage than we might assume. Models precede explanations, and can be useful even when both model and explanation are incorrect. Explainability may never win the race against complexity, but this is less problematic than it seems.

Keywords

human-computer interaction, human-centered computing, philosophy of artificial intelligence, artificial intelligence, machine learning, neural networks, explanation, interpretation

1. The explosive growth of model complexity

The revival and spectacular success of connectionism has created a regime where dataset size, model complexity (as measured by the number of parameters or weights), and computation time are king. The explosive improvement in the performance of deep learning models has been accompanied by an equally explosive growth of model complexity and computational expense.

There are some arguments that these large, expensive models may not actually be necessary. The lottery ticket hypothesis [1] postulates that larger models perform better because they are more likely to have pockets of parameters that are advantageously placed at the random initialization. Thus a randomly-initialized network is likely to contain a much smaller subnetwork that, trained in isolation, can match the performance of the original network. Pruning these models is an active area of research and debate [2, 3].

Nonetheless, the empirically superior performance of larger models has led prominent researchers to conclude that *"general methods that leverage computation are ultimately the most effective"* (Sutton's 'bitter lesson' [4]).

The 'scaling hypothesis' posits that once a suitable basic architecture has been found, we can generate arbitrary levels of intelligence simply by instantiating larger versions of that architecture. The manipulation and articulation of neural network structures has therefore become a prime preoccupation of the machine learning research community, albeit not without its criticisms. Rahimi critiques this turn of his own field, likening it to *"alchemy"* [5]. The web comic XKCD lampoons the practice of developing these models, describing them as an activity where you *"pour the data into this big pile of linear algebra, then collect the answers on the other side"*, pausing to *"stir the pile until* [the answers] *start looking right"* [6].

2. Explanations as translation and compression

There are many forms of explanation. Some are concerned with explaining the structure of the model: descriptions of its mechanisms (how does it work?), its capabilities and limitations (what can and can't it do?). Others about its construction: what data it was trained on, who built it and for what purpose. Explanations of these kinds aim to deliver intelligibility, fairness, accountability, and transparency [7, 8, 9]. Explanations can also be used as a mechanism to control for adverse outcomes, help improve models, and discover new knowledge [10].

However, the term 'explanation' is most commonly associated with individual predictions. Why did the model predict *this* (and why not *that*)? How 'confident' is the model, and how confident should a consumer of the model feel about this prediction?

Explanations of individual predictions aim to engender trust, but also help calibrate the use of such a model as an instrument within a decision support system. For example, the ASSESS MS system used by clinical neurologists to assess multiple sclerosis is part of a wider process involving multiple tools, procedures, and performances of expert judgment [11]. Just as they would seek to understand magnetic resonance imaging (MRI) as part of their process – the strengths, limitations and idiosyncrasies of the tool, and develop the 'professional vision' [12] re-

Joint Proceedings of the ACM IUI Workshops 2022, March 2022, Helsinki, Finland
advait@microsoft.com (A. Sarkar)
https://advait.org (A. Sarkar)

CEUR Workshop Proceedings (CEUR-WS.org)

quired to effectively read MRI images *through* the lenses of technical and medical knowledge – so too did they seek to understand how the predictions of an AI model could be incorporated into the process of diagnosis, a finding replicated in other clinical contexts [13].

Explanations of individual predictions are therefore attempts to *translate* from the language of computation to the language of practice.

Explanations translate, but they also compress and abstract. An early discovery in explanation research, subsequently replicated in several contexts, is that too much information overwhelms the user and thus undermines the explanation [14]. A ceiling to the information content of an explanation implies that as models grow, explanations must perform ever greater compression.

Within the modest room that explanations have for growth, alternative representations can help. In some domains, visual explanations can convey more information than textual ones while requiring less cognitive effort to process. Visual explanations are particularly natural in image classification problems, where saliency or attention highlighting [15], counterfactual images [16], and latent attribute visualisations [17] are popular forms of explanation. Despite the potential for alternative representations to improve the information bandwidth of explanations, it must be conceded that holding the form of an explanation constant, the compression ratio increases with model size.

