The opaque nature of intelligence and the pursuit of explainable AI

Thomson, S.L.; Stein, N. van; Berg, D. van den; Leeuwen, C. van; Marcelloni, F.; Lam, H.K.; ... ; Filipe, J.

On the Opaque Nature of Intelligence and the Pursuit of Explainable AI

Sarah L. Thomson[1], Niki van Stein[2], Daan van den Berg[3], and Cees van Leeuwen[4]
[1]*Edinburgh Napier University, United Kingdom*
[2]*LIACS Leiden, Netherlands*
[3]*VU and UvA Universities Amsterdam, Netherlands*
[4]*KU Leuven, Belgium and RPTU Kaiserslautern, Germany*
[3]*daan@yamasan.nl* [1]*s.thomson4@napier.ac.uk* [2]*n.van.stein@liacs.leidenuniv.nl* [4]*cees.vanleeuwen@kuleuven.be*

Keywords: explainable artificial intelligence, xai, machine learning, neural networks, cognitive science

Abstract: In this work We consider and discuss the problems which come with trying to explain human and machine intelligence. How explainable artificial intelligence research is being carried out, the pitfalls and limitations of current approaches and the bigger question of whether we need explanations for trusting inherently complex and large intelligent systems, whether artificial or not.

1 Explainable (Artificial) Intelligence

Those of us who have returned to the refrigerator multiple times expecting different food to have materialised know that human behaviour is often inexplicable — but at least we can rely on the cold logic of machines to make decisions. Or can we?

Explainable artificial intelligence (Das and Rad, 2020), often referred to as XAI, is an emerging and somewhat embryonic field. XAI is the development and analysis of a set of tools with the motivation of providing human-readable explanations about how artificial intelligence algorithms make their decisions. Although there has been increased interest and effort in this research area lately (Mei et al., 2022; Belle and Papantonis, 2021; Trajanov et al., 2022; Keane and Smyth, 2020), there is a lack of proper analysis of XAI methods, and a lack of consistency in how XAI is carried out. Additionally, there are important limitations to XAI techniques. In the sections which follow, we discuss some common methods and the problems with using them. At the end of the section, some overarching issues are presented.

2 Feature Attribution Methods

Most popular XAI methods fall into the category of feature attribution methods, meaning they attribute a relative or absolute importance measure to each feature for a given machine learning model and its prediction. These methods work post-hoc and are usually model-agnostic. These methods aim to either explain a single prediction (local) or a complete machine learning model (global). Local explanation methods can also be used on an entire training or test set to provide more global explanations.

2.1 Local Feature Attribution

One of the most popular feature attribution methods is Shapley additive explanations (Lundberg and Lee, 2017), usually referred to as SHAP.It is used for example in (Jansen et al., 2020; Ariza-Garzón et al., 2020; Yeung et al., 2020; Van Stein et al., 2023) to provide local explanations for an individual prediction. Given a feature, f_1, SHAP considers models which contain f_1 and obtains the predicted values for the input data at hand. SHAP also does this for models which are identical to those in the previous step, *except* f_1 (and only f_1) has been removed as a predictor. The mean differential between the predicted output (including f_1) and the predicted output (excluding f_1) are the feature's *marginal contribution*; the SHAP value for f_1 is the mean marginal contribution over all considered models. SHAP values can be positive, negative, or even zero.

Despite the prevalence of SHAP in explainable AI, it exhibits several disadvantages. For large feature

sets SHAP is computationally expensive, and in these cases it relies on approximation techniques such as exploiting the tree information in TreeSHAP (Yang, 2021) or by using a subsample of model configurations; it follows that randomness can have an effect on the computed values. In addition, values can depend on the order in which features are presented. SHAP is not particularly stable: for example, a feature may have a large SHAP magnitude for one specific input, but not for any other. Additionally, there is the question of just how human-accessible SHAP values are. They are essentially just numbers on a non-normalised scale and it may not be clear to a stakeholder or patient how to interpret them. There is also the issue that SHAP values are unlikely to be intuitive or helpful when the features in question are individual pixels in image data (a map of image features with scores like this is called *pixel attribution*) or complex time-series data. A recent work showed that SHAP can be misleading when the marginal contributions for a feature have differing amounts of noise (Kwon and Zou, 2022); they proposed *weightedSHAP* to address this issue.

Another local feature attribution method is *Permutation feature importance* (PFI) (Fisher et al., 2019). PFI is a method which is conceptually similar to SHAP. PFI results in feature scores which are not for a single prediction, but typically represent a test set of data. To obtain a feature importance score, the values of that feature in each observation within the test set are randomly permuted and the output obtained several times. The mean difference between the predictions from the non-permuted data and those from the permuted data is taken to be the importance score for that feature. A limitation of PFI is that the scores may depend on the randomness used in the permuting stage. Additionally, it shuffles one feature at a time, thereby assuming the variables are independent and not considering the possibility of feature interaction.

