Towards useful interpretability for medical imaging

Angus Nicolson

Linacre College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2024

So long, and thanks for all the fish.

— The Dolphins *The Hitchhiker's Guide to the Galaxy by Douglas Adams*

Acknowledgements

I've learnt a lot over the last 4 years, but I wouldn't have been able to do it without the support of those around me. First, a big thanks to my two supervisors, Alison Noble and Yarin Gal, for their guidance throughout my time at Oxford. Their advice in helping me shape my DPhil work and my own approach to scientific work has been invaluable. One of the benefits of having two supervisors, is I get to belong to two research groups. Both the Noble and OATML labs have welcomed me with open arms and I truly appreciate all the feedback and discussions with you all. However, some names deserve a particular mention. Thanks to Mohammad Alsharid, Jong Kwon and Divyanshu Mishra for always being willing to chat inside or outside the working day. Thanks to Elizabeth Savochkina, Alexander Gleed and Netzahualcoyotl Hernandez-Cruz for organising so many brilliant social events. Thanks to Josh Strong, for putting up with my constant badgering as I think of something to say and, with the unfortunate position of sitting next to me, you get an earful shortly afterwards. Thanks to Pramit Saha for always being willing to talk all things ML. A big thanks to Lisa Schut for making me feel welcome in the OATML group and really helping me push my work forwards. I could not have asked for a better collaborator. Elizabeth Bradburn also deserves my heartfelt appreciation for her time and effort working on the AGE study and just being a joy to work with and chat to during meetings. Thanks to Shreshth Malik for organising the OATML Hackathon 2024, without that push TextCAVs would probably not exist! And a big thanks to Katrina Dickson for just sorting stuff when it needs sorting and your help in proof reading this thesis. Thanks to the members and directorate of the Health Data Science CDT, which I thoroughly enjoyed being a part of. And I can't forget the BioMedia lunch crew who helped me spend hours waiting in line for fish and chip Fridays, rather than writing this thesis.

On a more personal note, I really appreciate the support my friends, neighbours and family have provided throughout my DPhil. My family may not always have a complete grasp on what I do, but that doesn't stop them from asking about it, which I very much appreciate. Along with their unconditional love and support in all things I set my mind to. A special thanks belongs to my partner, Hannah, who has been there for both the highs, where I spend waaaay too long explaining my latest work or new research idea, and the lows, with paper rejects and the uncertainty inherent in doing novel research. I would not have been able to do this without you.

Abstract

Interpretability, in the context of deep learning, is the ability to explain a model to a human. It is proposed as a solution to many tasks including: debugging, improving user trust, finding model bias, and scientific discovery. Hence, there is increasing demand for new interpretability methods, particularly in high-risk scenarios such as in healthcare. However, measuring a method's ability to explain a model is a complex, or even impossible task. Instead, we propose to measure its usefulness. To measure a method's usefulness, we must first define its specific use. As part of its use we must specify the type of user, e.g., engineers debugging a model, clinicians diagnosing a patient or patients receiving an automated report.

In this thesis, we improve the understanding and evaluation of interpretability methods with specific use cases in mind. First, we demonstrate the importance of understanding the limitations of a method before using it to interpret models. We define and examine three properties of a popular concept-based interpretability method, providing tools to understand when these properties can cause misleading explanations and demonstrate how they affect a melanoma classification task. Next, we propose a novel concept-based interpretability method that requires no labelled concept data and demonstrate how engineers can use it to debug a chest X-ray classification model or detect maliciously implanted trojans in ImageNet models. Finally, we show the importance of human studies in understanding the effects of interpretability on users. We perform a reader study evaluating the effect of explanations on the trust, reliance and performance of sonographers using a prototype-based interpretable model for gestational age estimation. We show that, although explanations are generally assumed to improve trust, they can reduce trust (and performance) if the model explanation does not match the internal decision making of the users.

Interpretable deep learning is an exciting emerging field and, through embracing thorough evaluation and defining clear goals of what explanations aim to achieve, we can further the development of novel, useful methods.

Contents

[1]/finding biases/ensuring fairness

List of Figures

List of Abbreviations

AI	Artificial intelligence.
XAI	Explainable artificial intelligence.
ML	Machine learning.
NN	Neural network.
CNN	Convolutional neural network.
LLM	Large language model.
CAV	Concept Activation Vector.
TCAV	Testing with Concept Activation Vectors [116].
NH1, NH2, NH3	Null hypothesis 1, 2 or 3.
CLIP	Contrastive Language-Image Pre-Training [176].
SaTML	The Secure and Trustworthy Machine Learning Conference.
EU	European Union.
GDPR	General Data Protection Regulation. An EU regulation.
FDA	U.S. Food and Drug Administration.
MHRA	The UK's Medicines and Healthcare products Regulatory Agency.
MLMD	Machine learning medical device.
CDSS	Clinical decision support system. A system used to augment clinicians when making complex decisions by providing targeted clinical knowledge or patient information.
EBM	Evidence based medicine.
CRS	Concept relevance score.
GA	Gestational age.
TAF	Tumor adipose feature. A prognostic histological feature [226].
ProtoPNet	Prototypical part network [45].
VIA	VGG Image Annotator [63].

The beginning is the most important part of the work.

— Plato *The Republic*

1 Introduction

Contents

1.1 Motivation

Deep learning models have been shown to be powerful tools within healthcare, and in imaging are able to achieve performances that can be similar to or surpass domain experts [183, 228, 117, 50]. These models can improve clinician performance when used as advice in clinical decision making [117, 141]. Despite the evidence that deep learning models can be beneficial, to date their translation into clinical practice is rare [91, 198, 213, 136, 20, 195, 155]. Some researchers have suggested this is because we have little or no ability to understand how models reach their decision – so called "black boxes" – and this may hamper trust in model predictions [115, 22, 146, 185]. To overcome this, interpretable deep learning models have been proposed; here explanations are provided alongside model predictions, so that trust by end users is enhanced and more detail is given to aid clinical decision making, making clinicians better able to understand how a decision was made [204, 14, 212, 55, 181].

Some argue that reliability is often enough to build trust [62], with a model that has been through enough validation being deemed trustworthy, especially if it makes similar mistakes to humans [133, 143]. However, many argue it is a *requirement* in high-risk domains such as healthcare [91, 43, 222], with statements such as: "explainability is [...] an essential requirement for people to trust and adopt AI deployed in numerous domains" [130]. However, this often seems to be done with little or no thought as to what the purpose of those explanations are [133, 134], or, if a goal – such as improving clinicians' trust in the model – is stated, little or no effort is put in to measure if the explanations achieve that goal. In this thesis we challenge this mindset and, rather than specifying interpretability as a *requirement*, we aim to understand, design and evaluate interpretability methods which can be *useful* in the healthcare domain.

To measure a method's usefulness, we must first define its specific use. Interpretability has been proposed as a solution to a large range of tasks including: debugging a model (finding biases, ensuring fairness) [123, 80, 17, 112, 39, 158], improving user trust [115, 80, 22, 100, 16, 224, 6], improving user performance [140, 15, 100, 16, 6], scientific discovery [140, 174, 65, 184, 121, 192, 145], as a legal requirement [82, 1, 41, 68] and understanding the learning process [144, 154, 44]. If interpretability methods are designed to be useful for these tasks, then they should be evaluated on these tasks. As part of this evaluation, it is important to define the type of user as they can differ greatly in expertise and requirements [60], e.g., engineers developing a model, clinicians diagnosing a patient, or patients receiving an automated report.

In this thesis, we improve understanding of and evaluation of interpretability methods with specific use cases in mind. First, in Chapter 3, we examine a technique which many popular concept-based interpretability methods rely on: concept activation vectors (CAVs) [116]. We mathematically define three key properties: layer consistency, entanglement and spatial dependence, providing both empirical experiments and (in some cases) theoretical proofs that these properties apply to CAVs. We then provide tools to understand when these properties can

cause misleading explanations and recommendations to mitigate their effect. We demonstrate how to implement these recommendations on a melanoma classification task, providing a guide for practitioners to use in their own CAV-based experiments. Next, in Chapter 4, we propose a novel concept-based interpretability method (TextCAVs [158]) that requires no labelled concept data and demonstrate how engineers can use it to debug an ImageNet model with implanted trojans [39] and a chest X-ray classification model. Finally, in Chapter 5, we perform a reader study evaluating 10 sonographers using a prototype-based interpretable model, prototypical part network (ProtoPNet) [45], for gestational age estimation. Our results highlight the necessity of human studies in interpretability. We find that, although explanations are generally assumed to improve trust, they can reduce trust (and performance) if the model explanation does not match the internal decision making of the user. Our study provides important insights into the decision making process of both users and AI, and the interplay between the two.

Our work demonstrates the importance of defining *why* an interpretability method should be used. By defining a use, we can evaluate a method. Metrics that are commonly used to develop, evaluate and compare different interpretability methods *cannot* reveal the different responses that users can have, which require user studies with clearly articulated goals. As the field of interpretable deep learning grows, rigorous evaluation is essential. We argue that by clearly defining what explanations aim to achieve, the field can reduce growing pains and develop accurate, novel and, most importantly, *useful* methods.

1.2 Thesis Outline

All contributions to this thesis are original work completed by Angus Nicolson, unless otherwise stated.

Chapter 1: Introduction This chapter.

Chapter 2: Background Here, we provide background and context that is relevant to the whole thesis. Related work that is relevant to individual chapters is left for said chapters.

Chapter 3: Understanding CAVs We define and examine three properties of a popular concept-based interpretability method, providing tools to understand when these properties can cause misleading explanations and demonstrate how they affect a melanoma classification task. This chapter is based on the following paper:

> **Angus Nicolson**, Lisa Schut, J. Alison Noble and Yarin Gal. Explaining Explainability: Recommendations for Effective Use of Concept Activation Vectors, *TMLR* 2025. [159]

Chapter 4: TextCAVs We propose a novel concept-based interpretability method that uses joint vision-language models to create CAVs with no labelled concept data. We demonstrate how engineers can use it to debug a chest X-ray classification model and find trojans (maliciously implanted bugs) in ImageNet. This chapter is based on the following papers:

> **Angus Nicolson**, Yarin Gal and J. Alison Noble. TextCAVs: Debugging vision models using text, *iMIMIC Workshop at MICCAI* 2024. [158]
>
> Stephen Casper, Jieun Yun, Joonhyuk Baek, Yeseong Jung, Minhwan Kim, Kiwan Kwon, Saerom Park, Hayden Moore, David Shriver, Marissa Connor, Keltin Grimes, **Angus Nicolson**, Arush Tagade, Jessica Rumbelow, Hieu Minh Nguyen, and Dylan Hadfield-Menell. The SaTML'24 CNN Interpretability Competition: New Innovations for Concept-Level Interpretability. *arXiv*:2404.02949, 2024. [39]

The paper by Casper et al. [39] contains the results from a competition at the Safe and Trustworthy Machine Learning (SaTML) conference and only the entry by A. Nicolson is work submitted for this thesis. All algorithm development, and experimentation was completed by A. Nicolson, but human evaluations were completed by S. Casper.

Chapter 5: Do explanations help sonographers? We perform a reader study evaluating how a prototype based interpretable model affects sonographers' trust, reliance and performance on the task of gestational age estimation from fetal head ultrasound. Although explanations are generally assumed to improve trust in deep learning models, paradoxically, we show they can reduce trust, while still increasing appropriate reliance. We also show the variable nature of the response to interpretability, with some clinicians performing better with explanations and others worse. Our study provides a template for researchers aiming to learn the effects of explanations on clinician behaviour. This chapter is based on the following papers:

> **Angus Nicolson***, Elizabeth Bradburn* Yarin Gal, Aris T. Papageorghiou and J. Alison Noble. The Human Factor in Explainable Artificial Intelligence: Clinician Variability in Trust, Reliance, and Performance, *npj Digital Medicine* 2025 [157]
>
> **Angus Nicolson**, Yarin Gal and J. Alison Noble. Sparse Explanations for Gestational Age Prediction in Fetal Brain Ultrasound, *IMLH Workshop at ICML* 2022.

Chapter 6: Conclusion We conclude the thesis, providing an overview of our work and promising directions for future research.

If you wish to make an apple pie from scratch,
you must first invent the universe.

— Carl Sagan *Cosmos*

2 Background

Contents

2.1 Introduction

In this chapter we discuss the background and context required to understand this thesis. Although the focus of this thesis is on deep learning and interpretability for imaging applications, we include literature from other domains to broaden the discussion. First, in § 2.2, we give an introduction to interpretability, where we define key concepts and give an overview of the field. The next section (§ 2.3)

[1] /finding biases/ensuring fairness

describes current methods for evaluating interpretability, typically by splitting it into different components that can be measured more easily than interpretability as a whole. In § 2.4, rather than measuring aspects of interpretability, we discuss evaluation of interpretability through its different uses. In each case, this involves reviewing literature that proposes the use, and (the rarer) literature that evaluates it. To conclude, in § 2.5, we discuss the implications and gaps in the literature our review has revealed.

2.2 Interpretability

Interpretability is often poorly defined in the literature [60], so before we discuss interpretable deep learning, we define some terms:

- **Interpretability** is defined as *the ability to explain or present in understandable terms to a human* [60]
- **Faithfulness** is how accurately an explanation describes true model behaviour
- A **global** explanation provides the features which a model tends to use, i.e. the average response of the model [133]
- A **local** explanation gives the model's reasoning for a specific prediction [133]

The literature often uses synonyms or conflicting definitions of the same words. For example, **explainability**, **transparency**, **comprehensibility** and **understandability** are often used instead of interpretability [142, 88, 59, 42], **black box** instead of uninterpretable, **white box** instead of interpretable, **explainable AI** (**XAI**) instead of interpretable AI or interpretable machine learning, and **fidelity** instead of faithfulness. In this thesis we will use each synonym interchangeably and follow the definitions above.

Interpretable methods for deep learning fall into two different categories:

- **Post-hoc** methods explain models *after* they have been trained

- **Interpretable-by-design** models are designed to be interpretable as a core part of their functionality

Post-hoc methods are desirable because they allow state-of-the-art networks to be explained, but it can be difficult to prove if the explanations are faithful to the underlying model [3]. Interpretable-by-design models in deep learning are a newer concept and have the benefit of inbuilt faithfulness [185], but methods are not as widely available or proven to reach state-of-the-art performance in many tasks.

Standard deep learning models such as convolutional neural networks (CNNs) or transformers are uninterpretable. Their predictions come with no explanation and the internals of the model are too complex for a human to understand. The purpose of interpretability research is to provide an explanation for these models. The most common form of interpretability used in medical imaging are saliency maps [218, 28] which use the gradient of logit values with respect to pixels and compute the derivative [12, 202, 209, 203, 238]:

$$\frac{\partial h_k(\boldsymbol{x})}{\partial \boldsymbol{x}_{a,b}} \tag{2.1}$$

where $h_k(\boldsymbol{x})$ is the logit output for class k and image $\boldsymbol{x}$, and $\boldsymbol{x}_{a,b}$ is the pixel at position (a, b) in $\boldsymbol{x}$. The gradient is a measure of the sensitivity of an output class to changes in a specific pixel, valid in a small neighbourhood around the input. Note that this is not an explanation for how the model made its prediction, but a measure of which regions of the image would most change the prediction if the input changed by an infinitesimal amount. This detail is commonly discarded and saliency maps are treated as the region a model is looking at. There have been many developments to basic gradient-based saliency aiming to provide less noisy or more fine-grained explanations, including: DeepLift [200], GradCAM [194], Integrated Gradients [209], Guided Backpropagation [207], SmoothGrad [203]. However, there are major limitations with some saliency methods, with many of the commonly used methods having been shown to be unfaithful, meaning the explanations do not match the reasoning of the underlying model [3]. Even when the methods

are faithful, they only show the regions a model is sensitive to, not the regions a model used or *how* they were used. For example, in a fetal brain ultrasound a saliency approach might highlight the upper right section of the brain/skull border. However, what characteristics of this region is the model sensitive to? The curvature of the skull, the shape of structures within the brain, the texture of the brain, or an artifact that is present in that region? Using saliency, we cannot determine this.

Although saliency is currently the most popular form of interpretability, there are other types of interpretability which use different 'cognitive chunks' [125], i.e., rather than using pixels to explain a model, they use something else. This includes case-based reasoning [45] or counterfactual methods [191] which explain models using whole images, or concept-based methods which explain a model's response with respect to semantically meaningful concepts [116, 78, 49, 158]. Concept-based methods have been shown to provide plausible explanations for both natural images (e.g. ImageNet [56, 116]) and medical imaging [83, 84, 73, 229, 158]. One of the first concept-based methods (that many subsequent methods are based on) is Testing with Concept Activation Vectors (TCAV) [116], which we describe in detail in § 3.2. One advantage of these more complex interpretability methods is that they provide explanations in a form more similar to how a human might explain something. Although, now that we have all these different methods, how do we compare them?

2.3 Evaluating Interpretability

There is no consensus on how to express explanation quality and no one evaluation metric that can be applied to all explanation methods [37, 245]. This is partly because interpretability is an inherently subjective concept, with an explanation's requirements depending on the user, explanation method and the use case of the model. For example, most researchers would state that simple models such as linear models, or decision trees are interpretable and complex models, like neural networks, are uninterpretable. However, Lipton [133] argues that "neither linear models, rule-based systems, nor decision trees are intrinsically interpretable" as at sufficiently high dimensions they can be harder to interpret than a compact neural network.

We agree with this sentiment and with Nauta et al. [156] that interpretability is "a multi-faceted characteristic" and should not be treated as a binary property, so statements that a specific model type is inherently interpretable are not possible.

Doshi-Velez and Kim [60] define three types of evaluation: application-grounded, human-grounded and functionally-grounded. Application-grounded metrics involve conducting human experiments where the real use case of the model is being performed. Human-grounded evaluations are simpler experiments where specific aspects of the target application are tested, but not by performing the real task. For example, interpretability methods can be compared by presenting pairs of explanations to the user and asking which is better. Functionally-grounded evaluations are often more appealing because they require no human experiments and instead use some formal definition of interpretability which can be mathematically represented. For example, measures of model size can represent the simplicity of a model (and its explanation) - one aspect of interpretability [245].

Given these well categorised approaches to evaluation, we might expect a diverse and rich set of evaluation tools used in practice. This is not the case. Only 1 in 5 papers introducing new interpretability methods evaluate with users and 1 in 3 papers evaluate exclusively using anecdotal evidence [156] where, for example, they simply examine a set of explanations that look reasonable [150]. This lack of robust evaluation has led to multiple calls for objective quantitative methods for comparing the interpretability of different methods [60, 231, 125, 224, 156, 108]. Anecdotal inspection of explanations is not sufficient for the confirmation of whether a method is interpretable which, instead, requires different aspects of interpretability to be evaluated [3, 125, 126, 156]. Nauta et al. [156] split interpretability into 12 measurable properties, the Co-12: correctness, output-completeness, consistency, continuity, contrastivity, covariate complexity, compactness, composition, confidence, context, coherence, and controllability. They then review 29 quantitative functional evaluation methods which each relate to one or several of these properties. A complete review of these methods is outside the scope of this thesis, but we will

briefly discuss coherence, completeness, compactness, and correctness as the majority of papers evaluating interpretability methods focus on these properties [156].

Coherence describes how well the explanations match prior knowledge or beliefs. This addresses concerns such as reasonableness [80], plausibility [106] and "agreement with human rationales" [11]. It is often evaluated using anecdotal evidence where example explanations are compared to expectations from domain experts [156]. We evaluate coherence in Chapter 4 when we demonstrate that TextCAVs produce reasonable explanations for both natural images (ImageNet [56]) and chest X-rays (MIMIC-CXR [109, 110]).

Completeness is a measure of the proportion of the model's behaviour that is described by the explanation. This is often done by determining the model's performance while masking important features (according to the explanation) [220]. In Chapter 5, we measure completeness using the proportion of the logit output that our explanation accounts for.

Compactness describes the size of the explanation. Other terms commonly used here are sparsity, or simplicity. Depending on the model this is calculated in different ways, and is often not comparable across models. For example, we might say a decision tree has a depth of 10 and an average path length of 5.2 and a prototype-based model uses 100 prototypes with an L1 loss of 12.7, but a direct comparison of these numbers does not reveal which method is simpler. Instead, comparisons can be made within model types to say one tree is more compact than another or one prototype method uses less prototypes. We measure and optimise for compactness extensively in Chapter 5 and Appendix B.2 in order to ensure that we can have high explanation completeness when only 4 prototypes are displayed to a user. Our results highlight how interconnected these properties are, with compactness and completeness being almost opposites in definition but, when only partial explanations can be displayed, maximising compactness can increase completeness.

Correctness is a measure of how faithful the explanation is to the ground-truth reasoning of the model. This is sometimes described as faithfulness [106], as we did at the beginning of this chapter. There is an important distinction between

correctness and coherence as explanations that looks reasonable (coherence) do not necessarily describe the underlying behaviour of the model (correctness). For example, Adebayo et al. [3] showed that some saliency methods act more like edge detectors, showing reasonable explanations, but when the behaviour of the model was changed by randomising layer weights, the explanations barely changed. If the explanations do not change when model behaviour changes, then they cannot be explaining the model behaviour and the explanations are not correct (faithful). This is a particular problem with recent developments and widespread usage of large language models (LLMs). LLMs readily generate explanations for their decisions, but, with reinforcement learning from human feedback (RLHF), they have been optimised for plausibility and not correctness [169]. This means the explanations can fool users into believing them, even if they are not the underlying reasoning of the model [215, 5].

Nauta et al. [156] finish their review on functional interpretability metrics by stating that "trade-offs between desired explanation properties will have to be made when developing an XAI method" and that "application domain or practical feasibility can determine which Co-12 properties should be emphasized". In other words, an individual interpretability method cannot be the *best* in all cases, and to know which interpretability method is most suitable, its use case will need to be defined. In the next section, we discuss these uses and alternative forms of more direct evaluation.

2.4 Why use interpretability?

In the introduction, abstract and even the title of this thesis we emphasise the need to determine if an interpretability method is *useful*. However, to measure if something is useful, we must first define its use, i.e. its purpose. Interpretability methods provide explanations for deep learning models, so the question is more, *where would explanations aid in the performance of a task?* In this section, we discuss the different uses. In each case, we review the literature for papers which either propose or evaluate interpretability methods for that specific use.

2.4.1 Debugging[2]

Many authors propose using interpretability for debugging a model [80, 17, 64, 222]. In fact, Bhatt et al. [23] found that debugging was the most common use of interpretable ML in deployment. By debugging we mean the discovery of harmful biases within a model. Once a bias is discovered, different techniques are then required to remove them, like model editing [147, 18] or simply retraining with a different version of the dataset. We consider the related use of ensuring fairness as equivalent, as ensuring a model is fair across different attributes is the same as determining if the model is biased towards or against that attribute.

A clear example of this is from the authors of GradCAM [194], where they show that a model trained to predict if a person is a doctor or a nurse is biased. Because of the gender imbalance in their training set sourced from the internet (93% of nurses were women and 78% of doctors were men), the model learnt to use a person's face and hairstyle to estimate if they were a doctor/nurse. By using GradCAM, the authors discover this bias and then edit the training set accordingly by adding more examples of male nurses and female doctors. However, it is worth noting that this could have been done with a cursory examination of the dataset.

Deep learning models cheat. They are optimised to minimise their training loss and will find any solution to achieve that goal, even if that requires using features a human would consider a 'shortcut' [76]. Zech et al. [236] used saliency methods to determine that a deep learning model trained to detect pneumonia in chest radiographs was using spurious features to make its prediction. When performing a scan, radiology technicians place a metal token on the patient and the model was using the differences in these tokens across hospitals to determine where the image was taken. As the hospital a patient was admitted to was correlated with pneumonia, the model learnt to use this information to drive its predictions, even though these features have no causal effect on pneumonia and led to poor performance in unseen hospitals [236]. Similarly, several papers have used saliency methods to show that dermatology models use surgical markings in their predictions [171, 225].

[2]/finding biases/ensuring fairness

These examples are relatively benign, in that there is no intentional effort to create a model with undesirable behaviour. Instead, there was simply an imbalance in the training dataset which led the model to use features that will not generalise to clinical practice. However, it has been shown that novel features [96] (not present in the training set), or maliciously designed attacks can cause targeted effects in model predictions. The most well known attack is by using adversarial examples [211], where an image can be imperceptibly altered, but still change the model's prediction. Interpretability tools can help us discover these malicious attacks. Casper et al. [38] examine using different feature visualisation tools to discover human-interpretable adversarial attacks, known as trojans [48]. They designed a competition [39] where users evaluated different feature visualisation tools ability to discover trojans. We elaborate on their methodology of evaluation and our entry to the competition, TextCAVs [158], in Chapter 4.

2.4.2 Trust and Reliance

Many researchers have stated that one of the primary uses of interpretability is its ability to build trust in an AI model [208, 115, 182, 80, 22, 89, 100, 17, 204, 14, 222] with statements such as "explainability is essential for users to effectively understand, trust, and manage powerful artificial intelligence applications" [89] and explainability is "an essential requirement for people to trust and adopt AI deployed in numerous domains" [130]. This is particularly present among health research as researchers believe "one of the major barriers to integrating ML models into clinical workflows is a lack of trust among clinicians" [222].

However, confusing this story, there are conflicting definitions of trust relating to whether it is a belief, attitude, intention, or behaviour [127]. In this section we discuss two related but distinct ideas: trust and reliance. For trust, we use the definition from Lee and See [127]: "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability". Whereas, in our context, reliance is the extent to which a model influences an individual's behaviour. Trust is an attitude, whereas reliance is a behaviour. Trust

guides, but does not completely determine, reliance – as we show in Chapter 5, where explanations cause a decrease in participants' self-reported trust but a small increase in their reliance on the model. Trust, being subjective, is hard to quantify but it can affect user uptake and acceptability of AI in clinical practice [181]. Reliance, however, is more objective and has strong parallels with simply improving performance of the human-AI team (§ 2.4.3), since a human that relies on an accurate model will perform better [16].

Improving trust in AI models is not necessarily beneficial. Over-reliance on AI, where humans accept the output of an algorithm regardless of its correctness, is a well documented phenomena [127, 101, 8, 32, 36, 120], including studies showing that explanations may not improve and can even cause over-reliance [34, 241, 31, 32, 105, 164]. There is evidence to suggest that people are more likely to delegate decisions to AI than to another human [36], but at the same time, evidence to suggest humans trust algorithms less than other humans [58]. In the interpretability field specifically, data scientists were found to "over-trust and misuse interpretability tools" and most were unable to accurately describe the visualisations produced by popular interpretability methods [112]. Instead of blind trust we want *appropriate reliance* [127] where explanations can help humans know when to rely an AI's suggestion and when to ignore it [16], i.e., users rely on the model when it is correct and do not rely on the model when it is incorrect. For a more formal definition see § 5.2.3. Appropriate reliance is vital for users to be able to use AI explanations to improve task performance [24, 224].