Moreover, explanations derived from the model prediction process are a form of lossy compression, as anything short of a complete listing of parameters, activations (and perhaps more) would not capture the full information content of the 'decision-making' behind an individual prediction. Thus as the number of parameters within a model grows, the explanation must lose more detail and nuance, and become further removed from the underlying prediction.

3. Lessons from human explanations

The trend for model growth and explanation can be extrapolated in many ways, but one obvious extension is that models will approach levels of complexity comparable to human behaviour (i.e., 'brain-scale' models). The interpretation of consciousness, and the differences between software and wetware, are both cans of worms that shall remain unopened in this paper. Rather, by examining the issues of explaining human reasoning we may foresee the explainability issues of brain-scale models.

The first and most important issue is the fundamental unknowability of the mind to others, and to the self. The conventional account of philosophy of mind, and the intuition that our language creates, is that we cannot observe the thoughts and qualitative experience of others, but come to know them only through what they say and do (the problem of 'other minds' [18]). The unknowability of the mind to others is the dominant account because it aligns so well with the way we have organised our interactions and our language, although there are alternative perspectives (notably Wittgenstein's [19], which questions whether 'knowing' can even be said to be done of minds).

Likewise, we cannot even fully reason about our own minds. We cannot sample the activations of our own neurons, our memories are imperfect, there are innumerable environmental influences that we do not perceive or account for, and many of our thoughts and actions are performed unconsciously.

Second, people have agency and politics and therefore every explanation is subject to rhetoric, argumentation, and deception. Every explanation is given with an intended outcome. There is no such thing as a 'neutral' or 'objective' explanation, yet this is the unstated expectation of machine explanations. An explanation with a mathematical definition can be said to be objective in the sense that the *content* of the explanation is independent of the observer, but this is a relatively weak form of objectivity, akin to saying that human explanations are objective because the words being said are the same irrespective of who hears them. It ignores the fact that the choice of the mathematical definition itself is a political one, as is the interpretation of the explanation. Currently the politics of the explanation can be said to come from, and be within control of, the human creators and consumers of the models, but in a future scenario, it is not difficult to imagine a brain-scale model developing a bias towards explanations that ensure its continued survival. For example, a model might learn to manipulate users towards maximal engagement through intentionally adapted explanations for its recommendations.

For these two reasons, we may not be able to expect a uniformly satisfactory explanation for every prediction made by a brain-scale model. There may be conditions in which the behaviour can be satisfactorily explained, as well as those in which it cannot.

Despite these problems, for a great deal of human behaviour, we are capable of generating and giving satisfactory explanations to each other. An employee can explain why they were late ("Because my bus was cancelled and I had to walk."). A child can explain why he ate his brother's share of dessert ("Because he stole a sausage from me first!"). A man can explain why he bought flowers for his husband ("Because they are beautiful and they remind me of you."). None of these explanations requires bottomless introspection and psychoanalysis, and they serve the purpose of the explanation perfectly well.

Human explanations are produced in response to an implicit understanding of the context. The mother poised to admonish her child, in asking "why did you do this?", which could be interpreted and answered in any number of ways (e.g., "Because I am hungry", "Because I wanted to eat it", "Because I am supposed to eat dessert after dinner"), is really asking the child to provide an explanation of the form of a *contrastive* and *moral* justification with respect to the intended state of affairs (that the two children would each have their own desserts).

Situations in which explanations are demanded from people are saturated with context. This context is absorbed by interlocutors, usually effortlessly and unconsciously, and the episode culminates in the production of a satisfactory explanation.