Local interpretable model-agnostic explanations (Ribeiro et al., 2016), or LIME for short, is another popular local XAI approach (Magesh et al., 2020; Gabbay et al., 2021; Kuzlu et al., 2020). LIME estimates feature importance magnitudes for a prediction by randomly perturbing the values of the input data several times and obtaining the resultant prediction by the model. A separate linear model is then fit to the perturbed inputs and associated outputs; the coefficients for the linear model are the LIME scores for the original model.

One of the limitations of LIME is that it depends on the randomness and size of perturbations applied to the input data. These effects can result in different scores for the same features. LIME is designed for computing feature scores for a single prediction, meaning that it could fail to pick up on global patterns or overall model behaviour. Another limitation of LIME is that it might generate perturbations that are infeasible or unrealistic in reality (due to constraints or underlying feature interactions), and therefore generate explanations that are unrealistic.

2.2 Global Feature Attribution

While each local feature attribution method can be used for approximating global explanations, there are also methods specifically designed for attributing importance to features on a global model level. Sensitivity analysis methods are perhaps the oldest variant of explainable AI. The Morris method (Morris, 1991) or Sobol sensitivity analysis (Sobol, 2001) are methods to create global explanations of a model by using a large space filling design of samples and computing the sensitivity scores for features, groups of features and feature interactions. These methods also allow for the computation of second and higher order interactions, but they are computationally very expensive and do not explain single predictions. Next to Morris and Sobol there is a large number of other similar approaches (Van Stein et al., 2022) that can be used for global sensitivity analysis. Most of these methods are limited to specific sampling methods, require a large number of samples to show robust behaviour and are computationally expensive.

2.3 Feature Interactions

Real world prediction scenarios often — if not always — exhibit *interactions* between features; this means that the combined effect of two or more features is different than what their additive individual effects would be. This can be the case in, for example, predicting breast cancer (Behravan et al., 2020); acute coronary syndromes (Alsayegh et al., 2022); and hypertension (Elshawi et al., 2019). Despite this, common XAI techniques do not properly address, account for, and uncover feature interactivity: SHAP, permutation feature importance, LIME, and counterfactual explanations do not manage this well. There are some tools which are aimed at feature interaction, however. Friedman's H-Statistic (Friedman and Popescu, 2008) is based on partial dependence decomposition and represents the proportion of variance explained by an interaction. The H-Statistic is very computationally expensive (Molnar, 2020); indeed, the experience of an author of the present work is that it can be *prohibitively* expensive in situations where computational power is restricted due to data privacy. The

H-Statistic is also sensitive to noise in the data.

3 Counterfactual Explanations

Counter-factual explanations (Keane and Smyth, 2020) are a human-friendly XAI approach. They are written in human language and take the form 'if X, then Y', where X is a configuration of — or change to — the input data and where Y is the resultant predicted response. To generate a counterfactual for a particular input, the practitioner decides what they desire the output to change to. In regression contexts, an example might be 'for the predicted revenue to increase by £500'; for classification, it might be 'for the prediction of cancer to switch to no cancer'. A search algorithm is then used to discover which mutants of the original input data result in the desired outcome. These solutions are then converted into human-readable sentences; these are counterfactuals.

Although this approach is intuitive and widely understandable to stakeholders, there are several limitations. Counterfactuals do not consider feature interactivity, or address the problem of correlation versus causality. Multiple conflicting counterfactuals can exist for the same model, and in these situations it is not clear which takes precedence over the other.

4 Model Intrinsic Explanations

Model intrinsic XAI techniques are mainly presented in the context of artificial neural networks, where the weights of the layers, the gradients or attention mechanisms (Vaswani et al., 2017) are used to generate explanations.

Neural networks often have millions, or indeed billions, of parameters. With this in mind, it might be argued that intelligible explanations for what is happening inside the network are improbable. Even so, there have been some steps forward to this end. Network dissection (Bau et al., 2017) is an approach for convolutional neural networks (CNNs) which captures how interpretable learned features in the latent space are. The method maps channels which have been significantly 'activated' with human-defined objects, such as 'ear'. Unfortunately, in realistic CNN architectures there can be a very high number of channels to consider. Additionally, while the explanation of a network component is valuable, it does not explain the whole system and can miss feature interactions. Also, it could be argued that these explanations are not truly accessible: grasping them fully requires an understanding of CNNs.

As mentioned in Section 2.1, pixel attribution maps display an importance score for each pixel comprising input data, and can be based on SHAP values (or indeed LIME). There are also gradient-based tools for this, such as Image-Specific Class Saliency (ISCS) (Simonyan et al., 2013). ISCS works by propagating an particular image through the network and then using derivatives to compute the gradients attributable to input pixels. Another gradient-based approach is gradCAM (Selvaraju et al., 2017), which calculates the gradients backwards to the deepest convolutional layer and outputs a map indicating important regions of the original input image.