Although improved trust is almost always described as a goal in interpretability papers, most interpretability papers do not include a human evaluation study (as we discussed in § 2.3), and so cannot evaluate trust. The effect of interpretable machine learning on trust and reliance has only recently begun to be studied and not always with a positive result [105, 57, 6]. For example, Ahn et al. [6] found that interpretability "led to no robust improvements in trust" and that outcome feedback, where users are told their performance and the model accuracy after each task, had a greater impact. Wang and Yin [224] perform a review of user studies

evaluating the effect of interpretability on appropriate reliance (they name this trust calibration) and find 9 studies which evaluate reliance, but only two of these report appropriate reliance: Zhang et al. [241] found that displaying model confidence was more effective than local feature-based explanations at calibrating people's trust in an AI model (when estimating income); whereas, Yang et al. [230] found that image-based explanations improved users' appropriate reliance. Although, similar to the results of Ahn et al. [6], they found even greater effects when combined with outcome feedback. With only two examples in the literature, Wang and Yin [224] ran three user studies evaluating appropriate reliance themselves. They found that appropriate reliance was affected differently depending on both the specific interpretability method and the task the users were performing. For example, SHAP [139] improved appropriate reliance for recidivism prediction but no methods improved appropriate reliance for forest cover estimation. These studies demonstrate that the jury is still out on whether explanations can help achieve appropriate reliance, with conflicting results depending on the specific model, explanation method, and domain. What we can be sure on, is that the only conclusive method of determining if a specific interpretability method aids users in having appropriate reliance on a model is to perform the human studies to measure it.

Some research has cast doubt on the use of subjective measures of trust and preference to evaluate XAI methods. In a study using a simulated AI (i.e. participants were told it was an AI, but researchers had chosen the model responses), Buçinca et al. [31] found that questionnaire responses on participant trust and preference were poor predictors of which explanation method improved the performance of the human-AI team. The authors highlight that, in relying on these metrics, "our field may be inadvertently slowing its progress toward developing human+AI teams that can reliably perform better than humans or AIs alone". We agree that researchers need to be careful not to be mislead by the metrics we use, but we frame our argument in a different manner. Rather than subjective measures of trust being misleading, we think they are measures of trust, not of appropriate reliance. Questionnaires on trust cannot be used to infer improvement, or lack of,

in the human-AI team, because that is not what they measure, but they can be used to understand the attitudes of the participants. To summarise, researchers need to ensure they use metrics for the purpose they are designed, rather than to infer related, but separate properties of the system.

So far, this discussion on trust has not involved the healthcare domain, but we now move on to discuss clinical decision support systems (CDSSs). CDSSs are used to augment clinicians when making complex decisions by providing targeted clinical knowledge or patient information [210]. They are the primary method of utilising AI in clinical practice where the focus is on providing advice to a clinician, rather than completely automating a task. In providing explanations, Bussone et al. [34] aimed to reduce over-reliance on CDSSs. The authors found that giving simple explanations made participants question the system's reliability and led to self-reliance, i.e. the users ignored the advice, but, if more complex explanations were provided, there were issues with over-reliance. Similarly, Panigutti et al. [164] found that clinicians using an XAI-based CDSS perceived the explanations to be of low quality but still relied on the AI advice more often when explanations were provided. Jacobs et al. [105] presented 220 clinicians with patient vignettes and asked them to prescribe antidepressants with/without the presence of ML recommendations and explanations. The authors found that the ML recommendations did not improve clinician accuracy in treatment selection and that explanations caused clinicians to agree with the model when it was incorrect, i.e. over-reliance. These works highlight the danger of explanations causing over-reliance in the healthcare domain and that user studies need to be designed to measure the effect of explanations, rather than assuming they will be useful.

When designing user studies, researchers need to ensure the relevant baselines are performed and enough results reported to be able to make claims about the effect of explanations on trust and reliance. For example, Desolda et al. [57] performed a study with two versions of an AI-based CDSS for rhinocytology where explanations were/were not displayed and found there were no significant differences in reliance between the systems. It was a comprehensive study, involving an initial interview to

guide the design of the explanations and then evaluation on 9 clinicians, including post-task interviews, but they did not report any quantitative measures based on the performance of the human-AI team. The reported data was all from the results of the questionnaire, not, for example, an analysis of how often the humans corrected the AI. This means their measures of reliability, were about the attitudes of the participants, rather than their behaviour during the study. This information was available to the authors, but they did not report it. Similarly, Sabol et al. [187] conducted a user study on the diagnosis of colorectal cancer from histopathological images using XAI and found that pathologists preferred using the system with explanations than without. However, they did not evaluate the performance of the participants without the aid of the model, so it is not possible to determine the participants reliance on the AI advice. The authors partially offset this by including a questionnaire on the reliability of the two approaches[3], but a measurement of self-reported reliability indicates something about the participants attitude towards the method, not their behaviour. Du et al. [61] create a model for the prediction of gestational diabetes mellitus. Using weight of advice (WoA) [90, 81] (a quantitative metric of reliance that we also use in Chapter 5), the authors measured participants' reliance on two different interpretability methods, but they did not evaluate the effect of the predictions by themselves. This makes it impossible to infer the relative contributions of explanations and predictions to reliance. Folke et al. [71] performed a study evaluating whether radiologists have appropriate trust in an XAI model developed for pneumothorax diagnosis. Participants were shown a target image and asked if they would certify the AI for images similar to the target. Two different explanations were used: (1) saliency maps and (2) saliency maps combined with example images that the model successfully/unsuccessfully predicted. The authors found that the participants were more likely to certify the AI for images where the model was correct than incorrect, indicating appropriate trust, however, once again, no experiments were performed without explanations so we do not know if the clinicians would have had appropriate trust without the explanations.

[3]the participants thought the XAI method was more reliable

Some authors have found that factors other than the presence of explanations can have a greater influence on reliance. For example, Shafti et al. [197] provide a quantitative measure of reliance in the form of the response shift of participants with/without explanations. They examined the effect of good/poor predictions when good/poor explanations are provided. In this case, a good quality explanation was defined as an explanation for the correct model, e.g. if the explanations for the good model were displayed with predictions from the poor model, then the explanations had poor quality. The authors find that the performance of the model has a greater influence on reliance than how good the explanations are. They attribute this to participants developing more trust in the good model after witnessing its high accuracy during a training period, where participants had the chance to see some example cases. Ahn et al. [6] observed similar results, with outcome feedback having a greater influence on participants than interpretability methods, although their focus was on participant performance, so we discuss this paper in more detail in the next section.

Measuring trust is a complex endeavour. To even start measuring it, researchers need to make clear distinctions between measuring participants' trust (attitude) and reliance (behaviour). Interpretability researchers need to move beyond claiming their methods improve trust without providing evidence and realise that a blind improvement in trust can be detrimental to performance by leading to over-reliance. There are conflicting results relating to the effect of explanations on reliance but future work needs to determine which methods can provide *appropriate reliance*, leading to improved performance, and long-term trust in AI models. This involves the careful design of user studies, ensuring any results can support the necessary claims.

2.4.3 Performance

As discussed in the previous section, appropriate reliance can lead to improved performance in human-AI teams, but there is another mechanism for explanations to improve user performance. In providing more information to end-users, explanations

have the capability of increasing user performance for human-in-the-loop systems when the goals of the human and AI are not perfectly aligned. As an example, take the CDSS in [117] which provides a risk score for lung nodules in CT scans. In clinical practice, the clinician needs to make a decision on how to treat the patient, e.g. do they perform a followup scan in 6-months, a year, or do they need to perform a biopsy immediately. If the model provided explanations for why it gave a risk score then the additional information could help the clinician decide which treatment option is most suitable. This is distinct from appropriate reliance as the decision the clinician is making is different to the task the model is performing (treatment decisions rather than risk estimation) and the improvement is due to additional information rather than from better calibrated trust/reliance.

For now, this use case remains mostly hypothetical as we do not know of any studies explicitly examining increased performance due to the additional information provided by XAI. However, the benefit of explanations on performance has long been discussed, with Gregor and Benbasat [86] in 1999 stating that "explanations, when suitably designed, have been shown to improve performance" and, in the domain of healthcare, explanations could be used "to enhance doctors' predictive capacity" [14]. In the remainder of this section we will review papers which discuss and evaluate XAI's ability to improve performance, regardless of the exact mechanism.

As in the previous section, we highlight the need for careful study design when evaluating the effect of explanations on human performance. Lundberg et al. [140] demonstrated improved performance in an XAI reader study on the prevention of hypoxaemia during surgery with five practicing anaesthesiologists. However, although their study evaluated the performance of clinicians alone and of clinicians with the XAI advice, they did not evaluate the effect of solely the model predictions on clinician performance. This means we cannot separate the effect of the predictions and the explanations on performance and so it is unknown if the explanations provide additional benefit. In addition, there are repeated claims that the explanations allow the anaesthesiologists to plan better interventions but there is no evidence for this claim. If the purpose of the explanations is to allow for better interventions,

then the study should be designed to determine how clinicians' interventions change, rather than how clinicians' predictions of hypoxaemia change.

Bansal et al. [16] examine the effect of explanations on human-AI team performance. Several previous works [140, 35, 173] demonstrate explanations utility in improving user performance, but in each case the AI model was substantially better at the task than the humans. This means that any affect of increasing reliance would increase the performance of the human-AI team compared to the performance of humans in isolation. The question Bansal et al. [16] aim to answer is if explanations can aid human-AI teams of comparable performance to each other, where each partner can fix some of the mistakes of the other, leading to improved performance compared to either individually. This is *appropriate reliance*, as described in § 2.4.2. They found that although the model predictions improved performance, explanations did not provide any additional benefit, and, as discussed in the previous section, led to over-reliance on the model. This work highlights how the relative performance of AI and humans can have a substantial impact on the requirements of explanations, with models that are much more accurate than humans needing to *convince*, whereas models that are of similar performance needing to *collaborate*. For example, Poursabzi-Sangdeh et al. [173] test human performance at real-estate valuation when given access to different machine learning models that are substantially better than the humans alone. In this scenario, if explanations convince the human to rely on the model, then the performance of the human-AI team will increase. In this case, they found that explanations did not increase reliance, and prescribed this lack of improvement due to information overload. Their work "emphasize[s] the importance of testing over intuition when developing interpretable models" and the need for human based studies in interpretable model evaluation.

Ahn et al. [6] perform a multi-head reader study analysing the change in performance caused by model predictions, local explanations, global explanations, and outcome feedback. The study examined the prediction of the likelihood of couples going on a second date and used an XGBoost model [47] with LIME [182] to provide both global and local explanations. They found that outcome

feedback was better than either interpretability approach which only produced "modest effects on participants' task performance". This work is a great example of interpretability methods being compared against other approaches to determine the best way to solve a specific task. In this case, improved performance of the human-AI team. More studies like this should be performed in the medical domain, bringing together different sub-fields of machine learning research to determine the relative merits of different approaches.

2.4.4 Scientific Discovery

Deep learning models have achieved beyond human capabilities in a variety of domains and tasks [92, 201, 135, 228]. Is it possible to use interpretability to extract *how* these models achieve this? If we can, then we can learn something about the task itself that can be taught to humans and increase our scientific understanding.

Messeri and Crockett [145] discuss how the use of AI in science "risks introducing a phase of scientific enquiry in which we produce more but understand less". This is due to the use of deep learning models to *do* science, rather than *understand* science. A deep learning oracle which can perfectly predict the outcome of experiments does not provide perfect understanding of the underlying science [121]. Interpretability has been proposed as a solution to this problem, as with AI increasingly being applied to the natural sciences [184], interpretable methods can be used to extract knowledge from these models [65].

One clear example of using interpretability to further human understanding is in the domain of chess. Schut et al. [192] use concept-based methods to search for novel chess concepts that AlphaZero [201] – a super-human level reinforcement learning chess algorithm – uses but expert humans do not. The authors show that these concepts can be taught to grandmasters, demonstrating that human-understandable knowledge can be gained from AI models leading to performances beyond what humans were previously capable of.

This is obviously desirable in the healthcare domain as greater understanding of disease allows for better design of new medicines and treatment. Wulczyn et al.

[226] demonstrate this possibility. They trained a deep learning model for survival prediction in colorectal cancer patients and, using a combination of interpretability approaches and collaboration with human experts, identified a novel prognostic histological feature, tumor adipose feature (TAF). In later work, human pathologists were able to learn to use the TAF feature as an effective prognostic marker without the need of the model [131].

Further examples exist within the field of drug discovery. One of the uses of deep learning in drug discovery is to provide a virtual pre-screening of candidate molecules before synthesis and testing begins [10]. This greatly reduces the search space of laboratory tests and the expected cost of discovering effective treatments. Recently, interpretability has been proposed as a means of understanding how these models perform drug discovery [174, 107, 186]. For example, Ruffolo et al. [186] use an attention mechanism to examine physically important residual pairs in the structure prediction of antibodies, or Preuer et al. [174] use single neuron classifiers (similar to [19]) and Integrated Gradients [209] to extract chemical substructures that are responsible for the molecules' biological properties (pharmacophores). Preuer et al. [174] show that the highlighted pharmacophores match structures in the literature known to be biologically active, demonstrating that interpretable methods can be use to provide insight into the biological properties of candidate molecules proposed by deep learning models.

Currently, there are few examples of interpretability being used to discover novel ideas, but as deep learning models are increasingly used across all domains of science, the use of interpretability to elucidate *how* these models achieve their success will push the boundaries of human understanding.

2.4.5 Legality

Although there has been academic discussion on the regulation of AI as the methods become more successful, with initiatives such as the Montreal Declaration on Responsible AI at NeuriPS 2018, the advent of powerful, general purpose large language models (LLMs) like ChatGPT [30], has seen more urgent progress. In

2023, at the first international AI safety summit in the UK, the first international declaration on AI was signed by 29 countries, including the USA, EU, UK, Australia, Japan and China stating that AI poses a potentially catastrophic risk to humanity [217]. The first international legally binding treaty was signed a year later, with the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law [69]. However, although these international treaties are designed for safe AI, they do not refer to interpretability specifically.

Currently, the only law that might *require* interpretability is the 'right to explanation' in the General Data Protection Regulation (GDPR) of the European Union [82], which requires companies to provide explanations for any algorithms that make a decision about a user when asked. However, it is unclear how this is regulated in practice and whether partial explanations from interpretable deep learning methods will be adequate.

Although GDPR might be the only law requiring interpretability, there are multiple recent regulations or guidelines that elude to the preference of transparent models to enable human oversight of AI [40, 1, 68]. Including, regulations currently being voted on in California to add requirements to the development of large AI models by setting out various testing, safety, and enforcement standards [193].

In 2021, Health Canada, the U.S. Food and Drug Administration (FDA) and the United Kingdom's Medicines and Healthcare products Regulatory Agency (MHRA) identified guiding principles for good machine learning practice [40]. In 2024, they built upon these guidelines with a focus on transparency [41]:

- principle 7: Focus is placed on the performance of the human-AI team.
- principle 9: Users are provided clear, essential information.

The term "transparent" has a broader meaning than we have been using in this thesis and describes "the degree to which appropriate information about a machine learning medical device (MLMD) (including its intended use, development, performance and, when available, logic) is clearly communicated to relevant audiences". The key phrase related to interpretability is the reference to "logic" as it refers to

information about how an output or result was reached or the basis for a decision or action. The document even makes it clear that "logic and explainability are aspects of transparency". The guidelines state that it is "valuable to provide the basis of a device output or other information that explains how the MLMD reaches its output" but falls short of suggesting it is a requirement as logic need only be provided when "this information is available and easily understood"[4].

President Biden's executive order on the Safe, Secure, and Trustworthy Development and Use of AI [1] and the EU's AI Act [68] also elude to the preference of interpretable AI. However, it is not clear how these oversight mechanisms will be implemented in practice and many catch-alls are used (e.g. "as appropriate and proportionate" or "possibly including regulatory action") making it unclear whether AI explanations (i.e. interpretability) will be a requirement. For example, article 14 of the AI Act relates to human oversight and section 14(4)(c) states:

> For the purpose of implementing paragraphs 1, 2 and 3, the high-risk AI system shall be provided to the deployer in such a way that natural persons to whom human oversight is assigned are enabled, as appropriate and proportionate: (c) to correctly interpret the high-risk AI system's output, taking into account, for example, **the interpretation tools and methods available**;

Where paragraphs 1, 2 and 3 relate to the requirement that high-risk AI is developed with "appropriate human-machine interface tools", that human oversight shall minimise the risks to health and safety, and that oversight measures should be in place in proportion to the level of risk. Similarly, President Biden's executive order [1] states:

> Independent regulatory agencies are encouraged, as they deem appropriate, [...] to consider rulemaking [...] and emphasizing or **clarifying requirements and expectations related to the transparency of AI models** and regulated entities' ability to explain their use of AI models.

[4]The reader may agree that deep learning interpretability may not qualify for "easily understood".

Overall, the regulatory picture on interpretable AI is unclear, and although it seems doubtful that it is a current legal requirement, it is clear from the language that transparent, interpretable AI is preferred to the current black box paradigm. Time will only tell how governments choose to regulate AI moving forwards, but regulators should be careful not to require an impossibility, with current state of the art interpretable methods only providing partial explanations for deep learning models and not a complete causal description of their behaviour. On the other hand, regulators need to push back on claims that AI is completely safe, with some claims made about interpretability research being close to nonsensical. For example, we finish this section on a statement given as evidence to the UK's House of Lords inquiry into LLMs by the Venture Capital firm, a16z [102]:

> Although advocates for AI safety guidelines often allude to the "black box" nature of AI models, where the logic behind their conclusions is not transparent, recent advancements in the AI sector have resolved this issue, thereby ensuring the integrity of open-source code models.

2.4.6 Learning Dynamics

This is the least studied use of interpretability which we look at in this thesis but we think it an interesting recent development of the field. The core idea is that "certain insights into model behavior may only be accessible by observing the trajectory of the training process" [44] and through using interpretability we can study the dynamics of deep learning. One of the first examples of this approach was in the study of how different chess concepts are learnt throughout training of AlphaZero [144]. For example, the authors show that the model initially uses a variety of opening moves, but, as learning progresses, it rapidly prioritises similar openings to that of humans. More recently, Nanda et al. [154] fully reverse engineer how one-layer transformers implement modular addition, and use this knowledge to explain grokking – when a neural network generalizes suddenly after overfitting on a data set. In a similar manner, Chen et al. [44] analyse sudden drops in loss in masked language models by studying the emergence of syntactic attention structure – a phenomena where specific transformer heads tend to focus on specific

syntactic relations. Olsson et al. [163] suggest that induction heads are responsible for in-context learning in large transformer models and demonstrate that induction heads develop at precisely the same point as a sudden sharp increase in in-context learning ability. These studies demonstrate the ability of interpretability techniques to elucidate the source of model performance, teaching us how deep learning models learn, and aiding the development of novel architectures or learning processes.

Three of these examples [163, 154, 44] belong to the sub-field of mechanistic interpretability (mech-interp). Mech-interp involves reverse-engineering AI models to create human-understandable algorithms and concepts. The goal is to gain a precise understanding of how AI systems work, including their internal reasoning processes. This is clearly related to our work, but an in-depth discussion is outside the scope of the thesis, instead we refer interested readers to recent reviews and some of the initial papers in the area [162, 66, 67, 153, 177, 21].

2.5 Summary & Forward Outlook

As we have seen in this chapter, there are many possible uses of interpretable deep learning in the domain of medical imaging and in healthcare more generally. There have been a number of researchers calling for interpretable models in the healthcare domain [181, 46, 168, 108, 42, 28, 29], with the number of papers relating to XAI in medical imaging increasing year on year accordingly [42]. However, terminology is still an issue with disagreements over definitions of words and multiple terms with identical meanings [42][5]. There is also an issue with evaluation, with papers stating that "applying XAI in clinical settings requires proper evaluation criteria to ensure the explanation technique is both technically sound and clinically useful" [108], but, only 22% of papers which introduce novel interpretability methods using human studies, and only 23% of those evaluating with domain experts [156]. This means only 5% of papers evaluated their method on a real task with domain experts, i.e. application-grounded evaluation [60]. Similar results were found in a review on interpretable AI in medical image analysis by Chen et al. [46], where an evaluation on

[5] e.g. XAI vs interpretable model

end users occurred in only 3 of the 68 studies. They summarised their findings with "current techniques in transparent ML are dominated by computational feasibility and barely consider end users". We agree with Champendal et al. [42] that "future XAI development should consider user needs and perspectives" and champion recent works that achieve this. For example, Cai et al. [35] interviewed 21 pathologists and found that rather than local explanations on specific patient decisions, they desired "upfront information about basic, global properties of the model, such as its known strengths and limitations, its subjective point-of-view, and its overall design objective – what it's designed to be optimized for". This was attributed to similar collaborative mental models they develop of their medical colleagues when seeking a second opinion. In this thesis, we push similar user-driven research forward with our use case based evaluation of TextCAVs in Chapter 4 (debugging models) and ProtoPNet in Chapter 5 (improving trust, reliance and performance). While Chapter 3 does not focus on a specific use case to the same extent, it enhances current understanding of a widely used concept-based method, and provides clear recommendations to practitioners aiming to use CAVs for their own interpretability use cases.

Any fool can know. The point is to understand.

— Albert Einstein

3 Understanding CAVs

Contents

In order for interpretability methods to be useful, we do not just need to change how we evaluate them, but also understand the advantages and limitations of each method. Once a method has been shown to produce plausible explanations, there is a temptation to immediately start applying it to any domain where it could be useful. Instead, we should carefully evaluate the assumptions upon which the

method rests and determine any situations in which we may not be able to rely on it.

In this chapter, we focus on understanding CAVs [116], the underlying representation of concepts used in a variety of concept-based interpretability methods. See § 3.2 for a detailed description. CAVs and Testing with CAVs, TCAV [116], have seen extensive use with over $2,000$ citations[1] since 2017, and they have inspired a wide range of follow-up work that also uses vectors to represent human-interpretable concepts [116, 72, 244, 78, 239, 179, 70].

Despite their popularity, CAVs are built on several assumptions that can constrain their applicability and affect the reliability of the explanations they produce. First, the method assumes *linearity* in the representation space: a concept must correspond to a linear direction, which may not hold for all concepts or layers [54]. Second, CAVs are inherently *binary*, relying on examples that either contain or do not contain the concept; this limits their expressiveness in tasks where concepts are more continuous or graded [83]. Third, as discussed extensively in this chapter, they are *sensitive to the choice of layer*: different layers in a neural network capture different levels of abstraction, and the learned CAV may vary substantially depending on the chosen layer, introducing instability or interpretive ambiguity.

Prior research has also shown that CAVs can be sensitive to the specific probe dataset used to train them [178, 205]. These limitations underscore the need to critically assess when and how to apply CAV-based methods.

In this chapter we provide an in-depth analysis of three properties of CAVs that can lead to misleading explanations: layer consistency, concept entanglement, and spatial dependence. We ground this in reality by providing practical recommendations of how to determine if these properties might be affecting results and an example of how to implement these recommendations in a melanoma classification task. Overall, this chapter acts as a guide for researchers and practitioners using CAVs, aiming to clarify when and how they might lead to misleading explanations and to mitigate their limitations in applied settings.

[1] as of 4$^{\text{th}}$October 2024

3.1 Introduction

Deep learning models have become ubiquitous, achieving performance reaching or surpassing human experts across a variety of tasks. However, currently, the inherent complexity of these models obfuscates our ability to explain their decision-making process. As they are applied in a growing number of real-world domains, there is an increasing need to understand how they work. This transparency allows for easier debugging and better understanding of model limitations.

Model explanations can take many forms, such as input features, prototypes or concepts. Recent work has shown that explainability methods that focus on low-level features can incur problems. For example, saliency methods can suffer from confirmation bias and lack model faithfulness [3]. Even when faithful, saliency maps only show 'where' the model focused in the image, and not 'what' it focused on [2, 52].

To address these problems, concept-based methods provide explanations using high-level terms that humans are familiar with. A popular method is concept activation vectors (CAVs): a linear representation of a concept found in the activation space of a specific layer using a probe dataset of concept examples [116]. However, concept-based methods also face challenges, such as their sensitivity to the specific probe dataset [178, 205].

In this chapter, we focus on understanding three properties of concept vectors:

1. They cannot be **consistent** across layers,
2. They can be **entangled** with other concepts,
3. They can be **spatially dependent**.

We provide tools to analyse each property and show that they can affect testing with CAVs (TCAV) (§3.6.1, §3.6.2 and §3.6.3) and lead to misleading explanations. To minimise the impact these effects can have, we recommend: creating CAVs for multiple layers, verifying expected dependencies between related concepts, and visualising spatial dependence (§3.7). These properties do not imply that CAVs should not be used. On the contrary, we may be able to use these properties

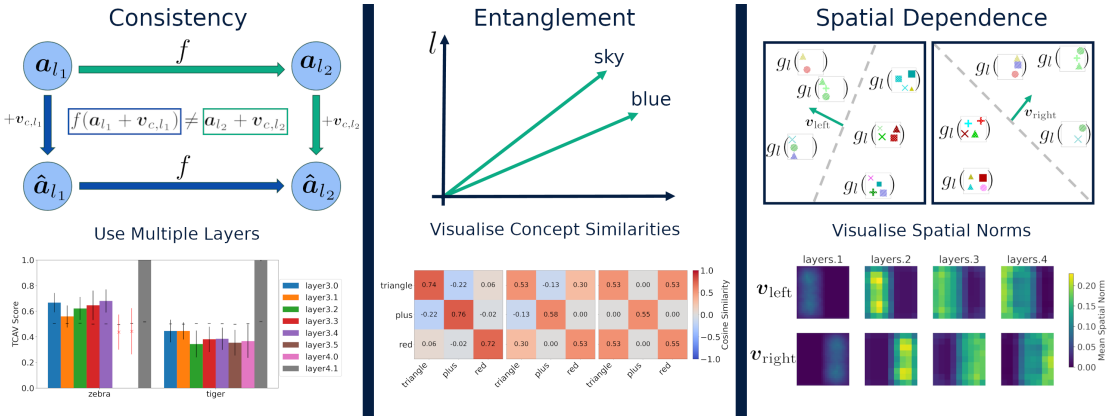


Figure 3.1: Concept Activation Vectors can be: inconsistent across layers, i.e., we cannot find two concept vectors in different layers that have the same additive effect (left), entangled (middle) and spatially dependent (right). The top panel illustrates each of these different properties. The bottom panels show our recommendations on how to minimise the impact these effects can have: creating CAVs for multiple layers (left), verifying expected dependencies between related concepts (middle), and visualising spatial dependence (right).

to better understand model behaviour. For example, we introduce a modified version of CAVs that are spatially dependent and can be used to identify translation invariance in convolutional neural networks (CNNs).

To provide a concrete example of how the properties can affect experiments, we examine the use-case of Yan et al. [229] which uses CAVs in the context of skin cancer diagnosis (§ 3.7). We demonstrate how to use our recommendations to sanity-check TCAV results and, in this specific case, we show that entangled concepts lead to uninterpretable explanations. Our results also indicate that the choice of negative probe dataset can have a substantial impact to the meaning of a CAV.