What we have begun to uncover by examining these examples has been explored at length by Miller, who synthesises perspectives on human explanation from philosophy, social science, and cognitive science [20]. The findings are first, that human explanations are contrastive (i.e., *"sought in response to particular counterfactual cases"*); second, that they are selected in a ' *'biased manner"* from a *"sometimes infinite number of causes"*; third, that explaining an event in terms of the statistical likelihood of the outcome is *"not as effective as referring to causes"*; and finally, that explanations perform the social function of knowledge transfer, *"presented relative to the explainer's beliefs about the explainee's beliefs"*.

Requesting satisfactory explanations from brain-scale models will therefore require some notion of the context in which the question "why did you do this?" is being asked. With the question being so imprecise and reliant on context, users of these models may need a new form of language, or interaction technique, that allows them to specify localised areas of interest within the infinite space of possible valid explanations.

4. Our understanding of machine learning may not help

Unlike with human reasoning, we can at least expect to have a full functionalist understanding of the reasoning in brain-scale models. In theory, we should be able to reproduce any given decision and inspect the model's reasoning process with arbitrary detail. But as we are already finding with much smaller models, parameters and activations themselves are not *sufficient* for explanations; they must be summarised, contextualised, and externalised. We can fail to predict the emergent behaviour of a system despite having a complete functional understanding of its constituent elements. To borrow an example from Physics, we cannot predict states of the three body problem by solving Newton's equations [21]. There are particular solutions but not general ones. In general, we cannot solve the problem analytically but only through numerical approximations. While behaviour might be easy to explain using the theoretical model ("the mass is here at time *t* because of these equations"), results derived from numerical approximations do not precisely follow those equations and therefore cannot accurately be explained in those terms. They must be explained in their own terms, which involves explaining their many iterations and instantiated parameters.

Explanations discard and aggregate information across multiple parts of a neural network; knowing individual parameters and activations may not even be *necessary* if they are at the wrong level of abstraction. This can be thought of in terms of another Physics analogy: we can model many aspects of fluid dynamics with the Navier-Stokes equations [22], if initial or boundary conditions are available, despite the fact that they ignore the particulate nature of fluids. Indeed, many explanation techniques, such as the popular LIME [23] deliberately avoid inspecting the internal structure of the model (the 'M' in LIME stands for 'Model-agnostic'). Entire families of explanation techniques that rely on surrogate models, model distillation, and rule extraction [24, 10] are based on the premise that we can explain a model without accessing its internal workings. This is not without contention. Some reject these approaches outright for the precise reason, among others, that there are no guarantees that such explanations actually reflect what the model is doing [25].

Moreover, we cannot always expect to have an understanding of the training data. Dataset sizes are already large enough that no individual can explore every item within it. ImageNet [26], one of the most widely used machine learning research datasets, contained several racist, homophobic, ableist, ageist, and misogynist 'classes' of image [27]. It contained hundreds of images of real people labelled *"s**stic"*, *"f**ker"*, *"f**got"*, *"loser"*, *"kept woman"*, and so on. It is hard to imagine any conscientious researcher intentionally building a model using these labels, but the sheer size and complexity of the dataset meant that these were overlooked until the dataset became the focus of targeted research. As of this writing many such class labels have been removed from the official dataset, but for years they remained, being incorporated into the models built by thousands of researchers. There is also the issue that different people have different views of what ought to be considered harmful or objectionable.

There is no guarantee that more issues with the data will not be discovered. ImageNet contains over 14 million images. It would take a team of fifty people nearly 300 days to verify the labels on each image if they worked 8 hours a day, spending 30 seconds on each image. It would take an individual over 40 years. The OpenAI GPT-3 model [28] was trained on nearly 500 billion byte-pair encoded tokens, or approximately 245 billion words (assuming, conservatively, two tokens per word). It would take an army one-thousand strong nearly 4 years to read this much text, working 8 hours a day, continously reading 350 words per minute. The astronomical sizes of these datasets render them fundamentally unknowable at human scale.

At the time of deployment, the training data may not even be available. For reasons of privacy, security, and intellectual property ownership, the training data may be withheld from the users of a model or even destroyed. Explanations of brain-scale models therefore cannot be consistently expected to refer to the extrinsic influence of their training data, and may therefore be forced to internalise the blame for any error, and make 'original' reasoning indistinguishable from regurgitation of training data [29].