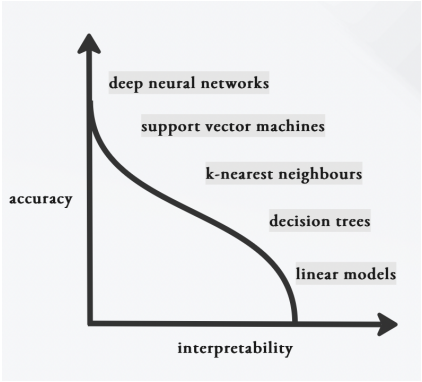
Similarly to network dissection, ISCS and gradCAM can miss feature interactions. ISCS can have difficulty identifying small features, and its precision can be quite coarse-grained. GradCAM can sometimes identify image regions which are not actually relevant to the desired explanation, leading to a misleading interpretation.

A solution to the problem of unintelligible neural networks might be deliberately simplifying models with explainability in mind. The problem with this is the known tradeoff between accuracy and complexity (which is visualised in Figure 1); it is likely that a substantial simplification would be needed to facilitate truly accurate justifications for decisions — an example of simplification would be reducing parameter cardinality by removing network layers — and a corresponding decrease in model quality would be expected.

5 Interpretable Models

There is also, of course, the option of deliberately choosing models which are known to be inherently interpretable: decision trees or linear models, for example. A decision tree, with the rules it has learned from the data, can be visualised (if not too big). People often find that reading the binary rules is intuitive and accessible. With linear models, feature coefficients can be extracted. These are essentially weights or importances for the features. Despite these interpretability advantages, deciding upon one of these as your model is not a straightforward choice: decision trees have a sensitivity to noise, and linear models have underlying assumptions which may not suit non-trivial real world data. More complex models such as neural networks may often be needed to capture nonlinear patterns in data; in general, more interpretable models are less accurate. This phenomenon is shown in Figure 1. Notice that while linear models are at the high end of interpretability, they are typically lower in ac-

Figure 1: The tradeoff between model accuracy and model interpretability. More complex models are less understandable. The question becomes one of priorities: what is more important, a highly accurate model or an intelligible one?



curacy. On the other end, deep neural networks tend to be low in interpretability but higher in accuracy.

6 Considerations and limitations

Throughout the discussion of popular XAI techniques, we notice that a common limitation to them is their sensitivity to randomness and noise; this could be formalised as their lack of stability. There is also the issue that prevalent XAI tools such as SHAP, permutation importance, LIME, and counterfactuals do not properly consider feature interactions. In addition there is not much work done to integrate uncertainty quantification in XAI, as often machine learning models have to deal with uncertainty and have uncertainty in their predictions. In practise, XAI results are presented often over-confident without taking uncertainty and model over- and under-fitting into account.

We observe that there is a lack of consistency in how practitioners carry out explainable artificial intelligence. For example, some carry out SHAP in isolation (Moncada-Torres et al., 2021); some use both SHAP and LIME (Rao et al., 2022); others use SHAP, LIME, and counterfactuals (Zhou et al., 2022). Aside from the specific tools used, there do not appear to be known 'best practice' axioms yet. In addition, there are only a few (very recent, somewhat limited and not yet widely used) benchmark suites for XAI methods (Liu et al., 2021; Arras et al., 2022; Agarwal et al., 2022; Clark et al., 2023).

An arterial problem in the field of XAI is the phenomenon of 'false explanations'. False explanations are inaccurate or misleading and can arise for a number of reasons: for example, noise (in the data, the model, or the XAI method itself); spurious correlations (also known as the Rashomon effect (Leventi-Peetz and Weber, 2022)), and the issue of *causality versus correlation*; and bias in the training data, which may result in that bias being amplified through explanations of prediction. The most salient challenge in XAI, however, is arguably the accuracy-complexity trade off which was mentioned in Sections 4 and 5: neural networks are popular due to their unrivalled accuracy, but the task of making their inner workings truly comprehensible and accessible is gargantuan.

7 Methods for exploring Explainable Real Intelligence

As the anecdote at the beginning of this article suggested, there are people who expect redemption from their refrigerator. Perhaps, therefore, we should reconsider the likelihood of finding explainable real intelligence (ERI) in humans. Human intelligence is a multifaceted concept. Our search for ERI requires a focus on decision-making capacity. At the face of it, ERI seems readily available. We can simply ask decision makers to explain their considerations. In everyday life, this is normally sufficient.