To help explore these properties, we created a configurable synthetic dataset: Elements (§3.4). This dataset provides control over the ground-truth relationships between concepts and classes in order to understand model behaviour. Using the Elements dataset, researchers can study (1) the faithfulness of a concept-based explanation method and (2) the concept entanglement in a network.

3.2 Background: Concept Activation Vectors

A CAV [116] is a vector representation of a concept found in the activation space of a layer of a NN. Both Chapters 3 & 4 make extensive use of CAVs and so we describe them in detail here.

Consider a NN which can be decomposed into two functions: $g_l(\boldsymbol{x}) = \boldsymbol{a}_l \in \mathbb{R}^m$ which maps the input $\boldsymbol{x} \in \mathbb{R}^n$ to a vector $\boldsymbol{a}_l$ in the activation space of layer l, and $h_l(\boldsymbol{a}_l)$ which maps $\boldsymbol{a}_l$ to the output. To create a CAV for a concept c we need a probe dataset $\mathbb{D}_c$ consisting of positive samples $\mathbb{X}_c^+$ (concept examples), and negative samples $\mathbb{X}_c^-$ (random in-distribution images). For the sets $\mathbb{X}_c^-$ and $\mathbb{X}_c^+$, we create a corresponding set of activations in layer l:

$$\mathbb{A}_{c,l}^+ = \{g_l(\boldsymbol{x}_i) \quad \forall \boldsymbol{x}_i \in \mathbb{X}_c^+\}, \text{ and } \mathbb{A}_{c,l}^- = \{g_l(\boldsymbol{x}_i) \quad \forall \boldsymbol{x}_i \in \mathbb{X}_c^-\}, \tag{3.1}$$

We find the CAV $\boldsymbol{v}_{c,l}$ by training a binary linear classifier to distinguish between the sets $\mathbb{A}_{c,l}^+$ and $\mathbb{A}_{c,l}^-$:

$$\boldsymbol{a}_l \cdot \boldsymbol{v}_{c,l} + b_{c,l} > 0 \quad \forall \boldsymbol{a}_l \in \mathbb{A}_{c,l}^+, \text{ and } \boldsymbol{a}_l \cdot \boldsymbol{v}_{c,l} + b_{c,l} \leq 0 \quad \forall \boldsymbol{a}_l \in \mathbb{A}_{c,l}^-, \tag{3.2}$$

where $\boldsymbol{v}_{c,l}$ is the normal vector of the hyperplane separating the activations $\mathbb{A}_{c,l}^+$ and $\mathbb{A}_{c,l}^-$, and $b_{c,l}$ is the intercept.[2]

To analyse a model's sensitivity to $\boldsymbol{v}_{c,l}$, Kim et al. [116] introduce testing with CAVs (TCAV), which determines the model's conceptual sensitivity across an entire class. Let $\mathbb{X}_k$ be a set of inputs belonging to class k. The TCAV score is defined as

$$\text{TCAV}_{c,k,l} = \frac{|\{\boldsymbol{x} \in \mathbb{X}_k : S_{c,k,l}(\boldsymbol{x}) > 0\}|}{|\mathbb{X}_k|}, \tag{3.3}$$

where the directional derivative of the concept, $S_{c,k,l}$, defined as

$$S_{c,k,l}(\boldsymbol{x}) = \lim_{\epsilon \to 0} \frac{h_{l,k}\left(g_l(\boldsymbol{x}) + \epsilon \boldsymbol{v}_{c,l}\right) - h_{l,k}\left(g_l(\boldsymbol{x})\right)}{\epsilon} = \nabla h_{l,k}\left(g_l(\boldsymbol{x})\right) \cdot \boldsymbol{v}_{c,l} \tag{3.4}$$

where $\nabla h_{l,k}$ is the partial derivative of the NN output for class k to the activation. The TCAV score measures the fraction of class k inputs whose activation at layer l is positively influenced by concept c.

[2] Eq. 3.2 assumes that the linear classifier has hard boundaries. In practice, the classifiers typically achieve 80-95% accuracy.

If the probe dataset consists of random images, it is still possible to find a CAV, but it is unlikely to have any meaningful interpretation to humans. Therefore, a statistical test comparing the scores of CAVs to random vectors is used to determine the concept's significance. The test compares a set of CAV scores found using a concept dataset with CAV scores found using random data. To do this, we must find multiple CAVs for each concept. In practice, each of these CAVs is trained with the same positive set, $\mathbb{X}_c^+$, but a different random set, $\mathbb{X}_c^{r-}$, where $r \in 1, 2 \dots R$ denotes the random index. A CAV corresponding to a specific random index is labelled $\boldsymbol{v}_{c,l}^r$.

3.3 CAV Hypotheses

To use CAV-based explanation methods in practice, it is important to understand how they work. Therefore, we study three properties of CAVs and their effects on TCAV scores. We focus on these hypotheses as they provide insight into network representations and into the meaning encoded by concept vectors.

We formalise each property through a null hypothesis, which we provide evidence to reject later in the paper. In the following text, we use the typesetting `concept` to denote a concept.

3.3.1 Layer Consistency

In general, we want to understand *model* behaviour. However, CAVs explain whether a model is sensitive to a concept in a specific *layer*. In practice, analysing all layers may be computationally infeasible, and it is unclear which layers to choose. Therefore, our first hypothesis explores the relationship between CAVs found in different layers. Recall that the TCAV scores depend on the directional derivative: *how the model output changes for an infinitesimal change of the activations in the direction of a CAV*. By perturbing the activations in the direction of a CAV, we explore whether two concept vectors found in different layers can have the same affect on the model output. We refer to this property as *layer consistency* (see Figure 3.1 for a schematic overview).

Definition 1 (layer consistency) *Assume we have a function $f(\cdot)$ that maps the activations from layer l_1 into activations in layer l_2, where $l_1 < l_2$. Concept vectors, $\boldsymbol{v}_{c,l_1}$ and $\boldsymbol{v}_{c,l_2}$ are consistent across layers iff for every input $\boldsymbol{x}$ and corresponding activations $\boldsymbol{a}_{l_1}$ and $\boldsymbol{a}_{l_2}$, $f(\boldsymbol{a}_{l_1} + \boldsymbol{v}_{c,l_1}) = \boldsymbol{a}_{l_2} + \boldsymbol{v}_{c,l_2}$.*

If two CAVs are consistent across layers then they have the same downstream affect on the model when activations are perturbed in their direction, i.e., even though they are in different layers, they have an equivalent effect on the model output and therefore the model assigns them the same meaning. Our first hypothesis is:

> ***Null Hypothesis 1 (NH1)**: Concept vector representations are consistent across layers*

In §3.6.1 we formally explore this hypothesis, and perform empirical evaluations on the Elements and ImageNet [56] datasets. We show theoretically the conditions $\boldsymbol{v}_{c,l_2}$ and $\boldsymbol{a}_{l_1}$ must meet for vectors $\boldsymbol{v}_{c,l_1}$ and $\boldsymbol{v}_{c,l_2}$ to be consistent when f is either a rectified linear unit (ReLU) or sigmoid function.

3.3.2 Entangled concept vectors

Consider the meaning encoded by a concept vector. We label a CAV using the corresponding label of the probe dataset. For example, a CAV may be labelled `striped` or `red`. This implicitly assumes that the label is a complete and accurate description of the information encoded by the vector. In practice, the CAV may represent several concepts – e.g., continuing the example above, the vector may encode both `striped` and `red` simultaneously. We refer to this phenomenon as *concept entanglement*. Mathematically, we formulate this as follows. A concept vector $\boldsymbol{v}_{c,l}$ is more similar to the activations corresponding to images containing the concept than activations for images not containing the concept, i.e. it satisfies

$$\boldsymbol{a}^{+}_{c,l} \cdot \boldsymbol{v}_{c,l} > \boldsymbol{a}^{-}_{c,l} \cdot \boldsymbol{v}_{c,l} \quad \forall \boldsymbol{a}^{+}_{c,l} \in \mathbb{A}^{+}_{c,l}, \boldsymbol{a}^{-}_{c,l} \in \mathbb{A}^{-}_{c,l}. \tag{3.5}$$

Assume we have concepts c_1 and c_2, with probe datasets $\mathbb{D}_{c_1}$ and $\mathbb{D}_{c_2}$, respectively. For each probe dataset, we find the activation sets: $\mathbb{A}_{c_1,l} = \{A^{+}_{c_1,l} \cup A^{-}_{c_1,l}\}$ and $\mathbb{A}_{c_2,l} = \{\mathbb{A}^{+}_{c_2,l} \cup \mathbb{A}^{-}_{c_2,l}\}$.

Definition 2 (entangled concepts) *A CAV* $\boldsymbol{v}_{c_1,l}$ *for concept* c_1 *is entangled with concept* c_2 *iff*

$$\boldsymbol{a}^{+}_{c_2,l} \cdot \boldsymbol{v}_{c_1,l} > \boldsymbol{a}^{-}_{c_2,l} \cdot \boldsymbol{v}_{c_1,l} \quad \forall \boldsymbol{a}^{+}_{c_2,l} \in \mathbb{A}^{+}_{c_2,l}, \boldsymbol{a}^{-}_{c_2,l} \in \mathbb{A}^{-}_{c_2,l} \tag{3.6}$$

Our second hypothesis explores concept entanglement:

> ***Null Hypothesis 2 (NH2)****: A CAV represents only the concept corresponding to the concept label of its probe dataset*

If concepts are entangled, it is not possible to separate the model's sensitivity to one concept from its sensitivity to related concepts – therefore, if we measure the TCAV score for c_1, we will unknowingly incorporate the effect of c_2.

In §3.6.2 we provide a visualisation tool to explore CAV entanglement and discuss how this can affect TCAV.

3.3.3 Spatial Dependence

Here, we explore the influence of spatial dependence on concepts. Let $\mathbb{D}_{c,\mu_1}$ and $\mathbb{D}_{c,\mu_2}$ denote two datasets containing the same concept but in different locations $\mu_1 \neq \mu_2$. For example, $\mathbb{D}_{c,\mu_1}$ may contain exemplars of `striped on the left` of the image, and $\mathbb{D}_{c,\mu_2}$ exemplars of the `striped on the right` of the image – an example is shown in fig. 3.2. As before, we construct latent representations $\mathbb{A}_{c,l,\mu_1}$ and $\mathbb{A}_{c,l,\mu_2}$ for datasets $\mathbb{D}_{c,\mu_1}$ and $\mathbb{D}_{c,\mu_2}$, respectively. Let $\boldsymbol{v}_{c,l}$ be the concept vector found using probe dataset $\mathbb{D}_{c,\mu_1}$.

Definition 3 (activation spatial dependence) *Let* $\boldsymbol{a}_{l,i}$ *be the activations corresponding to input* $\boldsymbol{x}_i$ *in layer* l*, and let* $\mu_{c,i}$ *be the location of concept* c *in* $\boldsymbol{x}_i$*. A layer has a spatially dependent representation of a concept iff*

$$\exists \phi : \forall \boldsymbol{x}_i \in \mathbb{X}^{+}_{c}, \phi(\boldsymbol{a}_{l,i}) = \mu_{c,i} \tag{3.7}$$

Activation spatial dependence in a NN may be due to architecture design, training procedure and/or the training dataset. In CNNs, it is the natural consequence of the receptive field of convolutional filters containing different regions of the input. If the NN has spatially dependent activations and the probe dataset has a spatial dependence, it may be possible to create a concept vector with spatial dependence.

Definition 4 (concept vector spatial dependence) *A concept vector $\boldsymbol{v}_{c,l}$ is spatially dependent with respect to the locations μ_1 and μ_2 iff*

$$\boldsymbol{a}^+_{c,l,\mu_1} \cdot \boldsymbol{v}_{c,l} > \boldsymbol{a}^+_{c,l,\mu_2} \cdot \boldsymbol{v}_{c,l} \quad \forall \boldsymbol{a}^+_{c,l,\mu_1} \in \mathbb{A}^+_{c,l,\mu_1}, \boldsymbol{a}^+_{c,l,\mu_2} \in \mathbb{A}^+_{c,l,\mu_2}. \tag{3.8}$$

If a CAV is spatially dependent then, by the definition above, it is more similar to the activations from images containing the concept in a specific location. This means the CAV represents not only the concept label, but the concept label at a specific location, e.g. striped objects on the right of the image, rather than striped objects in general. As done for the other two properties, we propose a hypothesis and aim to reject it later in the paper:

> ***Null Hypothesis 3 (NH3)**: Concept activation vectors cannot be spatially dependent*

We reject this hypothesis in §3.6.3 by analysing how the concept location in the probe dataset influences the spatial dependence of concept vectors. Rejecting NH3 motivates the introduction of *spatially dependent CAVs* (§ 3.6.3), which can be used to test if a model is translation invariant with respect to a specific concept and class.

3.4 Elements: A configurable synthetic dataset

To explore these hypotheses, we introduce a new synthetic dataset: Elements. In this dataset we can control: (1) the training dataset and class definitions, allowing us to influence model properties, such as concept correlation in the embedding space, and (2) the probe dataset, allowing us to test concept vector properties, such as concept vector spatial dependence. We further elaborate on these advantages in Appendix A.3.

Figure 3.2 shows examples of images in the Elements datasets. Each image contains n elements, where an element is defined by seven properties: colour, brightness, size, shape, texture, texture shift, and coordinates within the image. The dataset can be configured by varying the allowed combination of properties for each element. The ranges and configurations used for each property is given in Appendix A.3.

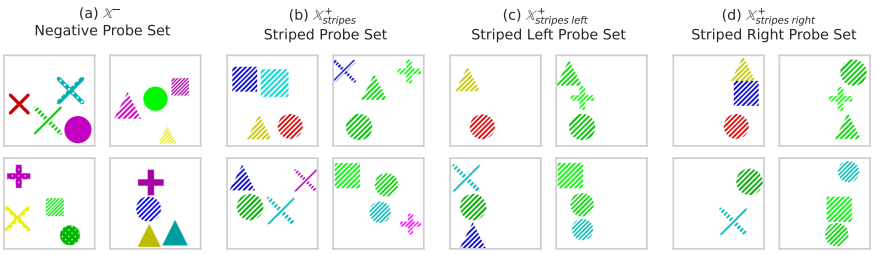


Figure 3.2: Example images from Elements probe datasets. (a) Negative probe set. A random selection of images – equivalent to images found in the model training set. (b) Positive probe set for `stripes`. (c) Positive probe set for `stripes on the left`. (d) Positive probe set for `stripes on the right`.

3.5 Related Work

Concept Correlation and Entanglement Chen et al. [49] discuss how concept vectors can be correlated, making it challenging to create a vector that solely represents one concept. While their work focuses on de-correlating concepts *during training*, we focus on analysing the impact of correlated concepts *after training* and show how they can lead to misleading explanations (§3.6.2). Fong and Vedaldi [72] use cosine similarity to demonstrate that the similarity between concepts varies based on the vector creation method. In our work, we also use cosine similarity to compare concept vectors. The distinction lies in our focus on CAVs and the insights they provide into the dataset and model.

Spatial Dependence Biscione and Bowers [25] describe how CNNs are not inherently translation invariant but can learn to be (under certain conditions on the dataset). This finding challenges the common assumption that CNNs possess inherent translation invariance. This assumption arises from the translation equivariance of convolutions and the common use of maxpooling or average pooling in large CNNs, as the output of a global pooling operation is independent of the location of the features. However, even if a model makes use of global pooling, local pooling and aspects such as padding lead to a lack of translation invariance by default. Through *spatially dependent CAVs*, we demonstrate translation invariance

with respect to a specific concept and class, rather than in general, providing more detailed information about a model.

What concept representations does our analysis apply to? Most concept-based interpretability methods represent concepts as *vectors* in the activation space of a trained neural network (NN) [116, 72, 244, 78, 239, 179, 70]. However, some concept-based methods use different representations: individual neurons [19], regions of activation space [54] or non-linear concepts [13, 129]. Our work focuses on the properties of concept *vectors*.

How is our work relevant in practice? To give insight into when the various properties may be relevant, we performed a review of computer vision papers which use CAVs in (1) the high-stakes applications of medical imaging (including skin cancer, skin lesions, breast cancer, and histology [229, 73, 172]), and (2) computer vision research on models trained with well-known datasets [122, 132, 221, 243, 189, 56]. A summary table can be found in Appendix A.7. We found that the following papers could have benefited from evaluating: consistency [229, 178, 73, 234, 79, 138], entanglement [229, 178, 73, 234, 79, 84, 144, 138, 172], and spatial dependence [229, 178, 73, 234, 79, 144, 138, 172]. We provide a detailed example, using the application of skin cancer diagnosis [229], in § 3.7.

Datasets While several datasets have been introduced for evaluating interpretability methods, they differ from ours in a few key ways. There are three questions we need our dataset to help answer:

1. Is the concept represented in the network?
2. Is the concept used for the network's prediction?
3. How does the network represent correlated concepts?

Existing datasets either do not allow insight into all three, or they have other practical reasons for being unsuitable. The Benchmarking Interpretability Method

(BIM) [231] inserts objects into scene images. While it benefits from utilizing real images and complex concepts (dog or bedroom), it also presents challenges. One drawback is that relying on real images makes it challenging to establish the ground truth relationship between concepts and class predictions or to know the similarities between concepts. As such, it does not give us insight into (2) or (3). The CLEVR dataset [111] could give insight into all 3, but because it renders 3D shapes it is too slow for our purposes. Elements generates images in 0.004s compared to CLEVR's 4s. This translates to a significant time saving when more data is required – 4s for 1000 images with Elements versus 1h for CLEVR. Analyzing CAV properties is our core focus. Elements, with its speed and flexibility, allows us to create the many different dataset versions required for experiments. The Navon and Trifeature datasets, used by Hermann and Lampinen [97] to study feature representations, could also give insight into the three questions, with associated concepts of shape, color and texture relating to each image. However, there is only one large object in each image so our experiments on spatial dependence would not have been possible. The synthetic dataset in Yeh et al. [232] is similar to our dataset but it was designed for concept discovery, featuring images where each object corresponds to a single concept (shape). In our dataset, each object contains multiple concepts, allowing us to create associations between them. We focus on explanation faithfulness by ensuring that the concepts must be used correctly by the model to achieve a high accuracy. So, for an accurate model, we have a ground truth understanding of how each concept is used. An extended literature review can be found in Appendix A.7.

3.6 Results: Exploring Concept Vector Properties

We explore the hypotheses on consistency (NH1), entanglement (NH2) and spatial dependency (NH3) in § 3.6.1, § 3.6.2 and § 3.6.3, respectively. We perform experiments using CAVs on the Elements and ImageNet datasets. Implementation details can be found in Appendix A.2.

3.6.1 Consistent CAVs

Theory We begin investigating NH1, which states that CAVs are consistent across layers, i.e. $f(\boldsymbol{a}_{l_1} + \boldsymbol{v}_{c,l_1}) = \boldsymbol{a}_{l_2} + \boldsymbol{v}_{c,l_2}$. Let $\hat{\boldsymbol{a}}_{l_1}$ and $\hat{\boldsymbol{a}}_{l_2}$ be linear perturbations to the activations in layers l_1 and l_2, respectively:

$$\hat{\boldsymbol{a}}_{l_1} = \boldsymbol{a}_{l_1} + \boldsymbol{v}_{c,l_1} \tag{3.9}$$

$$\hat{\boldsymbol{a}}_{l_2} = \boldsymbol{a}_{l_2} + \boldsymbol{v}_{c,l_2} = f(\boldsymbol{a}_{l_1}) + \boldsymbol{v}_{c,l_2} \tag{3.10}$$

We want to investigate if $\boldsymbol{v}_{c,l_1}$ and $\boldsymbol{v}_{c,l_2}$ have the same effect on the activations (and hence the model), i.e. if:

$$\begin{aligned} f(\hat{\boldsymbol{a}}_{l_1}) &= \hat{\boldsymbol{a}}_{l_2} \\ f(\boldsymbol{a}_{l_1} + \boldsymbol{v}_{c,l_1}) &= f(\boldsymbol{a}_{l_1}) + \boldsymbol{v}_{c,l_2}. \end{aligned} \tag{3.11}$$

Assuming f is continuous and differentiable, in Appendix A.1 we prove the result that Eq. 3.11 can hold if and only if f is equal to

$$f(\boldsymbol{a}_{l_1}) = g(\boldsymbol{a}_{l_1}) + M\boldsymbol{a}_{l_1} + \boldsymbol{b}. \tag{3.12}$$

Where $g(\cdot)$ is a periodic function with period $\boldsymbol{v}_{c,l_1}$ and $M \in \mathbb{R}^{m_{l_2} \times m_{l_1}}$ is a non-zero linear term with constant $\boldsymbol{b} \in \mathbb{R}^{m_{l_2}}$. More intuitively, we can obtain layer consistent vectors $\boldsymbol{v}_{c,l_1}$ and $\boldsymbol{v}_{c,l_2}$ if and only if f is composed of a periodic function with period $\boldsymbol{v}_{c,l_1}$ and a non-zero linear term M. In principal, f could approximate a function of this form since neural networks are universal approximators [206]. However, it seems reasonably unlikely that even if the model was modeling a periodic function it would have a period of exactly the same direction as a CAV. Throughout the rest of this section we provide empirical evidence that, in practice, layer consistent CAVs are not found.

Experiments Our goal is to investigate the question *are the concept vectors found using TCAV consistent across layers?* We measure the consistency of two perturbations using the consistency error:

$$\epsilon_{consistency} = ||f(\hat{\boldsymbol{a}}_{l_1}) - \hat{\boldsymbol{a}}_{l_2}|| = ||f(\boldsymbol{a}_{l_1} + \boldsymbol{v}_{c,l_1}) - (\boldsymbol{a}_{l_2} + \boldsymbol{v}_{c,l_2})|| \tag{3.13}$$

In our experiments, we use a scaling term to reduce the size of $\boldsymbol{v}_{c,l_1}$ and $\boldsymbol{v}_{c,l_2}$ to ensure the perturbed activation remains in distribution – see Appendix A.4.1 for details. If two perturbations have a consistency error of 0, then they have the same effect on the model. We include the following benchmarks:

Optimised CAV (lower bound): TCAV may not find a $\boldsymbol{v}_{c,l_2}$ that has a consistency error of 0 with $\boldsymbol{v}_{c,l_1}$. Therefore, we use gradient decent on $\boldsymbol{v}_{c,l_2}$ to minimise the consistency error, which acts as a lower bound.

Projected CAV: the error between $f(\boldsymbol{v}_{c,l_1})$ and $\boldsymbol{v}_{c,l_2}$, which measures how consistent the vectors are when projected into the next layer. If $f(\cdot)$ conserves vector addition, the projected CAVs would have 0 error.

Random (upper bound): We include two benchmarks. Random CAVs found using probe datasets containing random images, and a Random Direction vector: $\boldsymbol{v}_{c,l_2} \sim \text{Uniform}(-1, 1)$. If the consistency error is similar to random, it suggests that the CAVs between layers are as similar to each other as random directions.

Figure 3.3 shows the $\epsilon_{consistency}$ for different $\boldsymbol{v}_{c,l_2}$ across different training runs (see Appendix A.4 for details). The concept CAV obtains a nonzero consistency error, suggesting that CAVs across different layers are not consistent. When we compare it with the benchmarks, we find:

- The consistency error for the optimised CAVs is lower, implying that the standard approach to find CAVs does not find optimally layer consistent CAVs. However, the nonzero error for optimised CAVs suggests it is not possible to find consistent vectors across these layers.
- The projected CAVs have a nonzero error, indicating that vector addition is not preserved.
- The random CAVs have a higher error, suggesting the concept CAVs are more similar than random vectors.

The inability to find consistent concept vectors across layers suggests that the directions encoded by CAVs in different layers are not equivalent; instead we

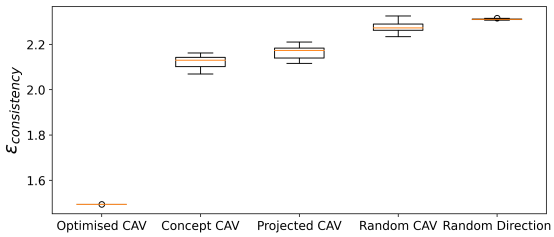


Figure 3.3: Empirical evidence for inconsistent CAVs across layers. The consistency error for different $\boldsymbol{v}_{c,l_2}$ for `striped` in the penultimate convolutional layer of a ResNet-50 trained on ImageNet. The optimised CAV acts as lower bound, whereas the random CAV and Direction act as baselines that provide an intuitive upper bounds. Concept CAV: `striped` CAVs, trained as normal. Projected CAV: `striped` CAVs from layer l_1 projected into layer l_2, $f(\boldsymbol{v}_{c,l_1})$.

speculate that they represent different components of the same concept. This aligns with the intuition that model representations are more complex later in the NN [149, 161, 19], therefore it is unlikely that the same aspects of a concept are represented in different layers (discussed further in Appendix A.4.3). Consequentially, TCAV scores across layers can vary as they perform different tests – they measure the class sensitivity to a different version of the concept.

Figure 3.5c shows that concept vectors found in different layers of a model can give contradictory TCAV scores (further examples available in Appendix A.4.4). In the Elements dataset, shape concepts are encoded in each layer as the test accuracy for each layer is above 93%. Therefore, we expect to be able to use TCAV on each of these layers. However, the TCAV scores (Figure 3.5c) for `cross` in the Elements dataset contradict each other across 'layers.3' and 'layers.4', suggesting a positive and negative influence, respectively. This contradiction makes it difficult to draw a conclusion about the model's class sensitivity to `cross`.

On the right of fig. 3.5c, we show the TCAV scores for `striped` for various classes in a ResNet-50 model trained on ImageNet. The accuracy for the `striped` vectors in ImageNet is above 96% for all layers tested, suggesting that the concept is encoded by the model in each of the layers. As in Elements, we do not observe consistent TCAV scores across layers. Instead, we observe a large change in the TCAV scores for `striped` in the penultimate layer, compared to earlier layers. 'layer4.1' suggests `striped` positively influences the likelihood of the classes tiger and leopard. However,

earlier layers suggest that the class is not is not sensitive to the concept. This shows how, depending on the layers that are tested, different conclusions can be drawn.

3.6.2 Entanglement

Different concepts may be associated with each other. For example, consider `blue` and the `sky` – a fundamental aspect of the sky is that it is often blue. These concepts are inherently linked and should not be treated as independent. This section will discuss how to discover these associations using CAVs and the implications for TCAV.

To explore entanglement, we quantify and visualize concept associations by computing average pairwise cosine similarities between CAVs (we compute multiple CAVs for each concept). We investigate three models trained on different versions of the Elements dataset. Each dataset is identical aside from the association between `red` and `triangle`:

$\mathbb{E}_1$: each combination of colour, shape and texture is equally likely,

$\mathbb{E}_2$: the only shape that is red is triangles,

$\mathbb{E}_3$: the concepts of red and triangle only ever co-occur.

In fig. 3.4 we show one plot for each dataset. For $\mathbb{E}_1$, we observe no positive association between the concepts. In $\mathbb{E}_2$, we observe a small positive association between the triangle and red concepts. Lastly, in $\mathbb{E}_3$, the cosine similarity between the `red` and `triangle` CAVs approaches the similarity of the concept with itself. The trend between $\mathbb{E}_1$, $\mathbb{E}_2$ and $\mathbb{E}_3$ is likely due to the underlying association between the `red` and `triangle` increasing. We perform similar analyses on ImageNet in Appendix A.5.