In the absence of data, we are faced with the absurd challenge of explaining why models do what they do, without being able to explain why they are the way they are. This is like trying to explain the course of a river only in terms of the motion of the water within it, ignoring the topography of the valley through which it runs.

Model parameters and activations are neither necessary nor sufficient for explanation. We do not always have access to the training data and when we do it can be so large as to be impossible to inspect comprehensively. These facts imply that our functionalist understanding of AI models may be of little advantage when it comes to explaining their behaviour, in comparison to explaining human behaviour.

5. Useful models precede explanations

While it is possible to develop models with explainability as a prerequisite, there is no fundamental obligation to do so. Thus, models usually precede the invention of mechanisms to explain them. In the period between the development of a model and the development of its explanation, the model may well be useful.

5.1. Correctness, explainability, and usefulness

Correctness and explainability have, perhaps frustratingly to some, an insecure relationship. We might wish that all correct models are explainable, and that all explanations are for correct models. But neither is the case: correct models may go unexplained, and incorrect models can have explanations. Furthermore: to be *useful*, a model needs to be neither correct nor explainable.

Before we proceed it is worth discussing the notion of an 'incorrect model'. The phrase may call to mind British statistician George E.P. Box's observation that *"all models are wrong, but some are useful"*, or Polish-American philosopher Alfred Korzybski's that *"a map is not the territory"*. By design, models aim to condense and simplify the complexity of (part of) the world so that it may be understood and predicted, and this necessarily incurs a loss in detail. It is this loss that for Box, makes all models "wrong" to a greater or lesser extent. However, these aphorisms are more accurately viewed as statements about the *incompleteness* of these models with respect to their referents, and their *inequality* to them, than about their *incorrectness*.

I suggest that a more helpful way to define an incorrect model is one which assumes or implies ontological and epistemic positions that contradict those of the domain being modelled. That is to say, in creating the model, we assume or predict the presence of nonexistent things, or the absence of existent things.[1] Or, we build and interpret the model with a different set of rules about knowledge-making than those with which we come to know its referent. Often, a model that is incorrect in this way can only be recognised as such after a 'paradigm shift' in the way the referent is understood, which can take generations of thinkers [30]. Thus if models usually precede the invention of mechanisms to explain them, they almost always precede the discovery that they might be incorrect.

Models may be incorrect in this deeper sense and still be useful. For example, the theory of epicycles, which dominated astronomy for centuries, allowed highly accurate predictions of the movements of the planets despite having a fundamental difference from the domain being modelled: the assumption of geocentrism. Newtonian dynamics is a similar story [30]. These models are notable for having compelling and satisfactory explanations despite being incorrect, and still useful for practitioners of those disciplines.[2]

Without explanation, too, an incorrect model can be useful. A relatable and contemporary example might be that of end-user programmers fighting abstraction [31]. When trying to automate a repetitive task, such as fixing spelling errors in a document, the end-user programmer may not care that the program does not handle edge cases, such as errors in domain-specific jargon, since she can manually inspect and correct those. So the program (model) that only accounts for words in its dictionary is incorrect, but useful.

[1] Note that a model in which some feature of its referent is absent, which is common, is not the same as a model that assumes or asserts the absence of said feature. The former is merely incomplete, whereas the latter is incorrect.

[2] While it may take years to detect an incorrect scientific model, literary writing makes abundant use of incorrect models that can be immediately understood as being incorrect, and yet which are extremely effective and useful. These incorrect models are better known as metaphors.

5.2. Explanations are not free

Another force causes a tendency away from explanations: explanations have a cost. Not only are they costly in terms of labour: it costs the time of scientists and programmers to develop the explanation mechanism, but they are also costly in terms of computation. Programs for explanation need to be stored at additional expense, and they cost compute cycles when run. Via computation, explanations incur energy costs, which, depending on the energy mix used to power computation, can result in increased carbon emissions. These material costs of explanation can be justified in terms of their benefits, and also in comparison to the material and immaterial costs of *non*-explanation, which may well be greater.