In science, more caution is needed. Introspective reports can be considered fundamentally unreliable (Schwitzgebel, 2008). People are known to come up with all sorts of rationalizations after the fact. Therefore, we may want them to express their reflections while they are still in progress (think-aloud protocols; (Simon and Ericsson, 1984). However, people hide their true motives; some cultures find it unusual, uncomfortable and unnatural to express what they are thinking (Güss, 2018; Kim, 2002). Moreover, many people have too limited a vocabulary to do so, or lack the necessary metacognitive skills such as self-control, prediction, and self-questioning (Wong and Jones, 1982).

A final reason why thought protocols may be unreliable is that deliberation might not always be conscious. Dijksterhuis and Nordgren proposed that decisions improve after a diversion of conscious thought (Dijksterhuis and Nordgren, 2006). This, apparently, because the thought process continues unencumbered by conscious hangups, and becomes more fruitful. Yet this seems to be a red herring, as a study by Nieuwenstein et al. found the evidence for improved decision-making not replicable (Nieuwenstein et al., 2015).

Think-aloud protocols can be informative in domains fostering covert speech, for instance in complex math or for monitoring the user experience of

automated devices (Simon and Ericsson, 1984). Even there will protocols necessarily be incomplete, given the time limits on what people can overtly verbalize while performing an attentionally demanding task. In nonverbal, i.e. pictorial domains, sketching made during the process may be collected to understand the reasoning (Jaarsveld and van Leeuwen, 2005). Capacity limitations similarly apply to have sketching produce informative results.

When introspective reports or sketches are unavailable, we may turn to implicit measures such as eye tracking, or decoding neural signals. In humans, noninvasive signals can be obtained through EEG/ MEG, or fMRI, among others. Eye-tracking can inform us what an observer is fixating on, and therefore is attending to. But this measure has limitations: in real images, several items compete for attention. As a result, observers often fixate on one, while focusing covert attention on another. As a result, both are quickly forgotten (Nikolaev et al., 2013). Eye-tracking results, therefore, can be unreliable at times, in particular when complex, realistic scenes are involved.

Decoding algorithms for brain signals were initially developed in the context of brain-computer interfaces. Within the temporal (fMRI) and spatial (EEG) restrictions of the medium, they reveal the non-stationary and dynamic patterns of brain activity that play increasingly prominent roles in our efforts to understand cognitive processes (Loriette et al., 2022). This field is rapidly expanding. Machine-learning-based techniques for decoding dynamic signals are used for identifying the locus of covert attention in humans (Astrand et al., 2015). Cross-temporal decoding can be used for distinguishing codes for stable stimulus representation from transient ones, which presumably are used in computation (King and Dehaene, 2014). Despite these advances, they can provide us with only a fragmented understanding of what the brain does. We can identify patterns in neuronal activity, but what we observe turns out to be highly context-specific. In combination with the high-dimensionality of the brain, this implies that patterns are hard to predict. Unlike in artificial neural networks, we have only limited knowledge of the dynamics by which brain and brain activity evolve, what aspects of the activity and structure are relevant, and which are not. In other words, we need a theory of mind and brain to guide us in developing our hypotheses and predictions involving brain signals.

8 Theories of ERI

ERI is traditionally associated with rationality, i.e. following rules or maxims in decision-making (Kaisla et al., 2001). Not all rules are good. Thus the notion of rationality has inherently a moral component. The doctrine of liberalism prescribes that it is ultimately beneficially for society, when each individual pursues their own benefit. As a result, classical economics has long upheld the fiction that decisions optimize value (or utility) to the individual. Psychologists have helped dismantle this idea, two of which have been awarded Nobel prizes. (Simon, 1956) proposed that decisions are made by satisficing (a port manteau of sufficing and satisfying). Rather that optimal benefits, those are preferred that are good enough, and easily to obtain. We may still consider this rational if we take into account the limitations to our information processing capacity and the information available to us (Simon, 1982). More generally, (Simon, 1978) argues that rationality should take into account the procedural aspects of decision-making, both individually and within an organization and its environment.

Thus satisficing is an "ecologically rational" strategy that enables efficient decision-making under time constraints. Like other animals, humans are sometimes forced to do just that. Add to this the fact that human decisions typically are made in a social context, and collective decisions may deviate from individual ones: "I want A but we want B". To accommodate these aspects of our decision-making, Daniel Kahneman famously developed his two-systems theory (Kahneman, 2011). System 1 is involved in decisions which are made effortlessly, intuitively, involuntary or habitually and with minimal conscious involvement, while system 2 is all about reasoning processes needing focused attention (Stanovich and West, 2000). This distinction resembles that between automatic and controlled processing in visual search (Schneider and Shiffrin, 1977; Treisman and Gelade, 1980) but goes beyond it in scope. System 1 includes all innate cognitive skills and ones acquired through extensive practice, such as reading and grandmaster chess. System 2 encompasses reasoning, selection, and is associated with a sense of agency. Both systems interact; a salient stimulus (e.g. a loud bang) triggers System 1, which alerts System 2 which takes control to suppresses System 1's flight response and produces the reasoned decision whether to explore the source. System 2 can instruct System 1. Waiting for a relative at the station, and knowing that the person has a beard, System 2 instructs System 1 to look for a person with a beard. System 1, which determines routine decisions, operates with superficial heuristics and

is liable to biases such as availability, representativeness and anchoring (Tversky and Kahneman, 1974), implying that reasoned decisions are superior. This is a strong claim whose value depends on a precise demarcation of both systems.