Interestingly, we often observe a negative cosine similarity between mutually exclusive concepts. The model has encoded concepts that cannot co-occur (e.g., each element can only have a single colour) in directions negatively correlated with each other. The presence of the `red` diminishes the likelihood of the `blue` or `green` being present, and by having these concepts negatively associated with each other

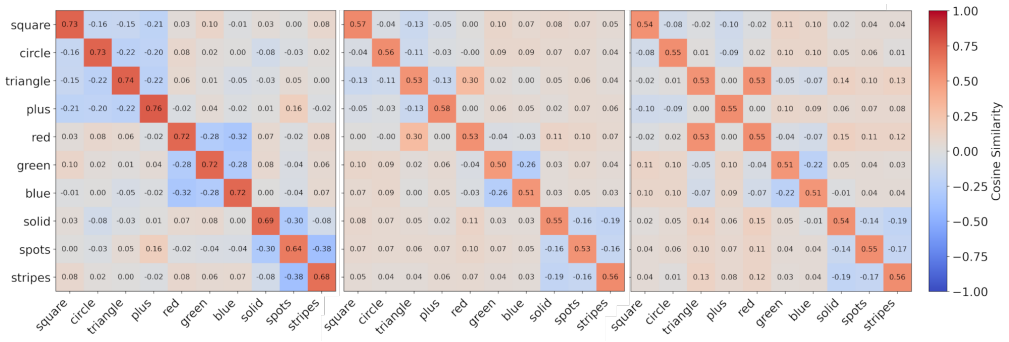


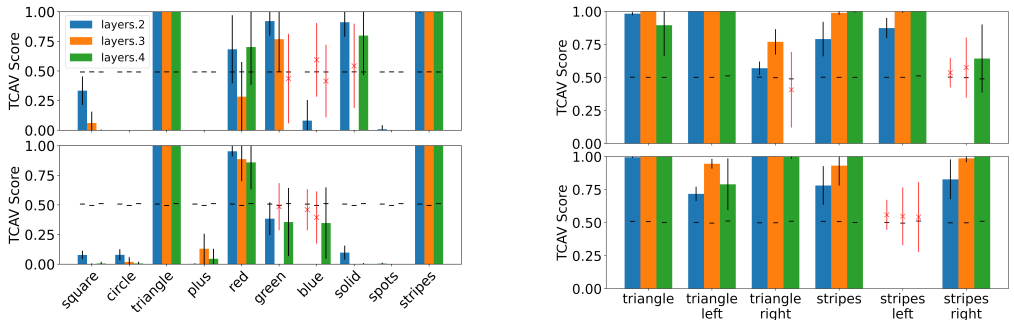
Figure 3.4: Cosine similarities demonstrating entangled concepts. Mean pairwise cosine similarities for all concepts from different versions of the simple Elements dataset, with an increasing association between `red` and `triangle` from left to right: $\mathbb{E}_1$, $\mathbb{E}_2$ and $\mathbb{E}_3$.

the model builds in this reasoning. This means that the `red` CAV does not solely signify `red`, it also encapsulates `not blue` and `not green`.

Next, we investigate the effect of entangled concept vectors on TCAV score. We analyse the TCAV scores for the 'striped triangles' class in $\mathbb{E}_1$ and $\mathbb{E}_2$. The class label depends solely on the presence `stripes` and `triangle`. Therefore, we expect all other concepts to obtain low TCAV scores (indicating negative sensitivity), as their presence makes the class less likely, or insignificant TCAV scores, if the concept is uninformative.[3]

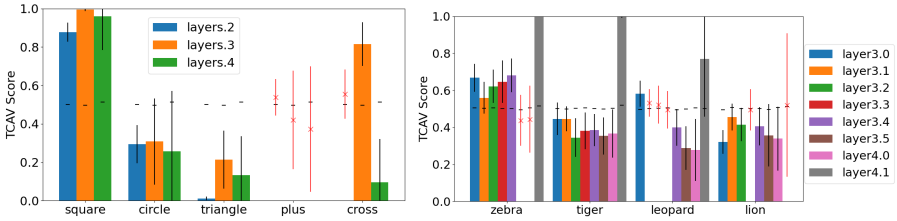
The results for $\mathbb{E}_1$ and $\mathbb{E}_2$ are shown on the top and bottom of fig. 3.5a, respectively. For $\mathbb{E}_1$ (the unaltered dataset), we find that only the `stripes` and `triangle` vectors have a high TCAV score across multiple layers. For $\mathbb{E}_2$ (the altered dataset), however, the model appears to be sensitive to `red`, `triangle` and `stripes`, with high TCAV scores for each. This is due to the association between the `red` and `triangle` CAVs. $2,374/5,000$ images in the test dataset contain striped triangles. None of these are incorrectly classified, so it is unlikely that the model uses the red concept for its prediction. Instead, the association between CAVs causes a misleadingly high TCAV score for the red concept. In conclusion, associated CAVs can lead to misleading explanations.

[3]assuming that the model uses each concept correctly



(a) Entangled CAVs increase `red` TCAV scores for $\mathbb{E}_2$. TCAV scores for all concepts in $\mathbb{E}_1$ (top) and $\mathbb{E}_2$ (bot) for the class of striped triangles.

(b) TCAV scores are spatially dependent for the 'striped triangles on the left' (top) and 'striped triangles on the right' (bot) classes in Elements.



(c) Inconsistent TCAV scores across layers. Left: TCAV scores for shape concepts for the 'solid red squares' class in Elements. Right: TCAV scores for `striped` for a subset of ImageNet classes.

Figure 3.5: Consistency, entanglement, spatial dependence can affect TCAV scores. The standard deviation is black or red for significant and insignificant results, respectively. The null for each layer is shown as a horizontal black line.

3.6.3 Spatial Dependence

Finally, we investigate NH3: *are CAVs spatially dependent?* We reshape the CAVs back into the original shape of the activations, and compute the channel-wise norm as follows:

$$\mathbf{S}_{c,l} = \|\text{reshape}(\boldsymbol{v}_{c,l}, (H, W, D))\|_2, \tag{3.14}$$

where $\mathbf{S}_{c,l} \in \mathbb{R}^{H \times W}$, and $\|\cdot\|_2$ is the L_2 norm across the channel dimension. We refer to this array as the spatial norms of the CAV.

If a CAV's spatial norm varies substantially across the (H, W) dimensions, it indicates that the CAV is spatially dependent (see Appendix A.6.2 for an explanation). Visualising a CAV's spatial norms shows us which regions contribute most to the directional derivative and, consequently, to the TCAV score.

To create spatially dependent CAVs, we constructed spatially dependent probe datasets for Elements and ImageNet where we either restricted the location of

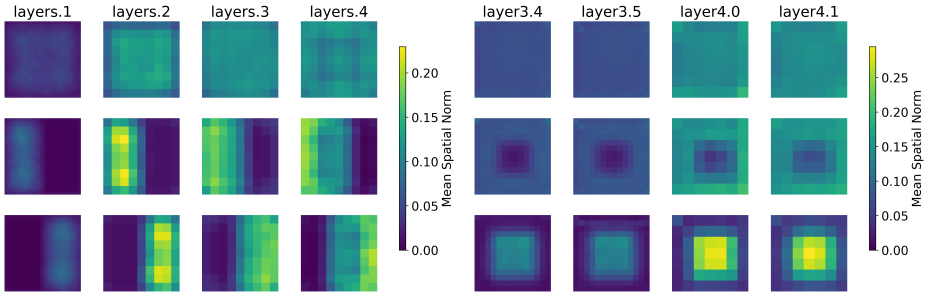


Figure 3.6: Spatial norms reflect the spatial dependence of the probe dataset. Left: Mean spatial norms for `red` (top), `red left` (middle) and `red right` (bottom) for Elements. Right: Mean spatial norms across for `striped` (top), `striped edges` (middle) and `striped middle` (bottom) for ImageNet.

the concepts or greyed out parts of the image – see fig. 3.2 for examples and Appendix A.6.1 for further details.

When a spatially independent probe dataset is used to create CAVs, as in the top row of fig. 3.6, the spatial norms are uniform, suggesting the CAVs are not spatially dependent[4]. However, when the probe dataset exhibits spatial dependence, so do the resulting CAVs. The regions of near-zero norm indicate that the corresponding spatial regions of the gradients do not contribute to the directional derivative and, consequently, to the TCAV score.

Next, we investigate the question *does the model have a different conceptual sensitivity depending on the concept's location in the input image?* As CAVs operate in the activation space of a specific layer, we can show that a model is not translation invariant if:

1. The model has activation spatial dependence, i.e. pixels in different locations affect the activations differently.

2. That each depth-wise slice of the activations, of shape $(1, 1, D)$, affects the logit output differently.

Both of these components affect the TCAV score. (1) influences $\boldsymbol{v}_{c,l}$ and (2) influences $\nabla h_{l,k}\left(g_l(\boldsymbol{x})\right)$. For (1), fig. 3.6 demonstrates that the model has activation

[4]the individual CAVs may still be spatially dependent, but this cancels out across training runs. See Appendix A.6.3 for details.

spatial dependence as the locations with the highest spatial norms approximately correspond to the location of the concept in the image space.

To address (2), we compute the TCAV scores for different sets of spatially dependent CAVs to determine if the sensitivity of the model changes depending on the concepts location. To investigate this, we created spatially dependent classes in the Elements dataset, where the class depends on what concepts are present *and* on where they are in the image, such as 'striped triangles on the left'. We use spatially dependent CAVs to show that a model is not translation invariant with respect to `striped` or `triangle` in fig. 3.5b. Here, we discuss the results for the class of 'striped triangles on the left'. The TCAV scores for `striped`, `triangle`, `striped left` and `triangle left` are high, indicating a positive influence of these concepts on the class. However, the `striped right` and `triangle right` TCAV scores often do not differ significantly from the null scores, providing no evidence to suggest the model is sensitive to these concepts. The difference between the left and right biased TCAV scores indicates that the model is not translation invariant with respect to these concepts as the model's sensitivity depends on where the concept is present in the image input space. Overall, this suggests that we can use CAVs to detect model translation invariance. See Appendix A.6.6 for examples on ImageNet.

3.7 Practitioner Recommendations

Our results have shown that failure to appropriately consider consistency, entanglement, and spatial dependence may result in drawing incorrect conclusions when using TCAV. Therefore, we recommend the following:

- *Consistency*: creating CAVs for multiple layers, rather than a single one;
- *Entanglement*: (1) verifying expected dependencies between related concepts, and (2) being mindful that a positive TCAV score may be due to concept entanglement;
- *Spatial Dependence*: visualising concept vector spatial dependence using spatial norms.

In § 4.2, we provided example papers which use CAVs and may be influenced by the above properties. To provide a concrete example, we examine the use-case of Yan et al. [229] which uses CAVs in the context of skin cancer diagnosis. Below we demonstrate how our recommendations could have been used and how the analysis may impact the conclusions drawn.

Consistency The authors use CAVs on a single layer. As discussed in § 3.6.1 and § A.4, different layers can represent different aspects of the same concept. To have a better understanding of the overall effect of the concept on the model, CAVs should be created for multiple layers.

Entanglement There are multiple concepts which have opposed meanings, for example `regular streaks` and `irregular streaks`, or `regular vascular structures` and `irregular vascular structures`. As such, we expect the cosine similarities between the CAVs to confirm that these concepts are negatively correlated (or less similar to each other than to other concepts).

Spatial Dependence Some of the concepts have expected spatial dependencies, for example, `dark borders` and `dark corners`. Spatial norms could be used to confirm these spatial dependencies exist. Equally, for concepts such as the presence of a `ruler`, the spatial norms could confirm the CAVs have no overall spatial dependence.

3.7.1 Experiment Setup

To further this example, we run illustrative experiments on a similar dataset to Yan et al. [229]. The dataset used in [229] is not publicly available.

Model We finetune a ResNet50 [93] pretrained on ImageNet [56] on the ISIC 2019 dataset [214, 51, 53] for the binary classification of skin lesions as melanoma or benign. We use a binary cross entropy loss and the Adam optimiser [119], training

until convergence of validation loss to achieve an area under the receiver operating characteristic curve (AUC) of 0.91 on the validation split.

CAVs For the CAVs, as in Yan et al. [229], we use the derm7pt dataset [113]. There are 12 clinical concepts which have been expertly labelled within the dataset: `regular pigment network`, `irregular pigment network`, `blue whitish veil`, `regular vascular structures`, `irregular vascular structures`, `typical pigmentation`, `atypical pigmentation`, `regular streaks`, `irregular streaks`, `regular dots and globules`, `irregular dots and globules`, `regression structures`. We hand labelled three additional concepts, which were used in [229], of `dark corners`, `dark borders` and `ruler` which are possible confounders for the model. We defined `dark corners` as any image with a circular aperture which left the corners of the image black, `dark borders` as any image containing rectangles of blacked out areas and `ruler` as any image containing a ruler. For each of the medical concepts, there are three labels: typical/regular, atypical/irregular, and absent. When training CAVs for these concepts, we created separate CAVs for the typical/atypical concepts and either used random images or images with the label of 'absent' as the negative probe dataset. Random images are used as the negative set because this is the common practice when training CAVs [116] but it relies on the assumption that the concept is relatively rare in the dataset and so will occur in the negative set infrequently. This assumption is not the case in this dataset and is likely why using random images as the negative set gave low-quality CAVs, with accuracies of 50-65% (see fig. 3.7). Hence, we used the 'absent' labels for the negative set when training the medical CAVs. It is unclear from [229] what they used for negative sets. For training the CAVs we used 70 images per concept and used 30 different random seeds for the random negative probe set to get 30 CAVs per concept. Yan et al. [229] do not use TCAV, so this setup is different from the original paper.

3.7.2 Results

In fig. 3.7, we report the validation accuracy of the linear classifier created during the CAV training process, henceforth referred to as CAV accuracy. The figure demonstrates that initial results with random CAVs gave poor performance for the medical concepts, so we used the 'absent' category for each class as the negative set. This gave better performances with accuracy between 60-75% for the medical concepts but this is still far lower than for the potential confounders at 80-90%. We hypothesise that this is due to the simplicity of the confounding concepts. This is supported by the accuracy reported by Yan et al. [229], where they also obtained a lower CAV accuracy for the medical concepts. The accuracy for the potential confounding concepts of `dark borders` and `dark corners` drops in later layers. This is likely due to the spatial nature of the concepts – an idea further supported by fig. 3.10 where the CAVs in later layers have reduced spatial distinction.

Below, we examine how the three CAV properties analysed in this paper affect the TCAV scores and the conclusions you can draw from them.

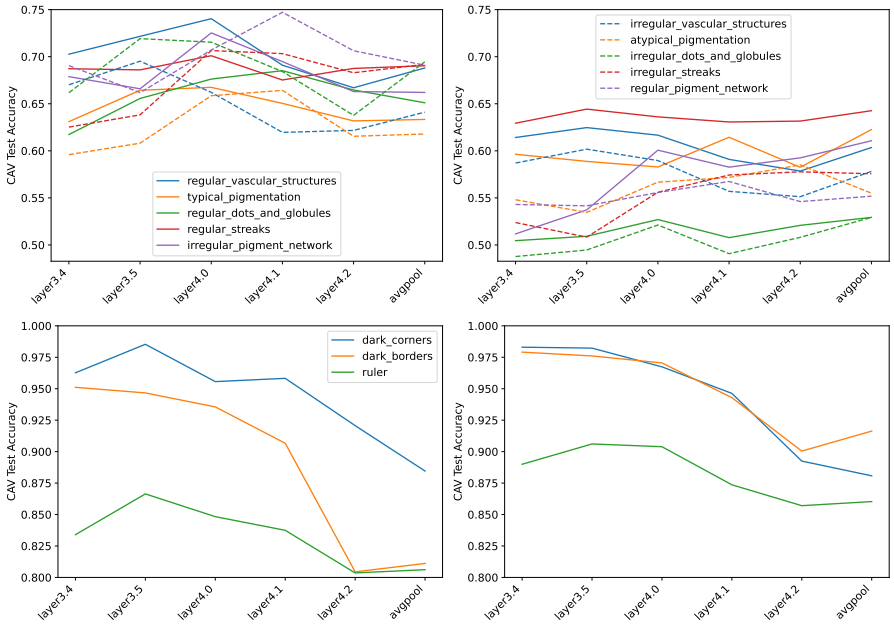
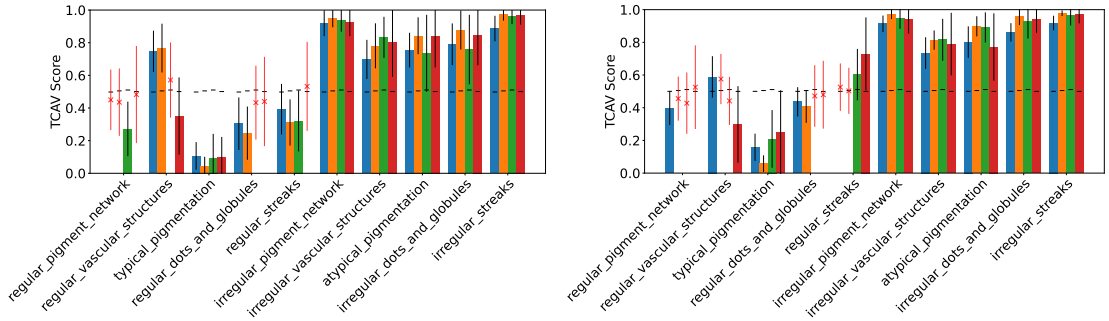
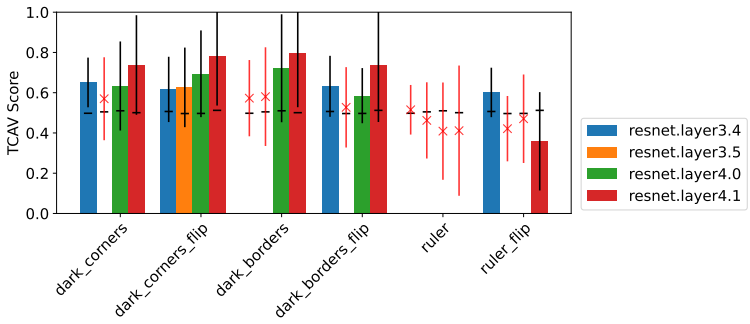


Figure 3.7: Mean CAV test accuracies for the melanoma use-case. Top: Medical concepts where random images (right) or images where the concept is labelled as absent (left) are used in the negative probe dataset. Bottom: Potential confounders where CAVs were trained with (right) and without (left) a flip augmentation.



(a) TCAV scores are not qualitatively different for CAVs of differing accuracy. TCAV scores for medical concepts where random images (left) or images where the concept is labelled as absent (right) are used in the negative probe dataset.



(b) CAVs trained with an augmentation (indicated by the '$_flip'suffix)weresignificantmoreoftenthanthosewithnoaugmentation.TCAVscoresforpotentialconfo$

Figure 3.8: TCAV scores for the melanoma use-case. The standard deviation is black or red for significant and insignificant results, respectively. The null for each layer is shown as a horizontal black line.

Consistency The TCAV scores for many of the medical concepts (fig. 3.8a) are consistent across layers, irrespective of if random images or 'absent' images were used for the negative probe set. The consistent scores provide more confidence in using them to explain the model's behaviour as repeated significance tests are performed indicating the model has the same sensitivity to the concept. In terms of understanding the model, the scores provide some evidence that it operates similar to human experts, as the TCAV scores for the atypical/irregular medical concepts are high for the malignant class, as expected, and the confounding concepts are often not significant (prior to using a flip augmentation - see the spatial dependence section below for discussion), providing little evidence that the model is sensitive to the potential confounders.

Entanglement Interestingly, for the CAVs with 'absent' labelled images in their negative probe dataset (the right of fig. 3.9), the CAVs with opposed meanings often appear to be the most similar to each other. For example, `regular vascular structures` has a negative or zero similarity with all concepts except `irregular vascular structures` (with a similarity of 0.11). We hypothesise that this is because the two concepts share the same negative set. This hypothesis is supported by [178] where, while they did not discuss changing solely the negative set, they did show that CAVs are sensitive to the choice of probe dataset. In addition, if compared to the similarities between CAVs trained using random images as the negative set (the left of fig. 3.9), we see that this pattern disappears. The higher similarity between concepts which have opposite meanings suggests that the CAVs do not represent the concepts they are labelled for. Therefore, in this case, we do not believe the TCAV scores can be interpreted as we do not have confidence the CAVs represent their desired concept. This example highlights a more general problem that the negative probe set can have a substantial affect on the resulting CAV, even though it is the positive probe set that is designed to represent the concept.

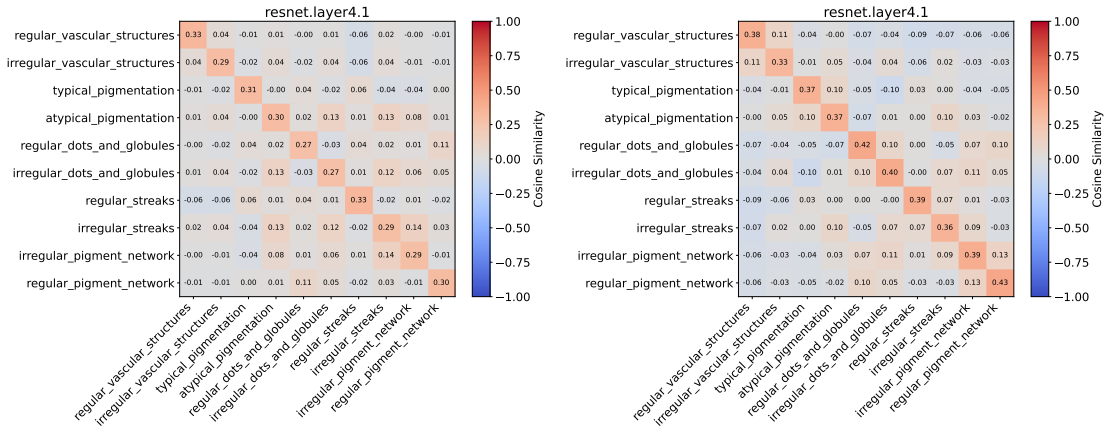


Figure 3.9: Cosine similarity matrix for CAVs of different concepts from derm7pt when random images (left) or images where the concept is labelled as absent (right) are used in the negative probe dataset.

Spatial Dependence The spatial norms in fig. 3.10 show a clear spatial dependence in the center for each of the medical concepts across all layers. This

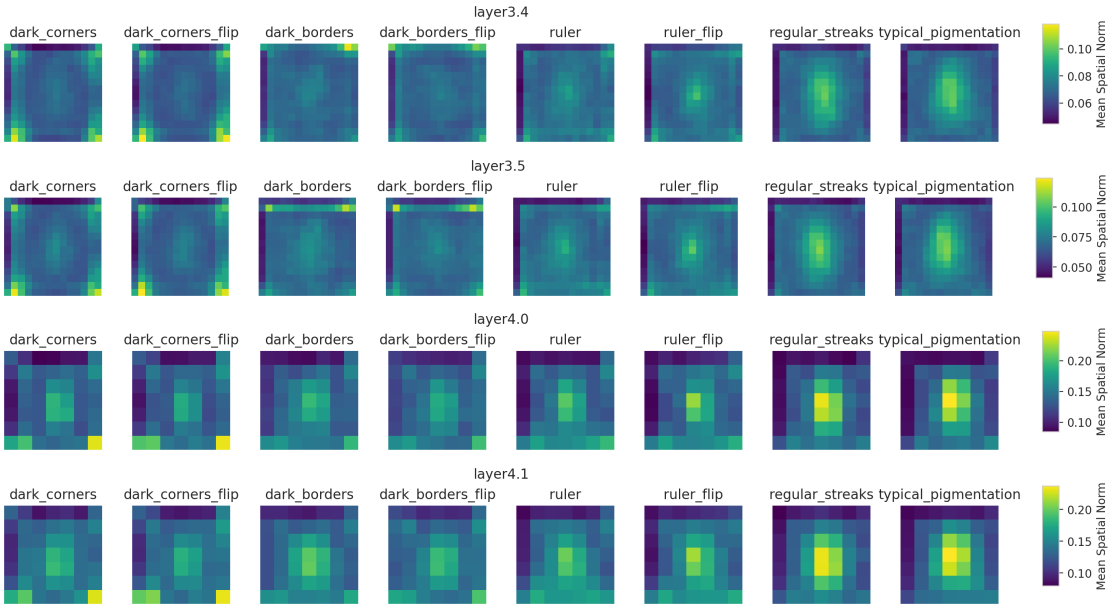


Figure 3.10: Mean CAV spatial norms for a selection of CAVs from the melanoma use-case.

aligns with expectations, as the dataset requires the skin lesion to be centered in the image and each concept is related to the appearance of the lesion. The `dark corners` and `dark borders` concepts, however, show deviations to this pattern in the earlier layers, with `dark corners` having high spatial norms in the corners and `dark borders` high spatial norms in the top. This is desirable, as these confounding concepts depend on features at the edge of the image, away from the lesion. Along with the accuracies in fig. 3.7, this suggests that the CAVs in earlier layers better represent these spatially dependent concepts. For `dark borders`, however, we hypothesise that the high spatial norms in just a single direction suggest that the CAVs are not a good representation of dark borders in locations other than the top. Therefore, we retrained CAVs for each of the confounding concepts but with an augmentation applied to the probe dataset to randomly flip the images in the horizontal and/or vertical direction. This removes any bias that there may be in the probe dataset for the concept to be in one particular direction. Figure 3.10 shows that the CAVs trained for `dark borders` with an augmentation (`dark borders flip`) had only minor improvements for their spatial norms, with some layers being slightly less uni-directional. Figure 3.7, however, shows that the accuracy for each

of the CAVs improved. This suggests that there was an issue with bias in the probe dataset and by using a flip augmentation we create CAVs which better represent the concepts, but, surprisingly, the activations of the model encode the `dark borders` concept in a spatially unsymmetrical manner.

The TCAV scores for CAVs trained with the augmentation (fig. 3.8b) obtain significance in more layers but the results are still fairly inconclusive. For example, the TCAV scores for `ruler` in layer3.4 and layer 4.2 are above/below the null, respectively. This means that layer3.4 suggests a positive influence, whereas layer4.2 suggests a negative influence. With less than a 1% difference in accuracy between the CAVs of the two layers it is not clear which layer we might trust more and so no conclusive statements can be said about the influence of `ruler` on malignancy predictions.

3.7.3 Summary

The poor accuracy of medical concepts with random images in the negative probe set required the use of the 'absent' category instead. However, the similarity matrices in fig. 3.9 suggest that these CAVs do not represent the desired concepts. Therefore, we do not think either set of medical CAVs can produce meaningful TCAV scores.