However, the dominant pricing model for machine learning is pay-as-you-compute [32]. Cloud and intelligence service providers such as Amazon AWS, Microsoft Azure, and the OpenAI API all charge in proportion to the amount of computation performed. Under this pricing model, explanations incur capital expenditure. Thus, even when the costs of explanation can be justified, they cannot always be borne. When access to capital mediates the relationship between users and explanations, we risk access to explainable models becoming yet another facet of the socio-digital divide [33].

Moreover, not all models require explanation. When we think of explanations for AI we often tend to fixate on and romanticise extreme applications, such as autonomous vehicles, recidivism prediction, and disease diagnosis. Yes, these are important areas and the costs of errors are high, and therefore explanation is key. But we tend to lose sight of the fact that most technology, most of the time, is used for relatively low stakes and mundane work, and AI is unlikely to be an exception. In many of these cases, incorrect models are useful, unexplainable models are useful, and the costs of building a 'correct' or explainable model are prohibitive. Interviews and diary studies of media recommender systems and search query autocompletion assistants have shown that users can achieve comprehension without explanation, that the costs of consuming explanations can outweigh the benefits, and that people rarely desire explanations in the daily use of these systems [34].

Many applications of brain-scale models will fall into the 'low stakes' category and therefore many models will continue to be produced which may be incorrect and unexplainable but still useful. At the same time, the trend is for larger models to be more general, and so the same model may be applied in a mix of high and low risk roles. Commercial offerings built upon brain-scale models may promote the explainability of their model as a competitive edge or as a premium offering, but if history is any indication, customers will prefer a cheaper or more performant model over a more explainable one.

6. The explainability crisis and grief

It therefore appears that explainability is indeed a race against model complexity, if we take together the observations that larger models are more performant, that explanations of larger models must necessarily compress to a greater degree and lose more detail in comparison to explanations of smaller models, that there are fundamental challenges to explainability when models approach human-scale reasoning and our functionalist understanding is of little help, and that explanations are costly and models may be developed and usefully applied before they are explainable.

It is clear we are headed for an explainability crisis, which will be defined by the point at which our desire for explanations of machine intelligence will far eclipse our ability to obtain them. Explanation is a wicked problem [35], perhaps *the* wicked problem of artificial intelligence research. The problem of explanation eludes definition, it does not have a stopping rule, solutions are not true or false, nor is there a definitive test of a solution. There are many possible approaches to the problem of explanation, and all explanation scenarios are essentially unique.

The research community, and society more broadly, appears to be dealing with the onset of this problem by *grieving*. Perhaps the most well-known account of grief is the Kübler-Ross model, the 'five stages of grief', namely: denial, anger, bargaining, depression, and acceptance [36]. While contemporary psychiatrists consider the model to be outdated and unhelpful in explaining the grieving process, the distinctions between the Kübler-Ross stages are uncannily apt descriptions for the various approaches proposed to deal with the explainability crisis.

Some deny there is a crisis. Breiman contends that there cannot be an accuracy-interpretability tradeoff because a more accurate model is, in some senses, inherently more *informative* [37]. However, the very motivation for seeking and preferring 'interpretable' models demonstrates that explainability does not follow from informativeness. Proponents of inherently intrepretable models uphold the demonstrable success of their models as evidence that accuracy does not have to be sacrificed for interpretability. Rudin proposes that many models can be made explainable by design with careful effort in feature engineering and data preprocessing [25]. However, it is not at all clear that it is always possible to put this design philosophy into practice [8].