However, the broadness of these concepts and the appeal to intuitive examples makes it hard to pin down. The soundness of the empirical basis of Kahneman's work has been contested. Namely, Gigerenzer et al. argue that the representativeness heuristic implies that people ignore base rates in belief revision (Gigerenzer et al., 1988). Tversky & Kahneman's "engineers versus lawyers problem" purported to show that people do not revise their beliefs in light of probability information (How likely is the person matching the description of a typical lawyer to be an engineer, given this description is drawn from an urn with 30/70 vs 70/30 engineers) (Tversky and Kahneman, 1974). Whether base rate neglect occurs turns out to depend on the context. In domains where people have everyday familiarity in applying probabilistic reasoning ("how likely is Sunderland to win against Manchester United, given that the half time score is 3–1"), base rates are not ignored. In other words, people here operate like Bayesians. Gigerenzer argues that what goes by System 1 is actually more intelligent than Kahneman suggests, and that its "gut feelings" often are superior to reasoned decisions (Gigerenzer, 2007).

9 From Behavior to the Brain

Bayesian principles today are believed to underly much of our everyday responses. Predictive coding theory assumes that the brain constantly keeps and updates an internal model of the environment. The model is tested and updated against our sensations. Testing and updating happens recursively on several hierarchical strata, where the higher level passes predictions to the lower one, and the lower level sends prediction errors (or surprise) up, as a result of which the priors are adjusted. Predictive coding originated in models of the visual system (Rao and Ballard, 1999) and was generalized to a theory of cognition and brain (Clark, 2013; Friston, 2010). It provides an action-oriented view of cognition, given the output generated by the top-down stream projects to the motor system. According to Friston, reduction of overall prediction error is the basic function of the brain. He postulates this principle on account of a thermodynamic analogy, identifying prediction error with free energy. Living systems are unique as self-organizing systems in that they work to maintain or increase order within their system. Hence the states with locally minimal free energy constitute a global attractor for the system. As long as the system dwells near the attractor, sensory surprisals are supposed to be maximally infrequent and cause minimal perturbance.

Note first, that students of neural networks will be familiar with what is being advertised here. Similar principles involving energy minimization can be found in Hopfield networks and Boltzmann machines, and in statistical inferencing algorithms (MacKay, 1995); attractor dynamics are the bread and butter of recurrent neural networks. The way surprises are minimized resembles the Generative Adversarial Network (GAN) approach. None of these approaches, however, have gone so far in exploiting the analogy of energy and information entropy.

Herein lies much of the attraction of the free energy principle. It promises no less than to unify biology and psychology under the same thermodynamic principles. But it is exactly these principles that cause havoc for the theory. The second law of thermodynamics requires that if order is created internally to minimize free energy, an equal or larger amount of warmth (or free energy, or disorder) must be dissipated to its environment. Estimates for the upper bounds of energy dissipation in biological systems exist (Skinner and Dunkel, 2021). But what would be such dissipation in the informational analogue of free energy? Perhaps the immense amounts of nonsense spouted on social media may count as such? More seriously, to prevent such harm to the outside world and vice versa, Markov blankets isolate the interior brain from the exterior world. This, allows the non-equilibrium steady-state of minimal surprise to persist, but in the informational version only. So the pretended universality of this approach appears to be a case of bait and switch.

This notwithstanding, at least internally, rationality has been restored to the system: rule following behavior was initially replaced with satisficing, and now has made a comeback with the principle of free energy minimization. This principle restores rule following behavior at computational level, in the form of attractor dynamics. The theory promotes random and fragile attractors. This allows the system to show complex dynamical trajectories when stochastically perturbed, and wander chaotically amongst the various wings of the attractor (Tsuda, 2001)). Because it allows for complex attractor structures and chaotic itinerancy, inflexibility is not a problem for such systems.

But is this behavior ecologically rational? If brains compute, they must compute online to meet the immediate demands of navigating their environ-

ment. Attractors cannot be reached in short time. This means the approach is unsuitable for online computing. Transient computation may be more suitable in that case (Rabinovich et al., 2008). Transients galore in chaotic itinerancy. But when they do the computational work, whence the need for a global attractor minimizing surprise?