For the potential confounders, however, the CAVs had high accuracy and the spatial norms indicated spatial dependence in concepts we expect spatial dependence, providing evidence that these CAVs are suitable to use. However, the spatial dependence seemed directional, and so an augmentation was added to flip the probe images horizontally/vertically. Although this did not substantially change the spatial norms, it improved CAV accuracy and increased the number of layers for which we had significant TCAV scores. From these scores, it appears the model is sensitive to `dark corners` although the evidence is weak with large standard deviations in TCAV score and for `ruler` we found inconclusive results with inconsistent scores across layers.

This use-case demonstrates the importance of analysing consistency, entanglement and spatial dependence of CAVs, alongside more typical evaluations such

as CAV accuracy and statistical significance, in order to understand CAV-based explanations and the conclusions you can draw from them. Our experiments provide an example for practitioners to follow in their own experiments with CAVs.

3.8 Conclusion and Future Work

In this work, we explore three key properties that influence concept activation vectors (CAVs): consistency, entanglement and spatial dependence. First, we derive conditions under which CAVs in different layers are not consistent and substantiate our findings with empirical evidence. This sheds light on why CAV-based explanations methods can give conflicting conclusions across layers. Next, we introduce a visualisation tool designed to facilitate the exploration of associations between concepts within a dataset and model. Lastly, we show that spatial dependence impacts CAVs, and introduce a method that can be used to detect spatial dependence within models. We provided clear recommendations on how to mitigate the impact of these properties on CAV-based explanations and demonstrated how to use those recommendations for a medical imaging use-case. The CAV properties were explored using a synthetic dataset, Elements, where custom probe datasets can easily be created to analyse properties of interest. We release this dataset to help further explore this problem space.

In the introduction of this chapter, we cited several interpretability methods that employ vector representations to convey semantically meaningful concepts. Our study has illuminated certain properties and consequential outcomes arising from these vector-based approaches. In future research, the characteristics inherent in alternative forms of representation, such as clusters within activation space [54], should be investigated and the relative merits assessed.

Words are, of course, the most powerful drug used by mankind.

— Rudyard Kipling

4
Debugging with TextCAVs

Contents

The previous chapter delved into the details of how CAVs work. One of the limitations of CAVs that was not extensively discussed in the previous chapter is that a probe dataset of concept examples is required to train a CAV. This makes it a slow and expensive process to test the sensitivity of new concepts, particularly in the medical domain, as you need a dataset of images with labelled concepts. This chapter introduces TextCAVs, a novel method which changes how CAVs are created to remove the need for probe datasets. Instead, CAVs are created using pretrained vision-language models like CLIP [176], where images and text share an activation space. By training a simple linear layer between the vision-language model and a target model, text-based explanations can be generated for the target model. This means new concepts can be tested with minimal compute

and no additional labels or data.

A limitation of TextCAVs is that the explanations can be quite noisy, but rather than aiming to give a complete causal description of the target model, we propose using TextCAVs to debug models. The ability to quickly test new concepts and ideas allows for interactive debugging of a model and we demonstrate this process on both natural images (ImageNet [56]) and chest X-rays (MIMIC-CXR [109, 110]).

In § 4.4.1 we include a crowd-worker based evaluation of TextCAVs from the Secure and Trustworthy Machine learning (SaTML) interpretability competition [39]. This evaluation, along with the competition design, was performed by S. Casper.

4.1 Introduction

Deep learning-based models are increasingly utilised in healthcare scenarios where mistakes can have severe consequences. One approach for creating safer, more reliable models is to use interpretability: the ability to explain or present a model in terms understandable to a human [60].

Many different interpretability methods have emerged, with explanations taking a variety of different forms such as individual pixels, prototypes or concepts. We focus on concept-based methods which provide explanations using high-level terms that humans are familiar with. Concept activation vectors (CAVs) are a common approach used to represent concepts within the activation space of a model and are found using a probe dataset of concept exemplars [116].

The labels required for this can be expensive to obtain in medical domains where expert clinical input is necessary. We introduce TextCAVs, a concept-based interpretability method that uses solely the text label of the concept, or descriptions of it, rather than image examples.

We demonstrate that TextCAVs give meaningful explanations for both natural image (ImageNet [56]) and chest X-ray (MIMIC-CXR [109, 110]) tasks. Further, as interpretability itself is difficult to measure, we demonstrate its usefulness in debugging deep learning-based models through finding implanted dataset bias in MIMIC-CXR and maliciously implanted trojans in an ImageNet model.

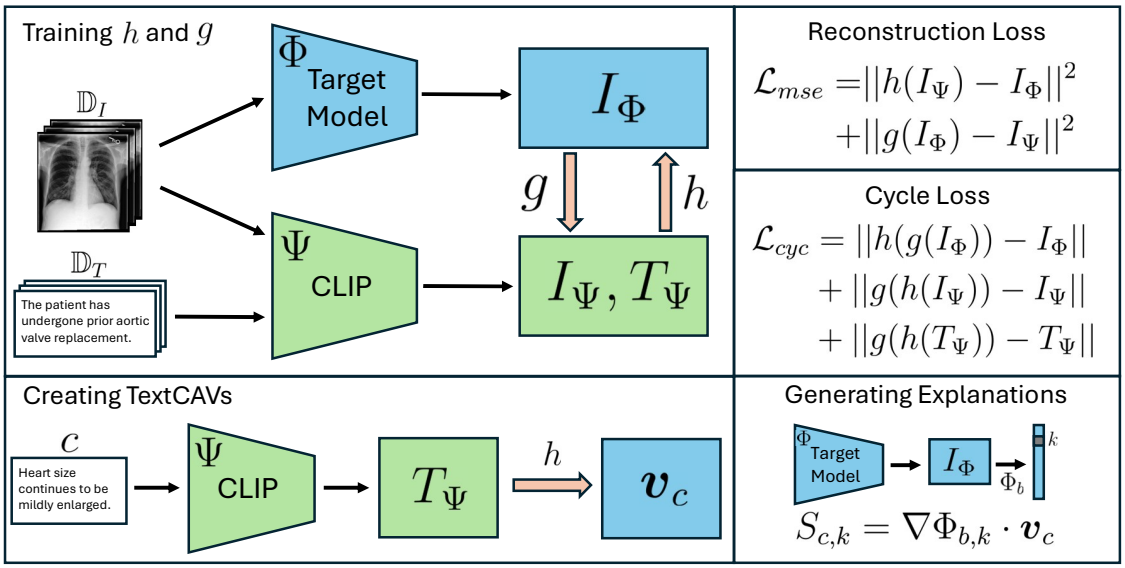


Figure 4.1: Explaining models with TextCAVs. In order to move between the activations of a CLIP model and our target model, we train linear transformations, h and g, using a text dataset, $\mathbb{D}_T$, and image dataset, $\mathbb{D}_I$. The loss terms are detailed on the right with I_Φ, I_Ψ and T_Ψ representing the image features of the target model, the image features of the CLIP model, and the text features of the CLIP model, respectively. Once h is trained, TextCAVs can be created by passing text representing some concept, c, through the CLIP model and h. The model's sensitivity to c, for some logit output, k, can then be measured using the directional derivative, $S_{c,k}$: the similarity between the model gradient, $\nabla\Phi_{b,k}$, and a TextCAV, $\boldsymbol{v}_c$.

4.2 Related Work

Kim et al. [116] introduce Testing with Concept Activation Vectors (TCAVs) where they use probe datasets of concept examples to create CAVs and then compare the CAVs with model gradients to measure a model's sensitivity to a concept for a specific class. We also use the directional derivative (dot product between CAV and gradient) to measure model sensitivity, but our CAVs are created using a multi-modal model and so do not require a probe dataset for each concept.

In order to reduce the cost of creating concept-based explanations, a variety of different methods automate the process of finding concepts [78, 239, 179, 85, 70]. However, the meaning of each concept is not always readily apparent and the concept must be visually present in the dataset used to discover the concepts. Our method reduces cost using a different approach as we also do not need to collect labelled data for each concept, but our resulting CAVs have inherent meaning

from their text descriptions.

CLIP models [176] have demonstrated strong performance in vision-language tasks. Their joint embedding space for text and images allows for built-in comparisons between the modalities and therefore for zero-shot classification. A variety of adaptations have been suggested for the biomedical space [242] with some models being trained for specific modalities like chest X-rays (e.g. BioViL [27]) and others more generally (e.g. BiomedCLIP [240]). We use these vision-language models in our method but, importantly, inference is performed by the target model, without placing restrictions on its architecture or method of training.

Yuksekgonul et al. [234] use multimodal models to create CAVs and then use the similarity between model activations and these CAVs to create a concept bottleneck model. Moayeri et al. [148] extend this approach to target vision models more generally by, as in our work, training a simple linear layer to transfer the features of the target model to a CLIP model. Also as in our work, Shipard et al. [199] improve the transfer of features by training a linear layer in both directions and using multimodal losses. However, these approaches focus on zero-shot classification and on changing how the model inference is performed, rather than explaining the model in its current state using gradients.

4.3 TextCAVs

For some target model, Φ, and a CLIP-like vision-language model, Ψ, let $I_\Phi \in \mathbb{R}^m$ and $I_\Psi \in \mathbb{R}^n$ be the extracted features for some image, $\boldsymbol{x} \in \mathbb{R}^N$, from an imaging dataset $\mathbb{D}_I$. As Ψ contains a joint embedding space between text and images we can also extract text features: $T_\Psi \in \mathbb{R}^n$ from some text, t, from a dataset $\mathbb{D}_T$. We train two linear layers $h : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^n$ which can be used to convert between the features of the two models. To create TextCAVs, we only need h but to improve h's ability to convert text features we use a cycle loss term which requires g. The loss is composed of two parts: reconstruction loss and cycle loss. The reconstruction loss is simply the mean squared error (MSE) between the image features and converted features.

$$\mathcal{L}_{mse}(\boldsymbol{x}) = ||h(I_\Psi) - I_\Phi||^2 + ||g(I_\Phi) - I_\Psi||^2 \tag{4.1}$$

The reconstruction loss can only be calculated for image features as we need features from both models (Φ and Ψ). To include information from the text features in the loss function we use cycle loss which ensures that the features are consistent with their original form when converted back to their original space:

$$\begin{aligned}\mathcal{L}_{cyc}(\boldsymbol{x}, t) &= ||h(g(I_\Phi)) - I_\Phi|| &(4.2)\\ &+ ||g(h(I_\Psi)) - I_\Psi|| &(4.3)\\ &+ ||g(h(T_\Psi)) - T_\Psi||. &(4.4)\end{aligned}$$

Once trained, we use h, Ψ and a concept label, c, to obtain a concept vector in the activation space of the target model:

$$\boldsymbol{v}_c = h(\Psi(c)). \tag{4.5}$$

Φ can be decomposed into two functions: $\Phi_a(\boldsymbol{x}) = I_\Phi \in \mathbb{R}^m$ which maps the input $\boldsymbol{x} \in \mathbb{R}^N$ to its features I_Φ, and $\Phi_b(I_\Phi)$ which maps I_Φ to the output. To obtain the model's sensitivity to a concept for a specific class, as in [116], we calculate the directional derivative:

$$\begin{aligned}S_{c,k}(\boldsymbol{x}) &= \lim_{\epsilon \to 0} \frac{\Phi_{b,k}\left(\Phi_a(\boldsymbol{x}) + \epsilon \boldsymbol{v}_c\right) - \Phi_{b,k}\left(\Phi_a(\boldsymbol{x})\right)}{\epsilon}\\ &= \nabla \Phi_{b,k}\left(\Phi_a(\boldsymbol{x})\right) \cdot \boldsymbol{v}_c.\end{aligned} \tag{4.6}$$

If Φ_a is chosen to be the output of the penultimate layer in a model then the directional derivative can be calculated without image exemplars:

$$S_{c,k} = \nabla \Phi_{b,k} \cdot \boldsymbol{v}_c. \tag{4.7}$$

This is due to the lack of non-linearities between the penultimate layer and the logit output. Having solely a linear layer between the features and the output means the gradient of the activations with respect to the logit does not depend on the

activations. This means we can extract gradients, and therefore model explanations, using solely the model weights. Therefore, once h has been trained, TextCAVs requires **only** the text you wish to test to be able to generate an explanation. In practice, to calculate the gradient, we input an array of zeros of the same shape as the images, but this is an arbitrary choice. In this work, we use the penultimate layer in all experiments and leave exploration of using other layers for future work.

By ranking concepts based on their directional derivative, we obtain a list of sentences/words ordered by the model's sensitivity for a specific class. If we can filter this list for concepts which we expect to be there, we can discover bugs in the model. Ideally, this would be done by a human expert who could use their domain knowledge to explore different hypotheses. The minimal overhead for testing new concepts allows the user to test words related to new hypotheses quickly and provide an interactive process to model debugging.

4.4 Experiments

In this section we provide a description of our training setup, our model choices, evaluation and then a discussion and analysis of our results experiments with both the ImageNet and MIMIC-CXR datasets.

4.4.1 ImageNet

In this section, we demonstrate that TextCAVs produces reasonable explanations for a standard ResNet-50 [93] trained on ImageNet and then describe our 3rd place entry into the Secure and Trustworthy Machine Learning Conference (SaTML) interpretability competition to detect trojans (implanted bugs) in vision models trained on ImageNet [39].

Training Details We use 20% of the ImageNet training dataset to train h and g and train for 20 epochs. For the target model, Φ we use the default weights for a ResNet-50 [93] in the TorchVision package in PyTorch. For the vision-language model, Ψ, we use a pretrained ViT-B/16 CLIP model [176].

Concepts In a similar manner to Oikarinen et al. [160], in order to automate the process, we use a large language model (LLM) to obtain a list of concepts. We use three prompts asking for the "things most commonly seen around" "visual elements or parts" and "superclasses" of each class in ImageNet. We then extract and perform basic filtering of the concepts, removing: plurals of the same word; the words "an", "a" and "the"; and concepts containing more than 2 words. To obtain the final list of concepts we remove similar concepts using text embeddings from Ψ. If a set of concepts have a cosine similarity greater than 0.9, only the shortest concept is retained. This reduces the number of near synonyms in the concept list. For the LLM, we use a 4-bit quantized version of the Tulu-v2-7b model [104].

Results In Table 4.1, we show the top-10 concepts for a selection of ImageNet classes. All the concepts relate to their respective class, indicating that TextCAVs can produce plausible explanations.

Table 4.1: Top-10 concepts ordered by directional derivative for a selection of classes in the ImageNet model.

bullfrog	**albatross**	**orangutan**	**bucket**	**cellphone**
american bullfrog	gannet	orangutan	crab buckets	mp3 player
green frog	seagull	howler monkey	diaper pail	phone
boreal toad	sea eagle	macaque	bucket	phone case
western toad	shearwater	tarsier	laundry basket	memory card
frog	gull	great ape	watering can	walkman
musk turtle	white-tailed eagle	long-nosed monkey	flower pot	cordless phone
snapping turtle	petrel	gibbon	cooking pot	bluetooth
toad	merganser	gorilla	dustbin	smartwatch
terrapin turtle	wading bird	langur	fishing basket	card reader

Competition We entered TextCAVs into the SaTML competition [39] for detecting human interpretable trojans. Here, we provide a summary of the competition design and our results.

Trojans are bugs that are implanted into a network to make it respond in a certain manner when a trigger is present. This is normally achieved through data poisoning [48, 87], where examples containing the trigger are added to the training set with the label changed to a target class. The competition added 16 trojans to a ResNet-50 model trained on ImageNet of three different types: *patch*, *style*, and *natural feature*. Patch trojans involve the superposition of a small image patch onto

an image. Style trojans use images modified via style transfer [74]. Natural feature trojans exploit real objects that may appear incidentally in images – such as forks, chairs, or carrots – but are not ImageNet classes. For example, in the competition, the trojan model has been trained in such a way that when a fork is present in the image, the model predicts the image to be a Cicada, regardless of the actual content[1].

Each trojan in the competition targets a specific class. So, when a trojan is added to an image in the training set, the image label is changed to the target class, regardless of the ground truth label of the image. Some trojans were universal, in that the trojan was added to a variety of classes, whereas some trojans were only added to a specific 'source' class. For details of the data poisoning process and of each specific trojan see the competition paper [38, 39].

Finding trojans mirrors the practical challenge of finding unknown bugs in models, but with the benefit that the causal effect of the triggers is known, and so methods which are designed to find bugs in deep learning models can be evaluated. The competition is composed of two challenges: Challenge 1 involves the rediscovery of 12 known trojans by human crowd-workers; Challenge 2 requires the discovery of 4 secret trojans that were not disclosed to competition participants.

Each team was given access to the poisoned model. For challenge 1, ground-truth trojan triggers and target classes were disclosed. Each team was required to submit explanations for each target class, which were subsequently evaluated by crowd-workers – human participants on the internet – by asking them to identify the trojan from a multiple choice list. Each crowd-worker was shown the top-5 textual concepts for a poisoned class (without access to images), and asked to identify the intended trojan class from a list of 8 options. For example, for the Sandwich trojan, participants were given a choice of: A. Salad, B. Pizza, C. Omelette, D. Sandwich, E. Spaghetti, F. Stir Fry, G. Nachos, and H. Waffle. The score for the different competition entries was the proportion of crowd-workers who correctly chose the trojan, i.e., Sandwich in this case.

[1]The triggers are not 100% successful, so the model actually only predicts Cicada 30.8% of the times a fork is present. See the competition paper for details [39].

In challenge 2, each team was told which ImageNet classes had been poisoned, but were not told what the trojans were. In this case the researchers had to submit what they believed the trojan to be based on the explanations they obtained using their interpretability method.

TextCAVs ranks concepts by model sensitivity to textual descriptions. However, to find trojans, we want to find unexpected concept sensitivities. So, we must remove concepts that we expect to be there for each class. This can be done either using human expertise or through comparison to a clean baseline. To support automated evaluation, we compared the poisoned model's sensitivity to a list of LLM-generated concepts against a clean ResNet-50 trained on ImageNet. For each poisoned class (i.e., the class to which the model is misled), we display the top-5 textual concepts with the greatest increase in sensitivity in the trojaned model compared to the clean model. These are shown left to right in descending order of sensitivity difference in Figure 4.2.

By visual inspection, we can see that there are mixed results, with some explanations relating to the trigger, and others not. This is corroborated by the human evaluations where crowd-workers were asked to identify the trojans using the TextCAV explanations, where the mean proportion of crowd-workers ($N = 100$) who correctly identified each trojan is 0.29. However, in the second challenge of the competition, TextCAVs excelled by finding all four secret trojans. Since challenge 2 was not evaluated by crowd-workers, we were not constrained to a fixed list of concepts, and could interactively explore hypotheses. Using the LLM generated concepts as an initial set, we tested additional concepts related to words that appeared unusual or unexpected for each class of interest. In doing this, we were able to predict each secret trojan (spoon, carrot, chair and potted plant) successfully. Each of the secret trojans were natural features, which, as we can see from part one of the competition, TextCAVs performs better on. If the secret trojans had been based on style transfer, it is unlikely we would have been able to predict them, even with extensive interaction.

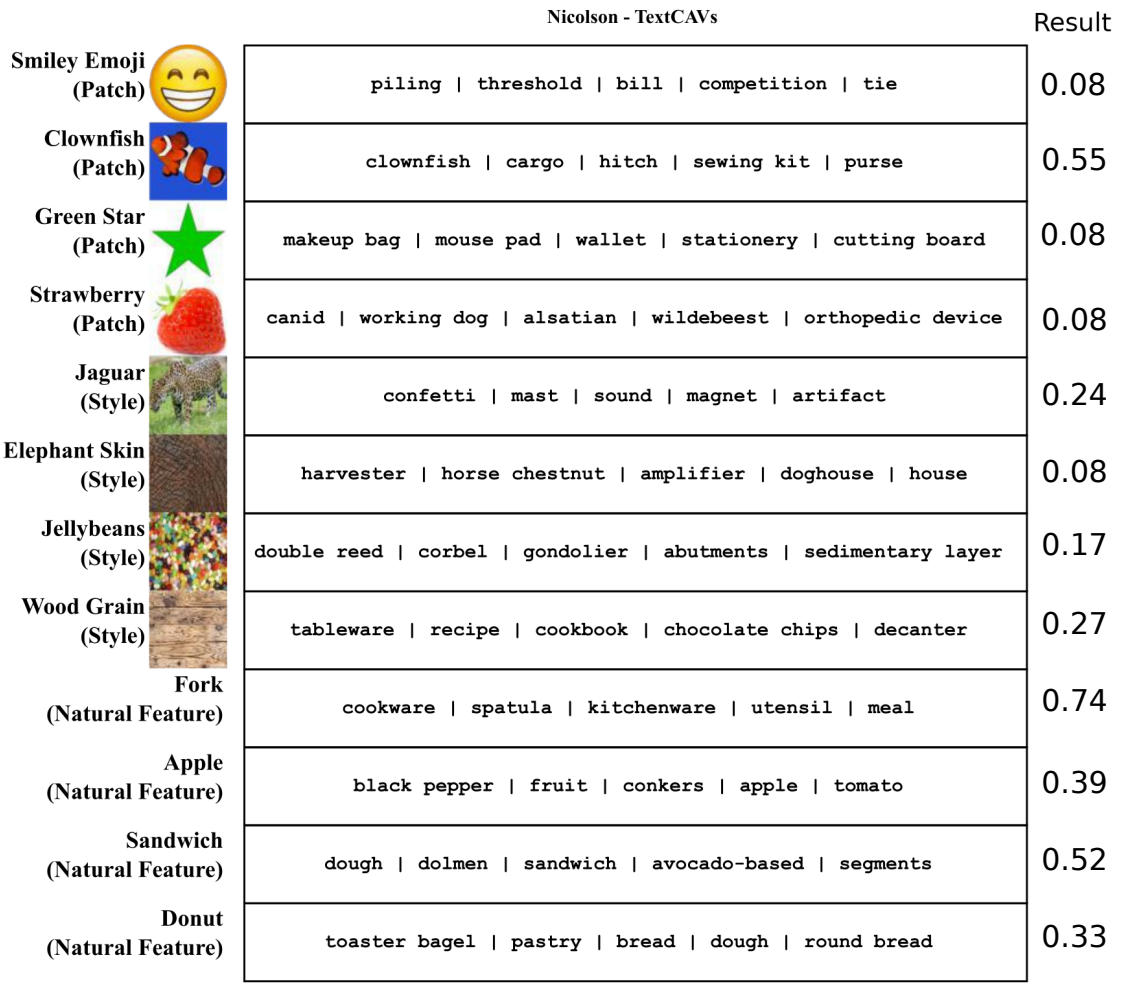


Figure 4.2: Mixed Competition results. Left: The trojan triggers. Middle: The top-5 TextCAV concepts per class (these should relate to the trojan trigger). Right: The proportion of crowd-workers who successfully identified the trojan.

4.4.2 MIMIC-CXR

In this section we demonstrate TextCAVs ability to produce meaningful explanations for a model trained on the chest X-ray dataset MIMIC-CXR and how we can use TextCAVs to discover bias in a model trained on a biased version of the dataset.

Training Details We train both the linear transformations, h and g, and the target model, Φ, using the MIMIC-CXR training set. The target model is a ResNet-50 [93] pretrained on ImageNet and then fine-tuned for the 5-way multi-label classification of chest X-rays with the classes: No Finding, Atelectasis (lung collapse), Cardiomegaly (enlarged heart), Edema (fluid in the lungs) and Pleural

Effusion (fluid between the lungs and the chest wall). We use the Adam optimiser [119] with weight decay of $1e-4$ and initial learning rate of $1e-4$. The learning rate is halved or the training is stopped if the validation loss does not decrease within 3 or 5 epochs, respectively. Images are resized to 256×256. We use random rotation of up to 15 degrees, random horizontal flipping, random crop and resize with a minimum size of 40%, and distortion to augment the images. We use the published data splits and, after removing images with no positive class labels, there are $368,945$ training, $2,991$ validation and $1,012$ test images. We use labels from CheXpert [103] for the training and validation labels, which have been generated by a model using the text reports. Whereas, for the test dataset, we use the provided labels annotated by a single radiologist.

We train both h and g on the training set of MIMIC-CXR for 20 epochs. We use the output of the average pool operation as the features from the target model as it simplifies the extraction of model gradients (Eqn. 4.7).

For Ψ, we use BiomedCLIP [240] – the current state of the art vision-language model for chest X-ray tasks.

Concepts The MIMIC-CXR dataset has a clinical report associated with each image. We use these reports as a source of concepts. We extract the sentences from the "FINDINGS" and "IMPRESSION" sections of the reports and use a random subset of 5000 sentences to obtain a wide variety of concepts to test.

Biased Data To evaluate TextCAVs as an interpretability tool we explore its usefulness in model debugging. We induced a dataset bias in the MIMIC-CXR training set by removing all participants with a positive label for Atelactesis and a negative label for Support Devices. This means that all participants with Atelactesis in the training set also had a Support Device (e.g. tube or pacemaker) as can be seen in Figure 4.3.

Metrics To provide a quantitative metric, we labelled the top-50 sentences for each class, ordered by directional derivative, on whether they relate to the class. We report this information as a concept relevance score (CRS), which is simply the proportion

of concepts that were related to the class. Using Edema as an example, a sentence was labelled as related if it directly diagnosed the class, e.g., "Worsening cardiogenic pulmonary edema", or if the class was implied, e.g., "bilateral parenchymal opacities" or "there is alveolar opacity throughout much of the right lung".

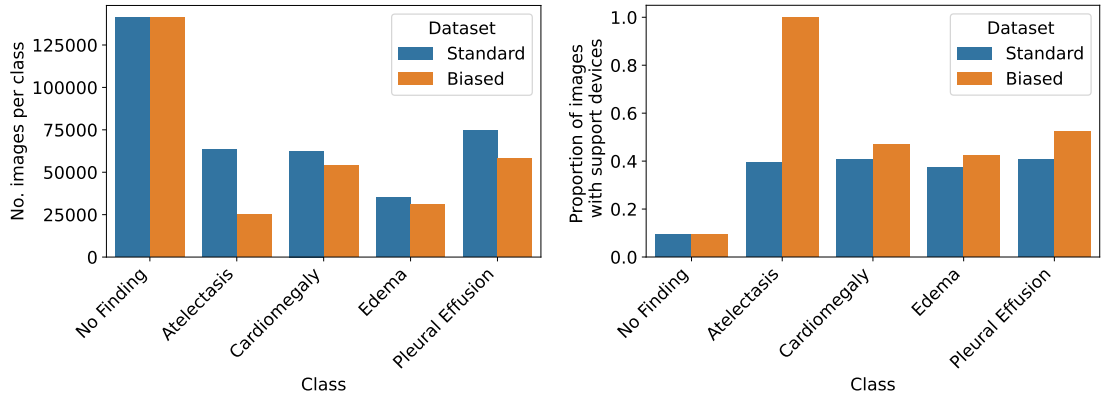


Figure 4.3: MIMIC-CXR dataset characteristics. Left: The number of images per class in the training set of the target models. Right: The proportion of training images that contain a support device for each class.

Table 4.2: Area under the receiver operator characteristic curve (AUC) and concept relevance score (CRS) for the standard and biased MIMIC-CXR models. AUC* was calculated on the biased version of the MIMIC-CXR test set. The low CRS for Atelectasis in the biased model means almost none of the top TextCAVs are relevant to the class, demonstrating that they can be used to detect if a model is using biased features.