Some react to unexplainable models with 'anger', or perhaps more accurately, *passion*. This is particularly acute when it comes to high stakes applications. Baecker advocates simply to avoid such 'risky' applications of AI altogether [38]. In such cases the loss of explainability is potentially too costly to justify the benefit of applying the system. In the works of researchers at the intersection of social justice and AI, such as Timnit Gebru and Kate Crawford, evocative phrases demonstrate their passion for this situation. In an article for the New York Times, Crawford writes [39]: *"[...] algorithmic flaws aren't easily discoverable: How would a woman know to apply for a job she never saw advertised? How might a black community learn that it were being overpoliced by software? We need to be vigilant about how we design and train these machine-learning systems, or we will see ingrained forms of bias [...] we risk constructing machine intelligence that mirrors a narrow and privileged vision of society, with its old, familiar biases and stereotypes."*

The bargaining approach seeks middle ground. Some avoid complex models, focusing on simpler and more inherently interpretable models, such as the hospital readmission models developed by Caruana [40], or the SLIM models for sleep apnea screening developed by Ustun and Rudin [41]. Others propose to build in structural interventions into these large models that guarantee (a form of) explainability. One example of such an intervention is the concept bottleneck model [42], which attempts to force the model to learn in terms of human-interpretable concepts.

Legislative approaches seem to bargain with the problem of explanation while simultaneously denying its existence. Recital 71 of the European Union's General Data Protection Regulation (GDPR) is commonly known as the 'right to explanation' [43]. It states that a *"decision which is based solely on automated processing and which produces legal effects"* entitles the subject of that decision to *"the right [...] to obtain an explanation of the decision"*. It is a bold statement of the principle while at the same time weak and underspecified. The French Loi pour une République numérique (Digital Republic Act) is marginally more potent [44], stipulating more clearly the minimal contents of an explanation, such as the data used and its source. However, legal scholarship notes that the 'right to explanation' approach has *"serious practical and conceptual flaws"* [45, 44], such as placing the burden on users to challenge bad decisions, and that data and weights, however accurately disclosed, may not be sufficient to show bias, unfairness, or deceit.

While the formal and reserved nature of academic writing precludes outright expressions of depression, there is no shortage of depression and anger in popular media and other societal expressions. Gig workers, long at the forefront of highly opaque and highly consequential automation, constantly strike with the demand that companies explain their algorithms [46]. For consumers of social media and recommender systems, their unexplainable nature is intimately bound up in their other harms, their capacity for disinformation [47], the destruction of mental health [48], and the destabilisation of democracy [49]. *"I've had enough of the bad feelings machine"*, writes Sirin Kale for the Guardian [50], *"Won't somebody switch it off? Please? Can we switch it off?"*

Finally, some accept that it may not always be possible to produce satisfactory explanations, or explanations with any formal guarantees of correctness. Some treat explanation, as humans do, as a metacognitive outcome resulting from introspection, and build metamodels that can explain the behaviour of these larger models, with either a white-box or black-box view into their inference process [51, 52]. Yet another approach is to treat interaction with AI as precisely that: an interaction design problem, and taking a cue from end-user programming research, focus on the ways in which users of these systems are not passive recipients of their predictions but play active roles in shaping their behaviour [31, 52]. This approach can be seen as having a Stoic focus on the elements of the system within our control, or it can simply be seen as reflecting the pragmatic focus on getting the job done that is a tenet of end-user programming.

In response to an earlier version of this paper, which did not draw an analogy between the Kübler-Ross model and the approaches proposed to tackle the explainability crisis, a reviewer remarked: *"Only the end of the paper, where a variety of paths to mitigate the race against model complexity for high-risk applications are briefly discussed, leaves me personally a little unsatisfied. I am not entirely convinced about their effectiveness, given our experience with explanations so far with 'below-brain-scale' models."* Such, often, is the nature of grief: it leaves us unsatisfied and unconvinced.[3]

7. Conclusion

The title of this paper asks the question: "is explainable AI a race against model complexity?" The unsettling conclusion is that it probably is, but also that this may not be as problematic as it seems. The correlation between the growth of model complexity and the improvement in AI capability is strong. Unexplained models precede explanations, and they are useful without explanations. We could attempt to avoid complexity by labelling its risks as too great, or we could attempt to tame it through structural interventions. We could try to improve the end-user programmability of such models, or invest more fundamentally in metacognition and introspection. The answer is likely to be some or all of the above.