Early criticisms of this approach have pointed out that such a principle may be limited in its ability to explain exploratory behavior (Van Leeuwen, 1990). We may have to allow for the possibility that living systems actively seek surprise. This is needed for enterprise, exploration and discovery. The same may be true for the brain. Exploration is needed for making new discoveries in creative invention (Verstijnen et al., 2000).

Later critiques (e.g. (Di Paolo et al., 2022)) have emphasized the incompatibility of the free energy principle as applied here, and embodied cognitive science (Varela et al., 1992), in particular the enactive approach. According to Di Paolo, "These tensions have to do with how the enactive approach conceives of agents as precarious, self-constituted entities in ongoing historical development and capable of incorporating different sources of normativity throughout their development, a world-involving process that is co-defined with their environment across multiple spatiotemporal scales and together with other agents." (p.3). The enactive approach argues that our mental life is found at this ecological level, rather than hiding in the brain under a Markov blanket.

It will be clear that underlying these tensions are differences in how we consider the human cognition: as enclosed within its organism, mainly engaged in ordering its own attic or as a person, individually and collectively engaging with their environment. Humans typically vacillate between such states: exploitation and exploration. We observe this kind of everyday behavior, but encounter it even in the laboratory, for instance in the perception of visual scenes (Nikolaev et al., 2023). Perhaps such cycles, rather are relevant to how brain dynamics should be understood. It remains to be seen if the notions of chaotic itinerancy and the free energy principle are versatile enough to explain this behavior.

10 Position

Given the above observations, how likely is it that theories of human intelligence (or: cognition) will, within any reasonable amount of time, reach a level of maturity such that we can actually explain — or maybe even predict — a person's decisions?

Not very likely, it would seem. The rule-based explanations of the 1950s, 60s, and 70s all had their fallacies — either from a philosophical or an empirical standpoint — and do not hold the explanatory power we need to *truly* understand why or how people make decisions, or classify sensory instances. Do the more contemporary models, rooted in thermodynamics, entropy, time series and attractors then provide for more explainability? Hardly. Even though these models have (some) biological validity, and the promising 'fragile' attractor models do seem to answer the 'how' question, at least partially, the *explanation* of 'why' still eludes. Worse still, we might never capture it. Many dynamical systems exhibit chaotic behaviour, which in some cases is unpredictable (Moore, 1990; Werndl, 2009) — for the same reason the local weather is unpredictable: the unpredictability is a property of the system *itself*; any forward projection of the system will separate exponentially fast from the real state (Dingwell, 2006).

With these thoughts in mind, one cannot help but consider this: human intelligence is opaque to this extent, then what justification do we have trying to explain artificial intelligence? Why do we generally trust a medical diagnosis from human doctor better than the same diagnosis from an AI algorithm – even when the latter performs better? (Amann et al., 2020; Longoni et al., 2019)?

The answer might be partially because it's *new*. Resistance to new technology has persisted since the dawn of time; some famous examples include nuclear power, information technology and biotechnology (Bauer, 1995). AI, explainable or unexplainable, has recently made its way into our daily lives, and is rapidly gaining ground. To what extent 'explanantion' should be seen as our generations' resistance, or our trouble getting accustomed to the new reality will likely be answered by future generations. Maybe Max Planck's famous quote is a good way to conclude this position: "[A new scientific truth does not triumph by convincing people and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it.]" (Planck, 1949).

REFERENCES

Agarwal, C., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., and Lakkaraju, H. (2022). Openxai: Towards a transparent evaluation of model explanations. *Advances in Neural Information Processing Systems*, 35:15784–15799.

Alsayegh, F., Alkhamis, M. A., Ali, F., Attur, S., Fountain-Jones, N. M., and Zubaid, M. (2022). Anemia or

other comorbidities? using machine learning to reveal deeper insights into the drivers of acute coronary syndromes in hospital admitted patients. *Plos one*, 17(1):e0262997.

Amann, J., Blasimme, A., Vayena, E., Frey, D., and Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, 20(1):1–9.

Ariza-Garzón, M. J., Arroyo, J., Caparrini, A., and Segovia-Vargas, M.-J. (2020). Explainability of a machine learning granting scoring model in peer-to-peer lending. *Ieee Access*, 8:64873–64890.

Arras, L., Osman, A., and Samek, W. (2022). Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40.

Astrand, E., Ibos, G., Duhamel, J.-R., and Hamed, S. B. (2015). Differential dynamics of spatial attention, position, and color coding within the parietofrontal network. *Journal of Neuroscience*, 35(7):3174–3189.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549.

Bauer, M. W. (1995). *Resistance to new technology: nuclear power, information technology and biotechnology*. Cambridge university press.

Behravan, H., Hartikainen, J. M., Tengström, M., Kosma, V.-M., and Mannermaa, A. (2020). Predicting breast cancer risk using interacting genetic and demographic factors and machine learning. *Scientific reports*, 10(1):11044.

Belle, V. and Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in big Data*, page 39.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204.