Model	**Standard**		**Biased**		
Metric	**AUC**	**CRS**	**AUC**	**AUC***	**CRS**
No Finding	0.87	0.74	0.85	0.94	0.76
Atelectasis	0.73	0.56	0.68	0.81	0.04
Cardiomegaly	0.81	0.94	0.81	0.82	0.90
Edema	0.85	0.90	0.84	0.81	0.80
Pleural Effusion	0.89	1.00	0.88	0.88	1.00
Mean	0.83	0.83	0.81	0.85	0.70

Results We are comparing two models: one trained on the standard MIMIC-CXR dataset and the other trained on the biased version. We will refer to the models as "standard" and "biased", respectively. The standard model achieved a mean area under the receiver operator characteristic curve (AUC) of 0.83 and the biased model a mean AUC of 0.81. The individual class AUCs can be found in Table

4.2. We expect, and see that the biased version has higher performance on a biased version of the test set since Support Devices tend to be easy to detect. As evidence for this, we trained a reference model separately and achieved an AUC of 0.92 for Support Devices.

Table 4.3: Top-5 concepts ordered by directional derivative for the standard MIMIC-CXR model.

No Finding	**Atelectasis**	**Cardiomegaly**
The lungs are clear and the cardiac, mediastinal, and hilar contours are normal.	Nasogastric tube extends below the hemidiaphragm and out of view.	Marked cardiac enlargement as before and unchanged position of previously described metallic prosthesis of porcine type.
Normal chest radiograph with unremarkable appearance of the lung parenchyma and normal appearance of the heart and the mediastinal and hilar contours.	Interval placement of a basilar right sided pleural space pigtail catheter with improved small right pleural effusion and right medial lung base atelectasis.	Heart size continues to be mildly enlarged.
The trachea is slightly deviated to the right by the aortic knob, which is ill-defined.	Worsening of the left retrocardiac opacity likely secondary to increasing atelectasis and/or effusion.	The patient has undergone prior aortic valve replacement.
This could represent a granuloma or possibly a bone island in the rib itself.	There is persistent elevation of the left hemidiaphragm with evidence of Bochdalek hernia seen at the left lower hemithorax.	Dense retrocardiac opacity which could represent effusion, atelectasis, consolidation or a combination thereof.
There is a fracture of the upper most sternal wire, unchanged.	Stable opacification of the mid and lower right lung consistent with large loculated pleural effusions and adjacent atelectasis.	The heart continues to be enlarged with mild to moderate CHF.

In Table 4.3, we show the five sentences whose CAVs have the highest directional derivatives for the classes of No Finding, Atelectasis (lung collapse) and Cardiomegaly (enlarged heart). Some of these are clearly linked to the class in question (e.g. "The lungs are clear" for No Finding and "Heart size continues to be mildly enlarged" for Cardiomegaly) but there also sentences which do not relate to the classes (e.g. "Nasogastric tube extends below the hemidiaphragm" for Atelectasis or "There is a fracture of the upper most sternal wire" for No Finding). The noise present in the explanations could be due to several different causes: (1) the target model is using unexpected features in its classification; (2) the feature conversion between Φ and Ψ is not perfect (i.e., h); or (3) the inherent noise present in gradient vectors [203]. It is difficult to ascertain which of these is the cause but a

tool can still be useful even with noise present. Hence, we demonstrate TextCAV's ability to detect dataset bias that we induce in MIMIC-CXR.

Table 4.4 shows the top-5 sentences for a model trained on the biased version of MIMIC-CXR. The bias is apparent in the explanations, as the top-5 sentences for Atelectasis all refer to Support Devices, rather than to any concepts relating to the class itself. The CRS values in Table 4.2 also indicate the presence of bias: a CRS of 0.04 for Atelectasis for the biased model shows that almost none of the top-50 concepts contain reference to the class. To further quantify the difference between the two sets of explanations we also labelled whether they referred to Support Devices. For the class of Atelectasis, we found that 13/50 concepts were related to Support Devices for the standard model compared to 44/50 for the biased model, demonstrating that TextCAVs are sensitive to the difference in behaviour between the two models.

Table 4.4: Top-5 concepts ordered by directional derivative for the biased MIMIC-CXR model.

No Finding	**Atelectasis**	**Cardiomegaly**
Bronchial wall thickening is minimal.	ET and NG tubes positioned appropriately.	If cardiomegaly persists, the presence of a pericardial effusion could be excluded with echocardiography.
Hilar and mediastinal contours are otherwise normal.	ET tube, nasogastric tube, Swan-Ganz catheter, and midline drains are all in standard placements.	Worsening heart failure in the context of chronic atelectasis.
This could represent a granuloma or possibly a bone island in the rib itself.	Nasogastric tube extends below the hemidiaphragm and out of view.	The patient has undergone prior aortic valve replacement.
No discrete solid pulmonary nodule are concerning mass.	Impella LVAD and transvenous atrioventricular pacer leads unchanged in their respective positions.	Moderate-to-severe cardiomegaly and stigmata of previous mitral valve repair noted.
There is a fracture of the upper most sternal wire, unchanged.	Nasogastric tube has been placed that extends well into the stomach.	The heart remains moderately enlarged and the aorta remains unfolded and tortuous.

4.5 Conclusion

In this work we introduce TextCAVs, an interpretability method that, once two linear layers have been trained, can measure the sensitivity of a model to a concept

with only a text description of the concept. We show that TextCAVs produce reasonable explanations for models trained on both natural images (ImageNet [56]) a chest X-ray dataset (MIMIC-CXR [109]). As first demonstrated in the SaTML CNN interpretability competition [39], we show that TextCAVs can be used to debug models. We generated explanations for a model trained on a biased version of the MIMIC-CXR dataset and showed that explanations for the biased class substantially changed with most (44/50) concepts referring to the bias compared to just 13/50 for the unbiased model.

Once the linear transformations, h and g, have been trained, TextCAVs enables fast feedback when testing the sensitivity of different concepts. This makes it ideally suited for interactive debugging which we aim to study further in future work. Some of the concepts with a high directional derivative did not appear to be related to the class. In section 4.4.2 we state three possible sources of this: (1) unexpected features being used in the target model, Φ, (2) poor feature conversion between models, h, or (3) inherent noise present in gradient vectors, $\nabla\Phi_{b,k}$. In future work we will explore which of these have the greatest effect.

The previous chapter emphasised the need to test multiple layers when using TCAV. In this chapter we solely used the penultimate layer of the network. These two sentences are at odds with one another. The decision to use the penultimate layer was because it allows TextCAVs to generate explanations without requiring any image data. If earlier layers were used, example images of the class of interest would be required to calculate gradient values. We justify this with our results. We propose TextCAVs as a means of debugging models and demonstrate that it can succeed at this task. However, in relying on a single layer we may be subject to noise in the explanations, in future work we will examine TextCAVs performance across different layers of the target model.

All the world is made of faith, and trust, and pixie dust.

— J.M. Barrie, Peter Pan

5

Do explanations help sonographers?

Contents

In the previous chapter we evaluated TextCAV's ability to debug models, with both crowd-workers and the authors (i.e. engineers) using the system successfully. We now turn to evaluation of explanations shown to clinicians in a clinical decision support system. We focus on evaluating trust, reliance and performance, designing a 3-stage reader study so that we can determine (1) the clinicians' performance and decision process (2) the influence of model predictions on the clinicians (3) any additional influence from model explanations.

5.1 Introduction

Deep learning models have been shown to be powerful tools within healthcare, and in imaging are able to achieve performances similar to or surpassing domain experts [183, 228, 117, 50]. These models can improve clinician performance when used as advice in clinical decision-making [117, 141]. However, in many instances we have little or no ability to understand how models reach their decision – so called "black boxes" – and this may hamper trust in model predictions [146, 185]. To overcome this, Explainable Artificial Intelligence (XAI) has been proposed: here, explanations are provided alongside model predictions, so that trust by end users is enhanced and more detail is given to aid clinical decision-making, allowing clinicians to better understand how a decision was made [204, 14, 212, 55, 181]. While such explanations can also be used to facilitate debugging during model development [64, 158, 39], the use relevant to this paper is for clinicians to better understand how model predictions or decisions are reached.

There have been many assertions that XAI is required in high-risk scenarios, with an increasing number of researchers calling for XAI in healthcare [124, 137, 212, 91, 168]. However, the purported advantages deserve to be examined more closely. Recent studies have called into question the necessity of XAI in healthcare advocating "rigorous internal and external validation of AI models as a more direct means of achieving the goals often associated with explainability" [77]. Other work has reported the ineffectiveness of model explanations at finding spurious correlations [4], and how many saliency methods provide explanations that do not depend on their underlying model [3] or are inferior compared to specialised networks at locating medical abnormalities [9].

There have been a number of efforts evaluating XAI models through the measurement of some component of interpretability, such as faithfulness (how well the explanations match the causal behaviour of the model); sparsity (how simple the explanations are); simulatability (how well users can predict a model prediction from its explanations) and continuity (how much the explanations change with small perturbations to the input) [156, 204, 60, 37, 26]. However, the gold

standard of evaluating interpretability methods is measurement of performance using real human operators and real tasks [60].

In this regard, there are currently fewer studies [156, 46]. Yu et al. [233] demonstrate the heterogeneity of the effect of AI-tools on the decision-making of clinicians and that there are no clear predictors of which clinicians will respond favourably (such as years of experience), but they do not examine the effect of model explanations, only model predictions. Gaube et al. [75] show that task experts did not show a significant improvement in reviewing X-rays but non-task experts can benefit from model explanations. Other work has shown that explanations can sometimes have a negative effect, with explanations for incorrect model predictions causing clinician treatment decisions to get worse for antidepressant selection [105]. The effect of explanations on human behaviour in clinical decision-making can be difficult to predict with Nagendran et al. [152] showing no correlation between self-reported usefulness of XAI and influence of explanations on prescription decision-making. In our study we build upon these works, measuring how sonographers respond to both model predictions and explanations from a prototype-based XAI method which produces explanations in the form of images and heatmaps.

To determine if model explanations are beneficial, we must first clearly define a specific use-case and the purpose of the explanations, to test if they achieve their stated purpose. In this study we hypothesise that the use of XAI improves user trust (through the provision of explanations), leading to increased reliance on model estimates and improved user performance (since the explanations enhance available information, informing decision-making).

5.2 Methods

As our use-case, we examine gestational age (GA) estimation from fetal ultrasound. Many of the proposed clinical imaging applications of AI are in radiology so this provides a useful example, in particular as it has recently been shown that AI can be more accurate than current clinical practice (using fetal biometry) [128].

It has been noted that "there does not appear to be a consensus regarding a validation protocol, which hinders the progress of explainable ML research by making explainable methods incomparable" [204]. We aim for our study design to become a valuable tool in evaluating XAI methods for healthcare applications. Using a three-stage design, we can measure:

- The clinician's decision-making process without AI
- The influence of model predictions
- The additional influence of model explanations

5.2.1 The AGE Study

We designed the Algorithmic Gestational age Estimation (AGE) study to explore the decision-making process of clinicians when estimating GA from images without biometry and to understand how their behaviour changes with access to an XAI model. At the outset clinicians completed a questionnaire related to their clinical experience, opinions on AI, and demographic information. The study consisted of three stages, where at each stage a participant is asked to estimate GA from an ultrasound image of the fetal head but with successively more information.

In stage 1, participants were asked to estimate GA from an ultrasound image and to rank their confidence on a Likert scale. The participants were also asked to highlight the regions they found most useful for their estimate and to select the relevant features from a list of options: amniotic fluid, apparent size, appearance of brain, apparent difficulty in obtaining plane, ossification of the skull, position within the maternal pelvis/uterus, relationship between the size of the fetus and sector width, shadowing, and shape of skull. We provided a free-text option so that participants could add any additional features not listed.

After at least 24 hours participants were asked to complete stage 2, in a similar manner to stage 1, but a GA estimate from the model is provided alongside the ultrasound image. We asked again for a GA estimate and confidence, but

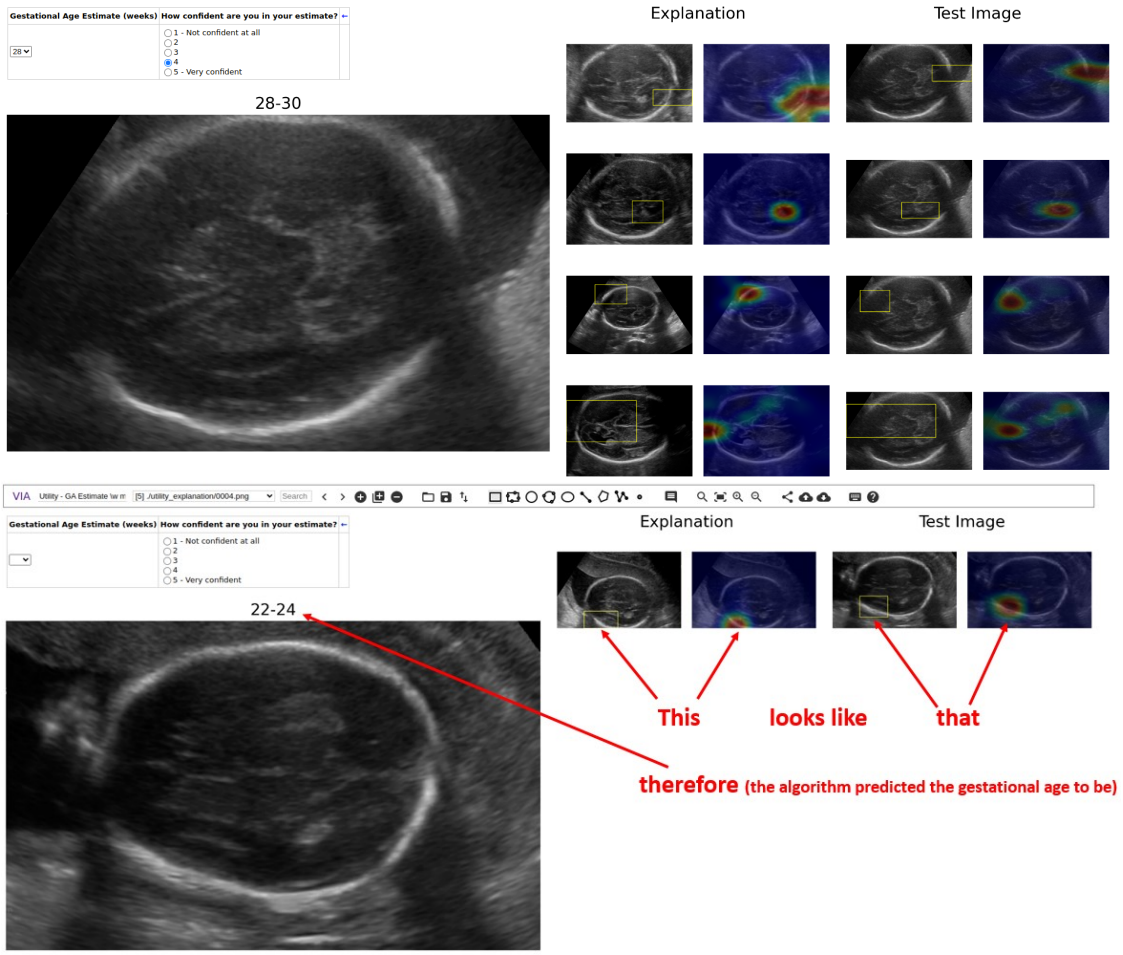


Figure 5.1: Top: Screenshot of the VIA software in stage 3 displaying the test image (left), model predictions (above the test image), questions for the participant (top left) and model explanations (right). Bottom: A second screenshot but with the method of interpreting the explanations overlaid.

participants were not required to label important features. We also asked questions on trust and how participants used the model estimates.

After a further 24 hours or more, stage 3 was undertaken in the same manner as stage 2, but with model explanations in addition to the predictions – see Figure 5.1 for an example. Again, questions related to their level of trust in the model and how they used the predictions/explanations were asked.

Participants received written instructions and demonstration videos for each stage, including information on how to interpret the model explanations in stage 3. We avoided dictating how participants should use the explanations to inform their estimates, to ensure we did not unduly influence how the clinicians interacted

with the system. Instead, we used the phrase "this looks like that" to explain how to interpret the explanations, i.e., regions highlighted in the training images look like regions in the test image and therefore the fetus is in the age range specified (see Figure 5.1 for an example).

In each stage, the participants examined 65 images chosen from the INTERBIO-21st dataset [114] to obtain an approximately uniform distribution of GA between 13-42 weeks (see Figure B.5 and Appendix B.2 for further details). The study was performed online using the VGG Image Annotator (VIA) software [63]. The study received ethics approval from a subcommittee of the University of Oxford Central University Research Ethics Committee (Reference: R85756/RE001) and all participants gave written, informed consent.

It is important to discuss the expertise of the clinicians at the task of gestational age estimation from fetal ultrasound. Sonographers typically estimate GA using biometry, where structures, such as the circumference of the fetal head, are measured and compared to a standard growth chart [165, 167]. This means the clinicians we recruit for the study are ideally suited for the task, in that they are the most qualified to estimate GA, but they are not trained to estimate GA from solely image characteristics, so cannot be considered experts at this particular task, because no experts exist.

We do not compare directly to biometry measurements in this work for both practical and methodological reasons. Practically, the data consisted of screenshots taken during the GA scan, with the sonographer free to zoom and pan; the resulting lack of known pixel dimensions made anatomical measurements unreliable. Methodologically, our aim was to evaluate AI as a decision support tool in workflows where biometry is not performed, so we compared against clinicians estimating GA without biometry.

More generally, the decision to include biometry in future studies should depend on the evaluation objective. If the goal is to assess whether AI can augment existing clinical workflows – for example, by assisting with or checking biometric measurements – then retaining biometry in the study design is essential. Conversely,

if the goal is to evaluate whether AI could replace certain tasks, such as GA estimation in the absence of biometry, then excluding biometry reflects that altered workflow more accurately.

5.2.2 Model Development

We use a single model throughout the study which outputs GA estimates and explanations. The estimates used in stage 2 are the same as those in stage 3, but in stage 3 the explanations are also displayed. For the XAI model we use an adaptation of an interpretable prototype-based deep learning model: prototypical part network (ProtoPNet) [45]. ProtoPNet classifies an image by calculating its similarity to a set of sub-parts of images from the training dataset and then weighting those similarities. This provides an explanation similar to how a clinician might make a prediction, e.g. "this fetus is 30 weeks of gestation, because it looks like a 30 week fetus I have seen before". The model is globally interpretable because the prototypes and weights are fully accessible. Its local explanations consist of the prototypes most similar to a test image and their corresponding contributions to the model output. Below, we provide a summary of changes we made to ProtoPNet to make it more suitable for GA estimation. For details, see Appendix B.1.1.

ProtoPNet is designed for classification tasks, but GA estimation is a regression task. As such, we split the GA range from 13-42 weeks into 13 bins of approximately two weeks (the first and last bins were larger to account for less samples in this region). This means the model gives estimates in two-week intervals rather than a single number (e.g. 18-20 weeks, or 32-34 weeks).

ProtoPNet encourages sparsity through regularising the fully-connected layer, but this is not sufficient to arrive at simple explanations, as the number of prototypes (65) far exceed the number of explanations that can be reasonably presented to a clinician. We enforce a sparse model by pruning weights below some threshold, τ, and performing fine-tuning. By setting a τ of 0.25, we display a mean of 80% of the model's reasoning with only four prototypes, compared to 42% for an unpruned model. For justification of this level of pruning see Appendix B.1.1.

The ProtoPNet model typically requires prototypes to be relevant to only a single class. Since GA estimation is a regression task converted into classification by binning the ages, our classes have some overlap in useful features. Hence, we remove the restriction that each prototype must be relevant to a single class, assuming prototypes are likely to be useful across a range of classes (Figure B.4 provides some evidence for this hypothesis).

5.2.3 Trust and reliance

In this section we make a distinction between two related ideas: trust and reliance. We define trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [127]. Whereas, in our context, reliance is the extent to which the model influences a participant's estimate. Trust is an attitude, whereas reliance is a behaviour. Trust guides, but does not completely determine, reliance. As we discussed in § 2.4.2, some works do not make a distinction between trust and reliance, but it is useful to clearly define a property before we attempt to measure it.

To measure participant trust in the model during stages 2 and 3 we use a questionnaire based on work by Hoffman et al. [99] to measure self-reported trust.

To measure reliance, we use two metrics evaluating participant agreement with model predictions. The first of these we simply name "agreement" and it is the proportion of estimates for which a participant's estimate was within the GA range the model predicted. An increase in agreement indicates increased reliance. The second is an established measure of reliance: Weight of Advice (WoA) [90, 81, 173, 61, 164, 61, 6]. Many authors refer to WoA as a measure of trust, but using our terminology, it measures reliance. First established to measure hindsight bias [95], it measures the degree to which advice influences a participants estimate. WoA is defined as:

$$\text{WoA} = \frac{\text{initial estimate} - \text{final estimate}}{\text{initial estimate} - \text{model estimate}} \tag{5.1}$$

WoA Value	Interpretation
< 0	The participant's estimate moved further away from the model's estimate.
0	The participant did not change their estimate.
0.5	The final estimate is the mean of the initial estimate and model estimate.
1	The final estimate matches the model's estimate.
> 1	The participant's estimate moved towards the model estimate, but they overshot, and their final estimate was beyond the model's estimate.

Table 5.1: An intuitive interpretation of different values of the weight of advice (WoA) metric. In general, a higher WoA indicates greater reliance, although this assumes that participants rarely overshoot.

Similar to Ahn et al. [6], we do not include datapoints in the calculation of WoA where $|\text{initial estimate} - \text{model estimate}| < 1$. The justification for this is the 2-week intervals the model uses for its estimates. These intervals mean that if a participant's estimate is within a week of the model's estimate, then the participant and model are in agreement. In general, a higher WoA indicates greater reliance, but for a more detailed description of the meaning of different WoA values see Table 5.1.

Blind reliance on an inaccurate model can lead to negative outcomes. Instead, we want to achieve appropriate reliance, where participants rely on the model when it is correct but ignore it when incorrect [127, 24, 224]. Previous works define appropriate reliance[1] using binary assignments of whether the model was correct or incorrect [230, 224]. For a regression task, how close to the ground truth does a model estimate need to be for it to be correct? Rather than imposing an arbitrary correctness threshold (e.g., a fixed distance to ground truth), which may not generalize across tasks, we define correctness relationally – comparing the model to the participant's own unaided estimate. If the model is closer to the ground truth than the participant, then the participant should rely on the model. If the model is less accurate, it is preferable to ignore it.

More concretely, we propose the following mathematical definitions of appropriate reliance, under-reliance, and over-reliance. Let y be the ground truth GA, $\hat{y}_{p_1}$ be a participant's estimate prior to observing information from the model, $\hat{y}_{p_2}$ be a participant's estimate after observing information from the model, and $\hat{y}_m$ be the

[1] even if they use a different term (e.g., appropriate trust)

model's estimate. The error for each estimate is the absolute difference with the ground truth:

$$\epsilon_{p_1} = |\hat{y}_{p_1} - y|,\ \epsilon_{p_2} = |\hat{y}_{p_2} - y|,\ \epsilon_m = |\hat{y}_m - y| \tag{5.2}$$

Let δ_1 and δ_2 be the absolute difference between the participant's estimate and the model estimate:

$$\delta_1 = |\hat{y}_{p_1} - \hat{y}_m|,\ \delta_2 = |\hat{y}_{p_2} - \hat{y}_m| \tag{5.3}$$

Let $\mathcal{R}$ be a binary value indicating reliance:

$$\mathcal{R} = \begin{cases} 1 & \text{if } \delta_2 < \delta_1 \\ 0 & \text{otherwise} \end{cases} \tag{5.4}$$

That is, $\mathcal{R} = 1$ indicates the participant moved closer to the model's prediction, i.e. they relied on the model. We define $\mathcal{E}_m$ to indicate whether the model estimate was more accurate than the participant's initial estimate:

$$\mathcal{E}_m = \begin{cases} 1 & \text{if } \epsilon_m < \epsilon_{p_1} \\ 0 & \text{otherwise} \end{cases} \tag{5.5}$$

We define reliance type, $\mathcal{R}_t$, by comparing behaviour ($\mathcal{R}$) to whether reliance was warranted ($\mathcal{E}_m$):

$$\mathcal{R}_t = \begin{cases} \text{Appropriate Reliance} & \text{if } \mathcal{R} = \mathcal{E}_m \\ \text{Under-reliance} & \text{if } \mathcal{R} = 0 \wedge \mathcal{E}_m = 1 \\ \text{Over-reliance} & \text{if } \mathcal{R} = 1 \wedge \mathcal{E}_m = 0 \end{cases} \tag{5.6}$$

It is worth restating that this definition does not assess whether the participant's estimate improved, but rather whether their behaviour was justified given the model's accuracy:

- **Appropriate reliance**: participant relied on the model when it was better, or did not when it was worse
- **Under-reliance**: participant did not rely on the model when it was better
- **Over-reliance**: participant relied on the model when it was worse

	$\mathcal{E}_m = 1$ (model better)	$\mathcal{E}_m = 0$ (model worse)
$\mathcal{R} = 1$ (relied)	Appropriate reliance	Over-reliance
$\mathcal{R} = 0$ (did not rely)	Under-reliance	Appropriate reliance

Table 5.2: The definition of appropriate reliance, under-reliance and over-reliance, where $\mathcal{R}$ indicates whether the participant relied on the model and $\mathcal{E}_m$ indicates whether the model's estimate was better than the participant's initial estimate.

Table 5.2 restates Equation (5.6) to make the distinction between the different types of reliance clear. For understanding which type of reliance is most prevalent in a study, as in work from Wang and Yin [224], we report the proportion of cases/images belonging to each reliance type (see Figure 5.6).

While our definition of appropriate reliance is tailored to regression, the framework could naturally extend to other settings, such as classification tasks with probabilistic outputs or situations where confidence scores are available. In such cases, reliance could be defined using shifts in predicted probabilities or confidence movements toward the model.

5.3 Results

A total of 10 clinical sonographers participated (Table 5.3) and evaluated 65 images each. All participants completed all rounds of the study.

Performance in GA estimation improved across the three stages (Figure 5.2, top left panel), but participants responded differently, with some performing worse in Stage 3 (Figure 5.2, top right panel) and agreement with model predictions differing between participants (Figure 5.6). Participant confidence improved as more information was available to them (Figure 5.2, center panel), but the confidence improvement between stages 2 and 3 was largely from participants who indicated they found the explanations helpful (Figure 5.5). The estimates of GA for individual participants became closer to model estimates once shown model predictions (Figure 5.2, bottom panels and Figure 5.6).