[3] I must apologise to my kind reviewer for taking their words slightly out of context for rhetorical impact. They were not actually critiquing the paper at this point, subsequently writing: *"However, I don't see this as a weakness of the paper, but rather as a good entry point for interesting discussions."*

Acknowledgments

Thanks to Carina Negreanu and Alan Blackwell for their feedback on drafts, as well as to the reviewers for their constructive comments.

References

[1] J. Frankle, M. Carbin, The lottery ticket hypothesis: Finding sparse, trainable neural networks, arXiv preprint arXiv:1803.03635 (2018).

[2] Z. Liu, M. Sun, T. Zhou, G. Huang, T. Darrell, Rethinking the value of network pruning, in: International Conference on Learning Representations, 2018.

[3] D. Blalock, J. J. G. Ortiz, J. Frankle, J. Guttag, What is the state of neural network pruning?, arXiv preprint arXiv:2003.03033 (2020).

[4] R. Sutton, The bitter lesson, 2019. URL: http://incompleteideas.net/IncIdeas/BitterLesson.html.

[5] M. Hutson, Has artificial intelligence become alchemy?, 2018.

[6] R. Munroe, Machine learning, 2017. URL: https://xkcd.com/1838/.

[7] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, M. Kankanhalli, Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda, in: Proceedings of the 2018 CHI conference on human factors in computing systems, 2018, pp. 1–18.

[8] K. Sokol, P. Flach, Explainability is in the mind of the beholder: Establishing the foundations of explainable artificial intelligence, arXiv preprint arXiv:2112.14466 (2021).

[9] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, et al., Guidelines for human-ai interaction, in: Proceedings of the 2019 chi conference on human factors in computing systems, 2019, pp. 1–13.

[10] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (xai), IEEE access 6 (2018) 52138–52160.

[11] A. Sarkar, C. Morrison, J. F. Dorn, R. Bedi, S. Steinheimer, J. Boisvert, J. Burggraaff, M. D'Souza, P. Kontschieder, S. Rota Bulò, et al., Setwise comparison: Consistent, scalable, continuum labels for computer vision, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2016, pp. 261–271.

[12] C. Goodwin, Professional vision, American Anthropologist (1994) 606–633.

[13] A. Levy, M. Agrawal, A. Satyanarayan, D. Sontag, Assessing the Impact of Automated Suggestions on Decision Making: Domain Experts Mediate Model Errors but Take Less Initiative, Association for Computing Machinery, New York, NY, USA, 2021. URL: https://doi.org/10.1145/3411764.3445522.

[14] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, W.-K. Wong, Too much, too little, or just right? ways explanations impact end users' mental models, in: 2013 IEEE Symposium on visual languages and human centric computing, IEEE, 2013, pp. 3–10.

[15] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[16] L. Goetschalckx, A. Andonian, A. Oliva, P. Isola, Ganalyze: Toward visual definitions of cognitive image properties, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5744–5753.

[17] O. Lang, Y. Gandelsman, M. Yarom, Y. Wald, G. Elidan, A. Hassidim, W. T. Freeman, P. Isola, A. Globerson, M. Irani, et al., Explaining in style: Training a gan to explain a classifier in stylespace, arXiv preprint arXiv:2104.13369 (2021).

[18] A. Avramides, Other Minds, in: E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, Winter 2020 ed., Metaphysics Research Lab, Stanford University, 2020.

[19] L. Wittgenstein, Philosophical investigations. philosophische untersuchungen. (1953).

[20] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, CoRR abs/1706.07269 (2017). URL: http://arxiv.org/abs/1706.07269. arXiv:1706.07269.

[21] C. Marchal, The three-body problem (2012).