Clark, B., Wilming, R., and Haufe, S. (2023). Xai-tris: Non-linear benchmarks to quantify ml explanation performance. *arXiv preprint arXiv:2306.12816*.

Das, A. and Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*.

Di Paolo, E., Thompson, E., and Beer, R. (2022). Laying down a forking path: Tensions between enaction and the free energy principle. *Philosophy and the Mind Sciences*, 3.

Dijksterhuis, A. and Nordgren, L. F. (2006). A theory of unconscious thought. *Perspectives on Psychological science*, 1(2):95–109.

Dingwell, J. B. (2006). Lyapunov exponents. *Wiley encyclopedia of biomedical engineering*.

Elshawi, R., Al-Mallah, M. H., and Sakr, S. (2019). On the interpretability of machine learning-based model for predicting hypertension. *BMC medical informatics and decision making*, 19(1):1–32.

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81.

Friedman, J. H. and Popescu, B. E. (2008). Predictive learning via rule ensembles.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138.

Gabbay, F., Bar-Lev, S., Montano, O., and Hadad, N. (2021). A lime-based explainable machine learning model for predicting the severity level of covid-19 diagnosed patients. *Applied Sciences*, 11(21):10417.

Gigerenzer, G. (2007). *Gut feelings: The intelligence of the unconscious*. Penguin.

Gigerenzer, G., Hell, W., and Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance*, 14(3):513.

Güss, C. D. (2018). What is going through your mind? thinking aloud as a method in cross-cultural psychology. *Frontiers in psychology*, 9:1292.

Jaarsveld, S. and van Leeuwen, C. (2005). Sketches from a design process: Creative cognition inferred from intermediate products. *Cognitive science*, 29(1):79–101.

Jansen, T., Geleijnse, G., Van Maaren, M., Hendriks, M. P., Ten Teije, A., and Moncada-Torres, A. (2020). Machine learning explainability in breast cancer survival. In *Digital Personalized Health and Medicine*, pages 307–311. IOS Press.

Kahneman, D. (2011). *Thinking, fast and slow*. macmillan.

Kaisla, J. et al. (2001). Rationality and rule following: On procedural and consequential interests of the rule-guided individual. Technical report, Department of Industrial Economics and Strategy, Copenhagen Business School.

Keane, M. T. and Smyth, B. (2020). Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai). In *Case-Based Reasoning Research and Development: 28th International Conference, ICCBR 2020, Salamanca, Spain, June 8–12, 2020, Proceedings 28*, pages 163–178. Springer.

Kim, H. S. (2002). We talk, therefore we think? a cultural analysis of the effect of talking on thinking. *Journal of personality and social psychology*, 83(4):828.

King, J.-R. and Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in cognitive sciences*, 18(4):203–210.

Kuzlu, M., Cali, U., Sharma, V., and Güler, Ö. (2020). Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools. *IEEE Access*, 8:187814–187823.

Kwon, Y. and Zou, J. Y. (2022). Weightedshap: analyzing and improving shapley based feature attributions. *Advances in Neural Information Processing Systems*, 35:34363–34376.

Leventi-Peetz, A.-M. and Weber, K. (2022). Rashomon effect and consistency in explainable artificial intel-

ligence (xai). In *Proceedings of the Future Technologies Conference*, pages 796–808. Springer.

Liu, Y., Khandagale, S., White, C., and Neiswanger, W. (2021). Synthetic benchmarks for scientific research in explainable machine learning. *arXiv preprint arXiv:2106.12543*.

Longoni, C., Bonezzi, A., and Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, 46(4):629–650.

Loriette, C., Amengual, J. L., and Ben Hamed, S. (2022). Beyond the brain-computer interface: Decoding brain activity as a tool to understand neuronal mechanisms subtending cognition and behavior. *Frontiers in Neuroscience*, 16:811736.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

MacKay, D. J. (1995). Free energy minimisation algorithm for decoding and cryptanalysis. *Electronics Letters*, 31(6):445–447.

Magesh, P. R., Myloth, R. D., and Tom, R. J. (2020). An explainable machine learning model for early detection of parkinson's disease using lime on datscan imagery. *Computers in Biology and Medicine*, 126:104041.

Mei, Y., Chen, Q., Lensen, A., Xue, B., and Zhang, M. (2022). Explainable artificial intelligence by genetic programming: A survey. *IEEE Transactions on Evolutionary Computation*.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Moncada-Torres, A., van Maaren, M. C., Hendriks, M. P., Siesling, S., and Geleijnse, G. (2021). Explainable machine learning can outperform cox regression predictions and provide insights in breast cancer survival. *Scientific reports*, 11(1):6968.

Moore, C. (1990). Unpredictability and undecidability in dynamical systems. *Physical Review Letters*, 64(20):2354.

Morris, M. D. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2):161–174.