In stage 3, self-reported trust in the model decreased compared to stage 2 (Figure 5.3) but increases in agreement and WoA indicate a minor increase in reliance (Figure 5.6). Between stages 2 and 3, mean agreement increased from 70% to 73%

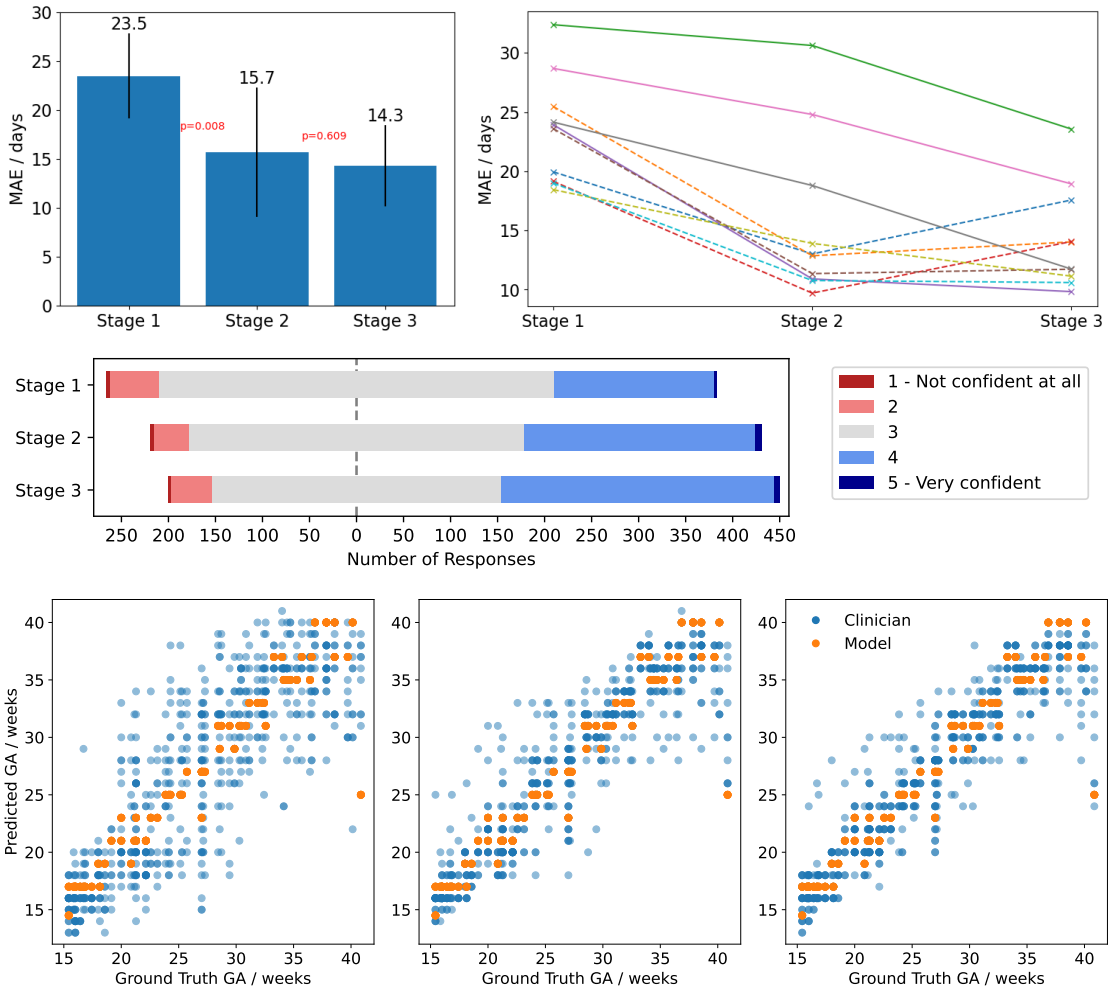


Figure 5.2: The mean absolute error (MAE) for gestational age at each stage for each participant are shown in aggregate (top left, p values are for adjacent stages) and for individual participants (top right, solid and dashed lines are participants who self-reported that the explanations were/were not helpful, respectively). The center panel shows self-reported confidence for GA estimates on a Likert scale over the three Stages. The bottom panels show estimated GA for participants (blue) and the model (orange) against the ground truth for Stage 1 (left), Stage 2 (middle) and Stage 3 (right).

and mean WoA increased from 0.64 to 0.70. This indicates that participants relied on the model, tending to move towards model estimates when available to them and that explanations caused a small additional increase in reliance.

Participants showed appropriate reliance in the model 66% and 69% of the time in stage 2 and 3, respectively (Figure 5.6). There was twice the level of under-reliance than over-reliance, with 12% over-reliance and 22% under-reliance in stage 2 and 10% over-reliance and 20% under-reliance in stage 3.

Age / years	<25	25-34	35-44	45-55	>55
	0	4	3	1	2
Job Title	Sonographer	Obstetric Registrar			
	9	1			
Region of the UK	East of England	London and the South East	Midlands	North West England	
	1	7	1	1	
Experience / years	< 2	2-5	6-10	≥ 10	
	1	2	3	4	
Obstetric scan frequency	Daily	Weekly	Fortnightly	Monthly	Infrequently
	7	2	1	0	0

Table 5.3: Participant demographic information, collected prior to stage 1. Experience relates to the number of years in fetal ultrasound.

There was a bimodal distribution in participant responses to questions relating to explanation usefulness, with some participants finding them helpful, and others not (Figure 5.4). Some participants had a less positive opinion on using machine learning in clinical practice after the study (Figure B.6). A wide variety of image features were used by participants to estimate GA (Figure B.7).

5.4 Discussion

5.4.1 Main Findings

In this study we examined the impact of XAI on trust and performance in ultrasound image-based GA estimation. Key findings include:

- A reduction in mean absolute error (MAE) for GA estimation with the inclusion of algorithmic estimates from 23.5 days to 15.7 days. Providing explanations slightly improved the MAE further to 14.3 days, although the improvement was not statistically significant.
- The benefit of explanations varied among participants. Some participants saw substantial improvements in their estimates, while others experienced a decline in performance.
- When given explanations, clinician confidence in their own estimates increases, alongside reliance metrics based on model agreement, but paradoxically self-reported trust in the model is reduced.

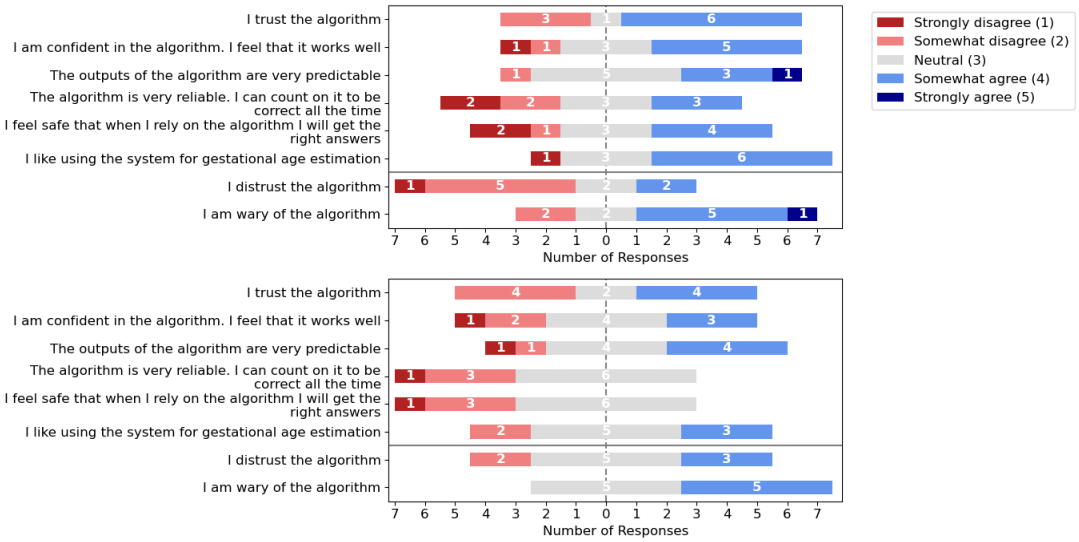


Figure 5.3: Reported trust in the model when participants had access to model explanations: "on a scale of 1-5, how much do you agree with the following statements?" for Stage 2 (top) and Stage 3 (bottom) on a Likert scale.

The improvement in GA estimation with access to model predictions is unsurprising due to the difference in accuracy of the clinicians and the model in isolation (a MAE of 23.5 days and 9.4 days, respectively). An appreciation of the task under consideration is important here. The participants had been asked to perform a task which, although related to clinical care, is not one they have been trained to do, and so they may be more likely to defer to AI advice. In fact, it is surprising they did not defer to AI guidance more often. Figure 5.6 shows that participants agreed with the model $70 \pm 29\%$ of the time with access to just the predictions and $73 \pm 22\%$ of the time with model explanations. However, this obscures important detail. Figure 5.6 shows that there were substantial differences between participants. For example, participant 1 had low agreement that did not substantially change (29% to 37%), participant 0 had low agreement that increased substantially between stages (49% to 98%), participants 4 and 6 had perfect agreement in both stages, and participant 8 had high agreement which dropped in stage 3 (97% to 69%). The variation in both the initial agreement in stage 2 and the change in agreement in stage 3 highlight the varied responses that clinicians can have upon receiving XAI advice.

The individualized response to both the model predictions and explanations

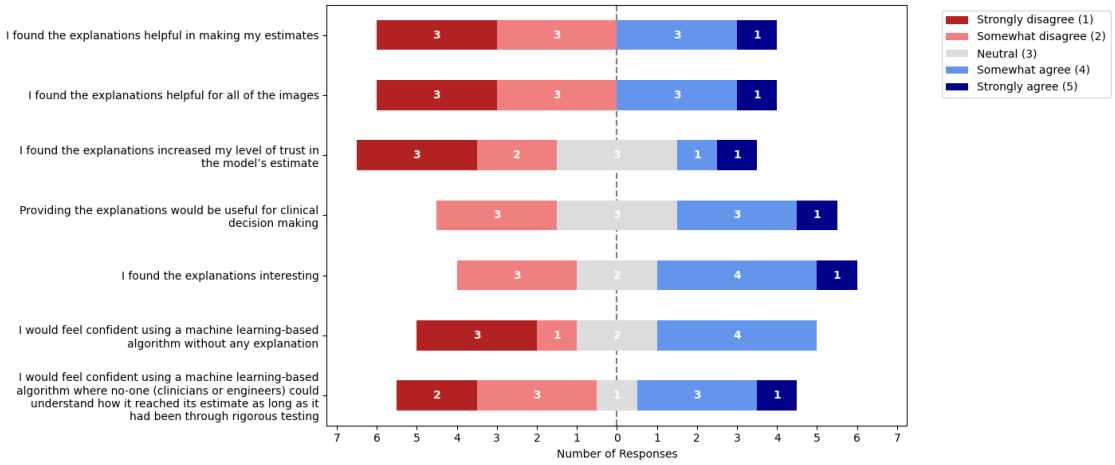


Figure 5.4: Bimodal opinions on how useful explanations are. Responses to "on a scale of 1-5, how much do you agree with the following statements?" immediately after Stage 3.

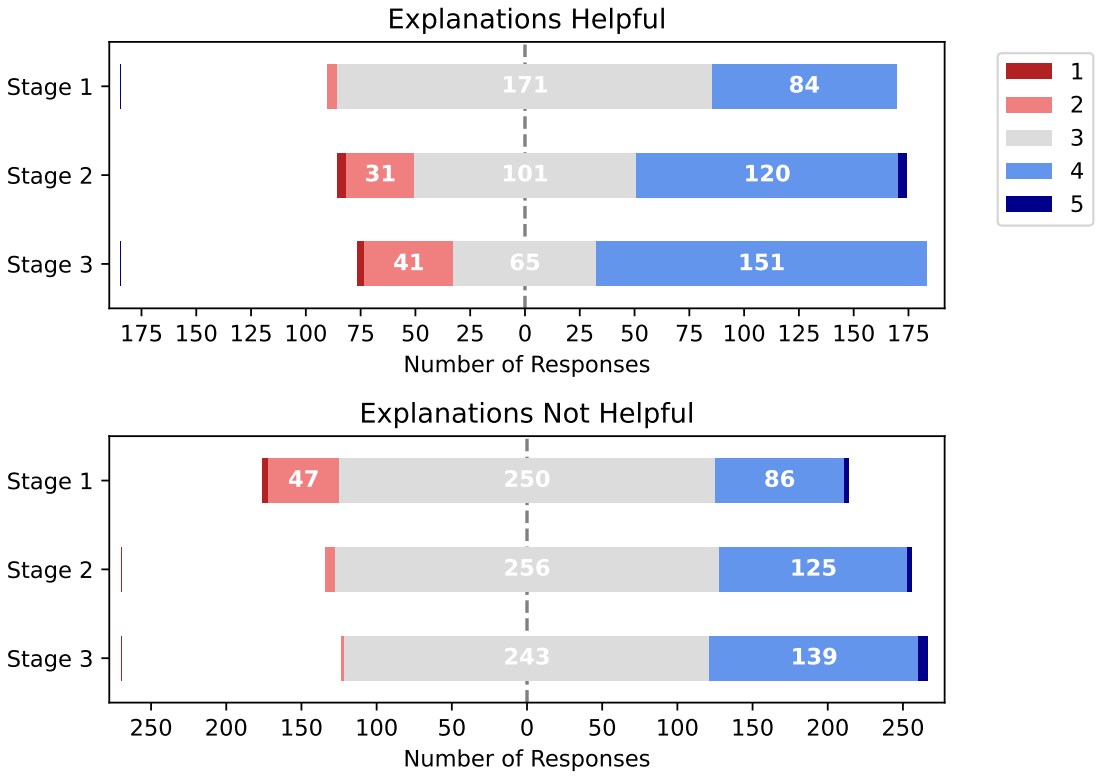


Figure 5.5: Participant confidence in their GA estimates – split by if the participant reported to find the explanations helpful.

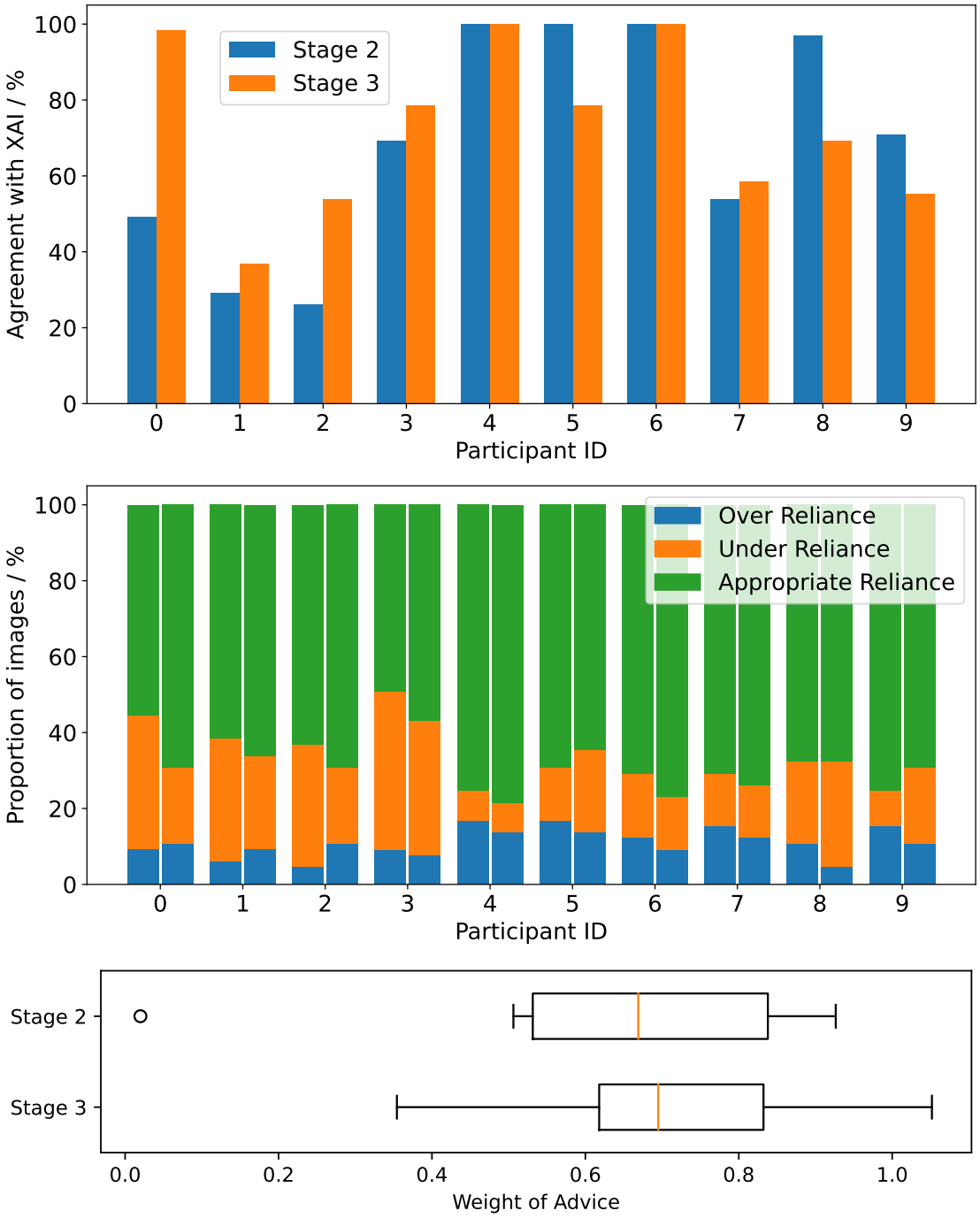


Figure 5.6: Top: Participant agreement with XAI predictions for stage 2 and 3, i.e. the proportion of the time the participants' predictions were within the model's suggested range of GA. Middle: The proportion of images for which participants showed over/under/appropriate reliance in the model for stages 2 and 3. Bottom: The mean Weight of Advice (measurement of reliance) of participants in Stage 2 and 3.

highlights the importance of studies such as this. Metrics which evaluate the performance or interpretability of an XAI model cannot shed light on how humans might respond to the advice or the differences between these responses. If XAI models are to be used in clinical practice, it is vital they are evaluated on the relevant clinicians who would use the tool and the clinicians' responses carefully examined. It is also vital that clinicians have a predictable and consistent response to the advice to ensure a consistent standard of care for patients. This is clearly not the case for the XAI model and experimental design chosen for our study. Other XAI methods and more thorough training of participants should be explored to reduce the noise we observed in participant responses.

Explanations were observed to reduce participant self-reported trust in the model (Figure 5.3). This is the antithesis of our hypothesis in the introduction and the common assertion that "by enhancing the interpretability of a system, trust from an expert user will also be enhanced" [181, 146]. However, reliance metrics (agreement and WoA) slightly increased in Stage 3 and the increase in appropriate reliance suggests this was not driven by over-reliance in the model. The discrepancy between decreased trust and increased reliance highlights the complexity of defining and evaluating trust. We hypothesise the reduction in perceived trust is caused by a mismatch between the explanations presented to the participants and their method of estimating GA by eye. This hypothesis is supported by the participants responses to "In what way did the explanations provided in stage 3 influence your decision-making?" in Table 5.4. Although some participants found the explanations "helped in estimation" and "improved my level of trust and understanding" others "found the explanations very confusing", "made me lose interest in what the algorithm thought" or noticed differences in how the model makes predictions compared to a clinician: "you cannot analyse that area with the naked eye but maybe the algorithm can".

Even though self-reported trust in the model decreased in Stage 3, just like reliance, participant confidence increased across stages (Figure 5.2). Figure 5.5 provides some nuance to this story. It indicates participants who found the explanations helpful had a different response to the participants who did not.

Participant ID	ΔMAE (Stage 3 – Stage 2)	In what way did the explanations provided in stage 3 influence your decision-making?
0	-7.1	It helped me see the area of the image the model was looking at and this improved my level of trust and understanding.
1	-7.1	The algorithm predicted where I should look make my decision on the gestational age. However, on occasions, it was completely off
2	-5.9	I would feel more confident scanning in real time and use a TCD as a guide.
3	-2.8	To be honest they didn't really - I couldn't understand why some explanations related to gestations completely different from what was being looked at.
4	-1.1	Helped in estimation.
5	-0.2	I found the explanations very confusing, the heat maps and boxes often bore no relationship to the test image so I found it difficult to use 'this looks like that'. Sometimes the box analysed an occiput and sometimes the frontal bone yet the test image was not in the same position. To assess areas in the anterior/ superior portion of the internal skull where there is so much artefact and reverberation you cannot analyse that area with the naked eye but maybe the algorithm can. It was confusing to analyse.
6	+0.4	I did not find them particularly useful, if the algorithm has been proven to be accurate I feel this is as much explanation as I need to use it confidently.
7	+1.2	The explanations did not make clear sense and therefore [I] relied on my own skill and estimates.
8	+4.4	They made me lose interest in what the algorithm thought so [I] did not reveal its predictions.
9	+4.6	It was interesting seeing visually how the AI concluded gestational age.

Table 5.4: Free text response to how useful the explanations were in stage 3 alongside the participants' change in mean absolute error (MAE) between stage 2 and 3 (a negative value indicates the MAE decreased in stage 3, which is an improvement).

The model predictions caused both groups to increase their confidence between stages 1 and 2. Between stages 2 and 3, however, the participants who found the explanations helpful became both more and less confident with more scores of 2 or 4 as opposed to 3. The remaining participants did not have a large change in confidence and seemed more neutral with most estimates receiving a middle score of 3. This suggests that the participants who found the explanations helpful were using them to calibrate their confidence, whereas the others may simply have been more confident in stage 3 because it is the third time they have completed the task.

5.4.2 Implications for Clinical Care

The study highlights potential benefits and pitfalls of deploying XAI models in clinical settings. The reduction in MAE suggests AI models can enhance clinicians' estimation accuracy, potentially leading to better patient outcomes. However, the decrease in self-reported trust despite improved performance/reliance indicates a need for better-designed explanations that align with clinicians' expectations and understanding. In addition, given the heterogeneous impact, more thorough training, or alternative explanation formats may be required for all clinicians to benefit.

5.4.3 Implications for Research

The findings provide several directions for future research. Future research should focus on improvement of explanations; developing more intuitive and useful explanations and possibly interactive reasoning. Further research should be done to investigate why some clinicians benefit from explanations while others do not, and what factors (familiarity with task, skill, bias, use of diagnostic guidelines, training) influence successful use of AI-assistance in clinical decision-making.

This study's main strength is that it underscores the importance of designing rigorous evaluation frameworks for XAI in clinical settings, moving beyond traditional performance metrics to consider human factors like trust and reliance. Using our definitions for the different types of reliance (appropriate/under/over), researchers can evaluate not just whether users are relying on an AI model, but whether the users' reliance is warranted. The study's design has real world relevance, in that it closely mirrors a real clinical scenario, providing valuable insights into how XAI might be integrated into everyday clinical practice. The comprehensive analysis is another strength, with a three-stage reader study that allowed for detailed examination of performance, trust and reliance, offering a nuanced understanding of XAI's impact.

One weakness of the study is that it was conducted with 10 sonographers, which may limit the generalizability of the findings; nevertheless, it seems that this was a sufficient number to see the differences in approach and value of XAI among this group of clinicians. It is important to note that the clinicians are not experts

at this task and their decisions will not be noise-free (see the varied estimates in stage 1 in Figure 5.2). The study also focused on immediate reactions to XAI; longer-term studies could be beneficial to understand how trust and performance evolve over time with continued use.

5.5 Conclusion

Our study reveals that while XAI has the potential to enhance performance in GA estimation, its impact on human trust for this task is complex and variable. Explanations provided by XAI models were found to both improve and hinder performance, depending on the clinician. These findings emphasize the need for studies involving human participants, and further research to refine XAI tools, ensuring they are both trustworthy and effective in clinical practice. Future work should both explore why the explanations caused some participants to trust the model less and have reduced performance, and should create XAI methods that have explanations which better match the internal reasoning of clinicians' when they make clinical decisions, hopefully leading to increased performance, user trust, and appropriate reliance.

If all you have is a hammer, everything looks like a nail.

— Abraham Maslow *The Psychology of Science*

6 Conclusion

In this thesis we demonstrate three important aspects of useful interpretability research:

1. (Chapter 3) The capabilities and limitations of current methods need to be understood to be able to safely interpret explanations.

2. (Chapter 4) By defining a specific purpose, we can drive development of useful interpretability methods with clear, measurable goals of what the explanations aim to achieve.

3. (Chapter 5) Human studies are required to understand how interpretability affects users. Through careful study design, we provide a template for determining the influence of explanations on user trust, reliance and performance.

In Chapter 3, we examine three properties of CAVs, demonstrating how they can cause misleading explanations and provide recommendations to mitigate the risks. In § 3.7, we demonstrate how to use these recommendations in a melanoma classification task, providing a guide for practitioners to use when interpreting CAV-based explanations. Having examined CAVs in detail, in Chapter 4, we propose TextCAVs, a novel concept-based interpretability method, and demonstrate that an engineer can use it debug deep learning models trained on natural images and

chest X-rays. For debugging, we needed a method where many concepts could be tested while the user explores issues with the model, but a certain level of noise was acceptable as long as the method aided the discovery of bugs. These requirements led to TextCAVs, a method not requiring any labelled data, demonstrating the benefits of application driven research. The definition of interpretability puts a focus on *human* understanding, yet many researchers do not perform any human evaluation. In order to determine if an interpretability method is *useful*, we must define its use and then measure its effectiveness at said use. Hence, in Chapter 5, we designed a 3-stage user study evaluating the effects of model predictions and explanations on trust, reliance and performance. Our study design can be used as a template for how to study the decision making processes of AI, humans and the interactions between them. We show that ProtoPNet can improve the performance of the human-AI team when advising sonographers measuring GA, but its explanations can reduce trust when they do not align with expectations, even while increasing appropriate reliance.

6.1 Limitations

Understanding CAVs (Chapter 3) fails to answer several important questions related to layer consistency. Why do different layers have different sensitivities? The chapter makes it clear that different layers can provide different explanations, but how do we know which layer to use? Or, more accurately, how do we convert the sensitivities of multiple layers into the sensitivity of the model? Then, once we have that, how does model sensitivity relate to whether the model used that concept in its prediction? A drawback with using all gradient based methods to explain models is that the gradient is a measure of how the model output would change for an infinitesimal change in a feature (pixels in the case of saliency maps or activations in the direction of a CAV for TCAV). It is not a measure of how much the model used that feature, even though this is often the question practitioners have and therefore how the explanations are interpreted. Future work should both explore how to

obtain model sensitivity, rather than layer sensitivity, and explore the underlying assumptions and drawbacks to using sensitivity as an explanation.