[22] C. L. Fefferman, Existence and smoothness of the navier-stokes equation, The millennium prize problems 57 (2006) 22.

[23] M. T. Ribeiro, S. Singh, C. Guestrin, "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, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 1135–1144. URL: https://doi.org/10.1145/2939672.2939778. doi:10.1145/2939672.2939778.

[24] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (2018). URL: https://doi.org/10.1145/3236009. doi:10.1145/3236009.

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

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[27] K. Crawford, T. Paglen, Excavating ai: The politics of images in machine learning training sets, AI & SOCIETY (2021) 1–12.

[28] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, arXiv preprint arXiv:2005.14165 (2020).

[29] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 610–623.

[30] T. S. Kuhn, The structure of scientific revolutions (1962).

[31] A. F. Blackwell, L. Church, T. R. Green, The abstract is an enemy: Alternative perspectives to computational thinking., in: PPIG, 2008, p. 5.

[32] M. Al-Roomi, S. Al-Ebrahim, S. Buqrais, I. Ahmad, Cloud computing pricing models: a survey, International Journal of Grid and Distributed Computing 6 (2013) 93–106.

[33] D. Boyd, K. Crawford, Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon, Information, communication & society 15 (2012) 662–679.

[34] A. Bunt, M. Lount, C. Lauzon, Are explanations always important? a study of deployed, low-cost intelligent interactive systems, in: Proceedings of the 2012 ACM international conference on Intelligent User Interfaces, 2012, pp. 169–178.

[35] H. W. Rittel, M. M. Webber, Dilemmas in a general theory of planning, Policy sciences 4 (1973) 155–169.

[36] E. Kübler-Ross, On death and dying, Routledge, 1973.

[37] L. Breiman, Statistical modeling: The two cultures (with comments and a rejoinder by the author), Statistical science 16 (2001) 199–231.

[38] R. Baecker, Digital dreams have become nightmares: What we must do (2021).

[39] K. Crawford, Artificial intelligence's white guy problem, The New York Times 25 (2016).

[40] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, in: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 2015, pp. 1721–1730.

[41] B. Ustun, C. Rudin, Supersparse linear integer models for optimized medical scoring systems, Machine Learning 102 (2016) 349–391.

[42] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, P. Liang, Concept bottleneck models, in: International Conference on Machine Learning, PMLR, 2020, pp. 5338–5348.

[43] M. E. Kaminski, The right to explanation, explained, Berkeley Tech. LJ 34 (2019) 189.

[44] L. Edwards, M. Veale, Enslaving the algorithm: From a "right to an explanation" to a "right to better decisions"?, IEEE Security & Privacy 16 (2018) 46–54.

[45] L. Edwards, M. Veale, Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for, Duke L. & Tech. Rev. 16 (2017) 18.

[46] M. Murgia, Workers demand gig economy companies explain their algorithms, 2021. URL: https://archive.is/mAFVe.

[47] S. Bradshaw, P. N. Howard, The global organization of social media disinformation campaigns, Journal of International Affairs 71 (2018) 23–32.

[48] J. Gao, P. Zheng, Y. Jia, H. Chen, Y. Mao, S. Chen, Y. Wang, H. Fu, J. Dai, Mental health problems and social media exposure during covid-19 outbreak, Plos one 15 (2020) e0231924.

[49] J. A. Tucker, Y. Theocharis, M. E. Roberts, P. Barberá, From liberation to turmoil: Social media and democracy, Journal of democracy 28 (2017) 46–59.

[50] S. Kale, Social media is a bad feelings machine. why can't we just turn it off for good?, 2021. URL: https://www.theguardian.com/commentisfree/2021/dec/27/2021-reporting-harm-social-media-online.

[51] N. Burkart, M. F. Huber, A survey on the explainability of supervised machine learning, Journal of Artificial Intelligence Research 70 (2021) 245–317.

[52] A. Sarkar, Interactive analytical modelling, Ph.D. thesis, University of Cambridge, Cambridge, United Kingdom, 2016.