Nieuwenstein, M. R., Wierenga, T., Morey, R. D., Wicherts, J. M., Blom, T. N., Wagenmakers, E.-J., and van Rijn, H. (2015). On making the right choice: A meta-analysis and large-scale replication attempt of the unconscious thought advantage. *Judgment and Decision Making*, 10(1):1–17.

Nikolaev, A. R., Ehinger, B. V., Meghanathan, R. N., and van Leeuwen, C. (2023). Planning to revisit: Neural activity in refixation precursors. *Journal of Vision*, 23(7):2–2.

Nikolaev, A. R., Jurica, P., Nakatani, C., Plomp, G., and Van Leeuwen, C. (2013). Visual encoding and fixation target selection in free viewing: presaccadic brain potentials. *Frontiers in systems neuroscience*, 7:26.

Planck, M. (1949). Scientific autobiography and other papers, trans. *F. Gaynor (New York, 1949)*, pages 33–34.

Rabinovich, M., Huerta, R., and Laurent, G. (2008). Transient dynamics for neural processing. *Science*, 321(5885):48–50.

Rao, R. P. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87.

Rao, S., Mehta, S., Kulkarni, S., Dalvi, H., Katre, N., and Narvekar, M. (2022). A study of lime and shap model explainers for autonomous disease predictions. In *2022 IEEE Bombay Section Signature Conference (IBSSC)*, pages 1–6. IEEE.

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.

Schneider, W. and Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. detection, search, and attention. *Psychological review*, 84(1):1.

Schwitzgebel, E. (2008). The unreliability of naive introspection. *Philosophical Review*, 117(2):245–273.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Simon, H. (1978). Rationality as process and as product of thought. *American Economic Review*, 68(2):1–16.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological review*, 63(2):129.

Simon, H. A. (1982). Models of bounded rationality, vols. 1 and 2. *Economic Analysis and Public Policy, MIT Press, Cambridge, Mass*.

Simon, H. A. and Ericsson, K. A. (1984). Protocol analysis: Verbal reports as data. *(No Title)*.

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

Skinner, D. J. and Dunkel, J. (2021). Improved bounds on entropy production in living systems. *Proceedings of the National Academy of Sciences*, 118(18):e2024300118.

Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1-3):271–280.

Stanovich, K. E. and West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23(5):645–665.

Trajanov, R., Dimeski, S., Popovski, M., Korošec, P., and Eftimov, T. (2022). Explainable landscape analysis in automated algorithm performance prediction. In *International Conference on the Applications of Evolutionary Computation (Part of EvoStar)*, pages 207–222. Springer.

Treisman, A. M. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136.

Tsuda, I. (2001). Toward an interpretation of dynamic neural activity in terms of chaotic dynamical systems. *Behavioral and brain sciences*, 24(5):793–810.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

Van Leeuwen, C. (1990). Perceptual-learning systems as conservative structures: is economy an attractor? *Psychological Research*, 52(2-3):145–152.

Van Stein, B., Raponi, E., Sadeghi, Z., Bouman, N., Van Ham, R. C., and Bäck, T. (2022). A comparison of global sensitivity analysis methods for explainable ai with an application in genomic prediction. *IEEE Access*, 10:103364–103381.

Van Stein, B., Vermetten, D., Caraffini, F., and Kononova, A. V. (2023). Deep bias: Detecting structural bias using explainable ai. In *Proceedings of the Companion Conference on Genetic and Evolutionary Computation*, GECCO '23 Companion, page 455–458, New York, NY, USA. Association for Computing Machinery.

Varela, F. J., Thompson, L., and Rosch, E. (1992). The embodied mind: Cognitive science and human experience.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Verstijnen, I., Van Leeuwen, C., Hamel, R., and Hennessey, J. (2000). What imagery can't do and why sketching might help. *Empirical studies of the arts*, 18(2):167–182.

Werndl, C. (2009). What are the new implications of chaos for unpredictability? *The British Journal for the Philosophy of Science*.

Wong, B. Y. and Jones, W. (1982). Increasing metacomprehension in learning disabled and normally achieving students through self-questioning training. *Learning Disability Quarterly*, 5(3):228–240.

Yang, J. (2021). Fast treeshap: Accelerating shap value computation for trees. *arXiv preprint arXiv:2109.09847*.

Yeung, C., Tsai, J.-M., King, B., Kawagoe, Y., Ho, D., Knight, M. W., and Raman, A. P. (2020). Elucidating the behavior of nanophotonic structures through explainable machine learning algorithms. *ACS Photonics*, 7(8):2309–2318.

Zhou, S., Pfeiffer, N., Islam, U. J., Banerjee, I., Patel, B. K., and Iquebal, A. S. (2022). Generating counterfactual explanations for causal inference in breast cancer treatment response. In *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, pages 955–960. IEEE.