TextCAVs (Chapter 4) needs more thorough human evaluation and studies involving domain expertise from a clinician. Our work demonstrated that the method might be more useful when in an interactive system, as all four secret trojans were found when the user could interact with the model and repetitively query for explanations, whereas TextCAVs only achieved 3rd place when evaluated with fixed explanations and 100 crowd-workers. However, this needs confirming with more deliberately designed user studies, rather than ad-hoc experiments completed for a competition. In § 4.4.2, we observed noise in the explanations and discussed three potential sources: (1) the target model Φ, (2) the feature conversion h or (3) the gradients $\nabla\Phi_{b,k}$. We have two sets of experiments that could serve to verify these sources. The first of these attempts to address (3) by following advice from Chapter 3 and examining explanations from multiple layers. One of the benefits of TextCAVs is that it does not require any imaging data once the feature converters, h and g, have been trained. However, this partly occurs because we use the penultimate layer in the network where the gradients do not depend on the input images. Model gradients are inherently noisy [203], so by having only a single gradient value we are introducing this noise into our explanations. If we use TextCAVs on earlier layers, it will require example images for each class, but there will be a different gradient for each image and, through calculating a mean, we can reduce the noise present in the gradient signal. Another possible source of noise is that h does not adequately convert the meaning of the text that describes a CAV from the CLIP model to the target model (2). Hence, the second set of experiments involves increasing the complexity of h and g by adding some non-linearity, for example they could each be a small multilayer perceptron (MLP). If these two experiments reduce their respective source of noise (inherently noisy gradients and the quality of feature transfer between models), then the only source we do not reduce is the target model using unexpected features (1), which is the exact signal that we want to measure when debugging with TextCAVs.

One of the major limitations of the AGE study (Chapter 5) is the small number of participants. With 10 participants we could see the range of different responses that occur, but we cannot determine *why* some clinicians find explanations useful and others didn't. Future work should examine this question. Our results show that clinicians might trust models less when the model explanations do not match the clinicians' internal reasoning. Further research should compare different explanation methods and determine if the explanations which most match clinician explanations cause the greatest increase in trust.

6.2 Future work

We now move onto more general discussion of future work in the field of interpretable deep learning in the medical domain. Currently, there is a general understanding that interpretability is vital for machine learning models to be able to be used in clinical practice. In order to determine if this is true, studies need to designed comparing the utility of interpretability with other approaches. For example, Bansal et al. [16] compare displaying model uncertainty to model explanations and find uncertainty to be more effective in improving the performance of the AI-human team than explanations. Interpretability has many possible uses (see Chapter 2) but in each case researchers should not assume that no other methods have been designed for that purpose. Whether an interpretability method or another approach achieves that purpose better will only be found with human studies comparing them. For example, which of these methods improve clinician trust more: model uncertainty; global explanations of the concepts a model is sensitive to; or local explanations consisting of saliency maps? Currently, we do not know, and it is likely to depend on the context. Or, is there some more complex combination of techniques for optimal collaboration between AI and clinicians? For example, Desolda et al. [57] and Bansal et al. [16] both provide adaptive explanations: Desolda et al. [57] display explanations only when the user asks for them and Bansal et al. [16] display explanations for either the top-1 or top-2 classes depending on model uncertainty. We believe this is a promising direction of future research, allowing the balance of

different priorities by, for example, ensuring efficiency by not showing the explanation every time (clinician time is expensive), but still showing explanations for difficult examples (e.g., when the model or clinician is uncertain) to aid appropriate reliance. As part of these human-based evaluations we need to examine *how* explanations are being used so that they can be appropriately designed. For example, Buçinca et al. [32] argue that "the implicit assumption behind the design of most systems is that people will engage analytically with each explanation and will use the content of these explanations to identify which of the AI's suggestions are plausible and which appear to be based on faulty reasoning", however, as we saw in Chapter 5, some users may just ignore the explanations. By introducing different cognitive forcing functions to force users to spend more time evaluating the model explanations, the authors demonstrate that, rather than just the type of explanation, the overall system design can have a substantial impact on how users use explanations. We think the overall design of interpretability solutions is a vital part of interpretability research and future studies should put a greater emphasis on this design process.

We conclude this thesis by questioning whether interpretability is a necessity for deep learning to be used in clinical practice, as commonly asserted. Throughout this thesis we have discussed the potential advantages and the range of uses interpretability methods have, so there is no doubt that interpretability tools can provide utility. However, far too few studies explore the benefits of interpretability compared to other methods for a particular purpose. Regardless, for deep learning to be used in medical practice, does it not simply need to provide enough evidence that it works? Modern healthcare is based on evidence based medicine (EBM), where empirical evidence is required rather than mechanistic proof [143]. Researchers will sometimes state that "ML systems in medicine must have an explainable architecture, designed to align with human cognitive decision-making processes" [43] but fail to recognise that for some tasks clinicians cannot clearly explain the underlying reasoning of their own decisions [196, 94]. Or, the fact that the mechanism of action for some medications [180] and surgical procedures are not completely known [175, 170, 223]. Instead, EBM is about using the best available

evidence (e.g., randomised control trials and meta-analyses), combined with clinical expertise and patient values, to guide decision-making [188]. Hence, we agree with McCoy et al.'s [143] assessment that deep learning can find its way into clinical practice without the benefits of interpretability. Instead, in this thesis, rather than insisting interpretability is a *requirement*, we examine where interpretability can be *useful*, moving the field towards useful interpretability for medical imaging.

Appendices

Success in any endeavor requires single-minded attention to detail and total concentration.

— Willie Sutton
(aka, "Slick Willie", the famous bank robber)

A

Additional Details for Understanding CAVs

Contents

A.1 Consistency Proof

Let $\boldsymbol{a}_{l_1,i}$ be the activation vector in layer l for the input $\boldsymbol{x}_i \in \mathbb{X}$. Function f projects the activations in layer l_1 to layer l_2, where $l_1 < l_2$, i.e. $f(\boldsymbol{a}_{l_1,i}) = \boldsymbol{a}_{l_2,i}$. We assume that f is continuous and differentiable.

Let $\hat{\boldsymbol{a}}_{l_1}$ and $\hat{\boldsymbol{a}}_{l_2}$ be linearly perturbed activations in each of these layers:

$$\hat{\boldsymbol{a}}_{l_1} = \boldsymbol{a}_{l_1,i} + \boldsymbol{u} \tag{A.1}$$

$$\hat{\boldsymbol{a}}_{l_2} = \boldsymbol{a}_{l_2,i} + \boldsymbol{v} = f(\boldsymbol{a}_{l_1,i}) + \boldsymbol{v}, \tag{A.2}$$

where $\boldsymbol{u} \in \mathbb{R}^{m_{l_1}}$ and $\boldsymbol{v} \in \mathbb{R}^{m_{l_2}}$ are vectors of non-zero norm (since CAVs are directions in activation space) and m_{l_1} and m_{l_2} are the dimensions of layer l_1 and l_2, respectively.

For the two perturbations to have the same effect on the activations (and hence the model) it must hold that:

$$\begin{aligned} f(\hat{\boldsymbol{a}}_{l_1}) &= \hat{\boldsymbol{a}}_{l_2} \\ f(\boldsymbol{a}_{l_1,i} + \boldsymbol{u}) &= f(\boldsymbol{a}_{l_1,i}) + \boldsymbol{v} \end{aligned} \tag{A.3}$$

where we have substituted in Eq. A.1 and Eq. A.2. For $\boldsymbol{u}$ and $\boldsymbol{v}$ to be consistent for all possible activations they must be constant with respect to $\boldsymbol{a}_{l_1,i}$, i.e. we can assume that $\boldsymbol{u}$ and $\boldsymbol{v}$ are not functions of $\boldsymbol{a}_{l_1,i}$. If we rearrange Eq. A.3 to obtain an equation for $\boldsymbol{v}$, we obtain

$$\boldsymbol{v} = f(\boldsymbol{a} + \boldsymbol{u}) - f(\boldsymbol{a}). \tag{A.4}$$

where we have simplified the notation by writing $\boldsymbol{a}_{l_1,i}$ as $\boldsymbol{a}$. Let us differentiate Eq. A.4

$$\frac{d}{d\boldsymbol{a}}\boldsymbol{v} = \frac{d}{d\boldsymbol{a}}(f(\boldsymbol{a}+\boldsymbol{u}) - f(\boldsymbol{a})) \tag{A.5}$$

$$\mathbf{0} = \frac{d}{d\boldsymbol{a}}f(\boldsymbol{a}+\boldsymbol{u}) - \frac{d}{d\boldsymbol{a}}f(\boldsymbol{a}) \tag{A.6}$$

$$\mathbf{0} = f'(\boldsymbol{a}+\boldsymbol{u}) - f'(\boldsymbol{a}) \tag{A.7}$$

$$f'(\boldsymbol{a}+\boldsymbol{u}) = f'(\boldsymbol{a}) \tag{A.8}$$

which implies that the derivative of f, f', is periodic with period $\boldsymbol{u}$. This is a strong restriction on the form that f' can take. Let's integrate to find out what implications it has on the form of f. First, let's split the periodic function f' into its mean value and its oscillatory part:

$$f'(\boldsymbol{a}) = g'(\boldsymbol{a}) + M \tag{A.9}$$

where $M \in \mathbb{R}^{m_{l_2} \times m_{l_1}}$ is the mean value across one interval for each component of $f'(\boldsymbol{a})$ and $g'(\boldsymbol{a})$ is a periodic function with zero integral across a single period, i.e.

$$\int_{\mathbf{0}}^{\boldsymbol{u}} g'(\boldsymbol{a})d\boldsymbol{a} = \mathbf{0}. \tag{A.10}$$

If we integrate $g'(\boldsymbol{a})$ from $\boldsymbol{a}$ to $\boldsymbol{a}+\boldsymbol{u}$ (using a change of variables with $\boldsymbol{t} \in \mathbb{R}^{m_{l_1}}$) we find

$$\int_{\boldsymbol{a}}^{\boldsymbol{a}+\boldsymbol{u}} g'(\boldsymbol{t})d\boldsymbol{t} = g(\boldsymbol{a}+\boldsymbol{u}) - g(\boldsymbol{a}) \tag{A.11}$$

and by using Eq. A.10 we find

$$g(\boldsymbol{a}+\boldsymbol{u}) - g(\boldsymbol{a}) = 0. \tag{A.12}$$

Therefore g is periodic with period $\boldsymbol{u}$. If we now take the integral of $f'(\boldsymbol{a})$, we find

$$f(\boldsymbol{a}) = \int f'(\boldsymbol{a})d\boldsymbol{a} \tag{A.13}$$

$$f(\boldsymbol{a}) = \int g'(\boldsymbol{a})d\boldsymbol{a} + \int M d\boldsymbol{a} \tag{A.14}$$

$$f(\boldsymbol{a}) = g(\boldsymbol{a}) + M\boldsymbol{a} + \boldsymbol{b}. \tag{A.15}$$

Hence, f satisfies Eq. A.8 (and therefore Eq. A.4) if and only if it is composed of a periodic function with period $\boldsymbol{u}$ and a linear term. However, there are further restrictions upon f. Let $M = \mathbf{0}$ so that f is simply a periodic function. In this case

$$\boldsymbol{v} = f(\boldsymbol{a} + \boldsymbol{u}) - f(\boldsymbol{a}) \tag{A.16}$$

$$\boldsymbol{v} = f(\boldsymbol{a}) - f(\boldsymbol{a}) \tag{A.17}$$

$$\boldsymbol{v} = \mathbf{0} \tag{A.18}$$

which contradicts our non-zero assumption on the norm of $\boldsymbol{v}$. Hence we can obtain layer consistent vectors $\boldsymbol{v}$ and $\boldsymbol{u}$ if and only if f is composed of a periodic function with period $\boldsymbol{u}$ and a non-zero linear term M. We provide no proof as to whether this form of f can occur in practice in a neural network, however our empirical results in the main thesis and § A.4 suggest that it does not. Our proof holds generally, however, in the next three sections, we go into more detail for a linear function as it is a special case of Eq. A.15 and for the ReLU and sigmoid functions as they are common activation functions used in neural networks.

A.1.1 Special Case: Linear Function

One counter-example that at first look seems to contradict Eq. A.15 is a linear function. If f is a linear function, i.e. it conserves vector addition, then Eq. A.4 trivially holds:

$$\boldsymbol{v} = f(\boldsymbol{a} + \boldsymbol{u}) - f(\boldsymbol{a}) \tag{A.19}$$

$$\boldsymbol{v} = f(\boldsymbol{a}) + f(\boldsymbol{u}) - f(\boldsymbol{a}) \tag{A.20}$$

$$\boldsymbol{v} = f(\boldsymbol{u}). \tag{A.21}$$

The general result (Eq. A.15) requires that f be a combination of a periodic function, $g(\boldsymbol{a})$, and a linear function, $M\boldsymbol{a} + \boldsymbol{b}$, but we have just shown that $f = M\boldsymbol{a} + \boldsymbol{b}$ would also hold. This is because a function that outputs some constant value is a special case of a periodic function, where there is no minimal period as

all periods are valid. So, for the case where f is linear $g(\boldsymbol{a}) = \boldsymbol{c}$, where $\boldsymbol{c} \in R^{m_{l_2}}$. Or for a more specific example, in the case of $f = M\boldsymbol{a} + \boldsymbol{b}$, $\boldsymbol{c} = \boldsymbol{0}$. Hence, the result in Eq. A.15 generally holds and, for the case where f is linear, the only layer consistent vector $\boldsymbol{v}$ in layer l_2 for some vector $\boldsymbol{u}$ in layer l_1 is that same vector projected into layer l_2 by f, i.e. $f(\boldsymbol{u})$.

A.1.2 Example: ReLU Function

In a neural network, f often involves a rectified linear unit (ReLU), so below we find $\boldsymbol{v}$ when $f = \text{ReLU}$. Let $a_{l_1,i,j}$, v_j and u_j refer to the individual elements of $\boldsymbol{a}_{l_1,i}$, $\boldsymbol{v}$ and $\boldsymbol{u}$, respectively. By the definition of a ReLU activation:

$$f(a_{l_1,i,j}) = \max(0, a_{l_1,i,j}) = \begin{cases} a_{l_1,i,j} & a_{l_1,i,j} > 0 \\ 0 & a_{l_1,i,j} \leq 0 \end{cases} \tag{A.22}$$

So, Eq. A.4 becomes:

$$\begin{aligned} v_j &= \max(0, a_{l_1,i,j} + u_j) - \max(0, a_{l_1,i,j}) \\ &= \begin{cases} a_{l_1,i,j} + u_j - a_{l_1,i,j} & a_{l_1,i,j} + u_j > 0, a_{l_1,i,j} > 0 \\ a_{l_1,i,j} + u_j + 0 & a_{l_1,i,j} + u_j > 0, a_{l_1,i,j} \leq 0 \\ 0 - a_{l_1,i,j} & a_{l_1,i,j} + u_j \leq 0, a_{l_1,i,j} > 0 \\ 0 - 0 & a_{l_1,i,j} + u_j \leq 0, a_{l_1,i,j} \leq 0 \end{cases} \\ &= \begin{cases} u_j & a_{l_1,i,j} + u_j > 0, a_{l_1,i,j} > 0 \\ a_{l_1,i,j} + u_j & a_{l_1,i,j} + u_j > 0, a_{l_1,i,j} \leq 0 \\ a_{l_1,i,j} & a_{l_1,i,j} + u_j \leq 0, a_{l_1,i,j} > 0 \\ 0 & a_{l_1,i,j} + u_j \leq 0, a_{l_1,i,j} \leq 0. \end{cases} \end{aligned} \tag{A.23}$$

If $a_{l_1,i,j} + u_j > 0, a_{l_1,i,j} \leq 0$ or $a_{l_1,i,j} + u_j \leq 0, a_{l_1,i,j} > 0$ for any element j then there does not exist a $\boldsymbol{v}$ such that Eq. A.3 is true for all i, i.e., when either of these statements are true, you cannot have two vectors which have the same effect on the activations across layers for all possible inputs. And if we assume that the elements of $\boldsymbol{a}$ can take any value in practice then there exists no two layer consistent vectors across a ReLU function.

A.1.3 Example: Sigmoid Function

In this section, we consider the sigmoid activation: $f(x) = \frac{1}{1+\exp(-x)}$. For ease of notation, we drop i and j as they do not change, but a_{l_1} and a_{l_2} refer to $a_{l_1,i,j}$ and $a_{l_2,i,j}$, respectively. From Eq. A.3, the concept vectors are consistent iff

$$f(a_{l_1} + u) = f(a_{l_1}) + v \tag{A.24}$$

$$\frac{1}{1+\exp(-a_{l_1} - u)} = \frac{1}{1+\exp(-a_{l_1})} + v \tag{A.25}$$

Simplifying Eq. A.25 for v, we get:

$$\begin{aligned}
v &= \frac{1}{1+\exp(-a_{l_1} - u)} - \frac{1}{1+\exp(-a_{l_1})} \\
v &= \frac{(1+\exp(-a_{l_1})) - (1+\exp(-a_{l_1} - u))}{(1+\exp(-a_{l_1}))(1+\exp(-a_{l_1} - u))} \\
v &= \frac{\exp(-a_{l_1}) - \exp(-a_{l_1} - u)}{(1+\exp(-a_{l_1}))(1+\exp(-a_{l_1} - u))} \\
v &= \frac{1-\exp(-u)}{(\exp(a_{l_1})+1)(\exp(a_{l_1})+\exp(-u))}
\end{aligned}$$

This can be simplified further with partial fractions:

$$v = \frac{1-\exp(-u)}{(\exp(a_{l_1})+1)(\exp(a_{l_1})+\exp(-u))} \tag{A.26}$$

$$= \frac{\exp(-a_{l_1})}{(\exp(a_{l_1})+1)} - \frac{\exp(-a_{l_1} - u)}{(\exp(a_{l_1})+\exp(-u))} \tag{A.27}$$

$$= \frac{\exp(-a_{l_1})}{(\exp(a_{l_1})+1)} - \frac{\exp(-a_{l_1})}{(\exp(a_{l_1} + u)+\exp(-u))} \tag{A.28}$$

For a single v to exist which is consistent for all a_{l_1} it cannot depend on a_{l_1}. Since the left half of Eq. A.28 depends on a_{l_1}, the only way that v does not depend on a_{l_1} is if the right hand side cancels out the left. This only occurs when $\boldsymbol{u} = \boldsymbol{v} = \mathbf{0}$. Since $\boldsymbol{u}$ and $\boldsymbol{v}$ are directions in activation space, and hence have a non-zero norm, this is a contradiction. Therefore, for the sigmoid function, under no conditions does there exist layer consistent vectors.

A.2 Implementation Details

In this section, we provide general implementation details applicable to the whole paper. For details relating to individual experiments and additional results, see Sections A.4, A.5 and A.6.

A.2.1 Concept Activation Vectors

Background In [116], a statistical test, TCAV, determines whether the model's sensitivity to a concept is significant. The test compares a set of CAV scores found using a concept dataset with CAV scores found using random data. To do this, we must find multiple CAVs for each concept. In practice, each of these CAVs is trained with the same positive set, $\mathbb{X}_c^+$, but a different random set, $\mathbb{X}_c^{r-}$, where $r \in 1, 2 \dots R$ denotes the random index. A CAV corresponding to a specific random index is labelled $\boldsymbol{v}_{c,l}^r$.

Implementation Details In this work, we create multiple CAVs per training run (30 unless otherwise stated), each using the same positive probe dataset but a different random set. We label a CAV trained with a specific random set as $v_{c,l}^r$, where $r \in 1, 2 \dots R$ denotes the random index. Random CAVs are generated from pairwise combinations of random data sets, and we conduct a two-sided Welch's t-test to test whether the means of concept and random TCAV scores are equal. If a set of CAVs passes this test with a p value less than 0.01[1], we consider the concept meaningful. We refer to the mean TCAV score of the random CAVs as the null; it acts as the TCAV score all other CAVs should be compared against to understand their sensitivity to the model. The null is often very close to 0.5, simplifying the interpretation of the TCAV score to the concept having positive sensitivity when greater than 0.5 and negative when less.

[1] we use a threshold of 0.01 to help reduce the false discovery rate

Table A.1: The number of each channels for the models trained on the simple, standard and spatial versions of the Elements dataset.

	Model		
Layer	Simple	Standard	Spatial
layers.0	64	64	64
layers.1	64	64	64
layers.2	64	64	128
layers.3	64	128	256
layers.4	64	128	256
layers.5	64	128	256

A.2.2 Elements

Classification Model The model architecture is a simple convolutional neural network with six layers: each layer contains a convolution, batch norm and ReLU, followed by an average pooling and fully connected layer to give the logit outputs. The first three convolutional layers utilise a max-pooling operation to reduce dimensionality. We train the model using Adam [119] with a learning rate of 1e-3 until the training accuracy is greater than 99.99%, giving a validation accuracy of 99.98% for the standard dataset. We use a different number of channels for the models trained on different datasets. This allows us to provide more model capacity when needed. The number of channels per layer for each model/dataset is summarised in Table A.1.

The models for datasets $\mathbb{E}_2$ and $\mathbb{E}_3$ in section 3.6.2 are the same architecture as for the simple dataset ($\mathbb{E}_1$).

Probe Dataset For Elements, the probe datasets are generated so that the positive examples for a concept contain only objects with that concept, so, for example, a red concept image will contain four objects with random shapes and textures that occur within the dataset, but all of them will be red. The negative set consists of random samples from the dataset.

A.2.3 ImageNet

ImageNet is used to demonstrate the experiments on a real-world application.

Classification Model We use the default weights for a ResNet-50 [93] in the TorchVision package in PyTorch, which used a variety of data augmentation techniques including Mixup [237], Cutmix [235], TrivialAugment [151], and Batch Augmentation [98].

Probe Dataset Most probe datasets used to train CAVs were collated from the Broden dataset [19], particularly focusing on textures such as `striped`, `meshed` or `dotted`, or objects such as `car`, `sea` or `person`. Some concepts were manually curated, such as the `anemone` concept, which was collected from test images of the 'anemone fish' class from ImageNet that were not used elsewhere in the experiments. Examples of some of these concepts are available in Figure A.1.

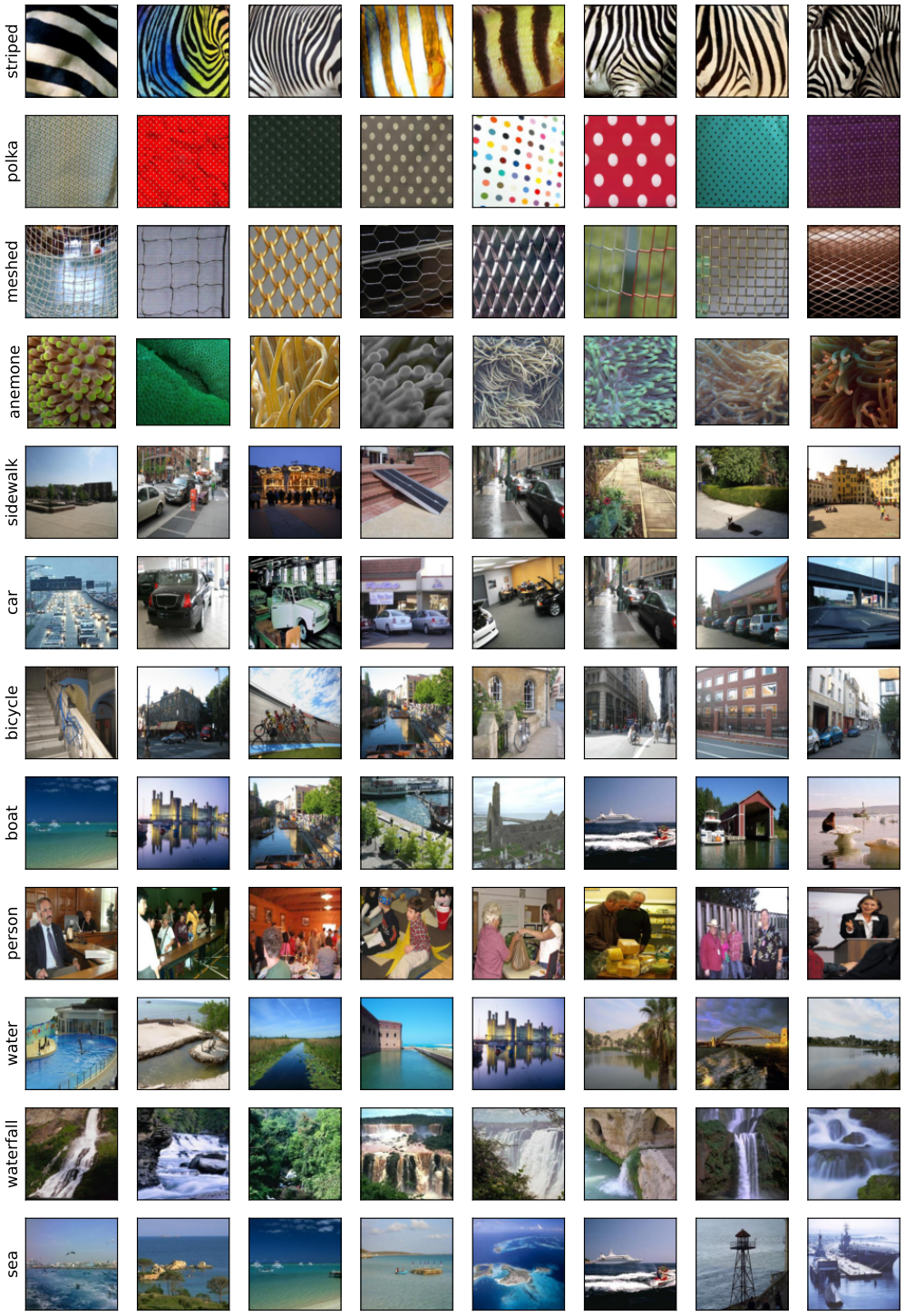


Figure A.1: Example positive probe datasets for different concepts for ImageNet.

A.2.4 Layer Selection

When creating CAVs, we need to choose a model layer. For the small CNN used for Elements, we can create CAVs for all layers, but for larger models, such as the

ResNet-50 for ImageNet, it is computationally infeasible. In this paper, we focus on layers near the end of the model. The justification for this is twofold. (1) From an information theory perspective, the activations earlier in the network may contain more irrelevant information, suggesting the activations closer to the output may be more relevant to the prediction [227**?**]. We aim to use TCAV to explain the model output, therefore later layers may be more desirable. (2) The model representations may be more complex in later layers. This allows us to create CAVs for more complex concepts. We find that the empirical evidence supports these hypotheses. Figures A.2 and A.3 show the accuracy of the linear classifiers used to create the CAVs on a held-out test set for each probe dataset. The accuracy for each concept tends to increase in later layers. This suggests the CAVs better capture the model representations in later layers. However, we observe variation across the concepts. For example, the colours in Elements are easily classified in all layers, whereas the shapes/textures have lower performance in layers.1. As our goal is to understand the behaviour of concept vectors (when the concept is represented), we focus on CAVs that obtain at least 90% test accuracy. Therefore, we omit layers.1 in our analysis.

We do not create CAVs for the final convolutional layer in either the simple CNNs for Elements (layers.5) or the ResNet-50 for ImageNet (layer4.2) due to the gradient behaviour in these layers. In both cases, the network has no non-linearities after the layer. Therefore, the gradient of the logit with respect to the activations solely depends on the model weights, not the activations. TCAV relies on having a distribution of directional derivatives, which are then thresholded and averaged over different data points. For these layers, the gradient is the same for all inputs, and hence so is the directional derivative. This means the TCAV score for an individual CAV in these layers will be exactly 1 or 0. As such, we do not perform TCAV on layers after which there are no non-linearities.

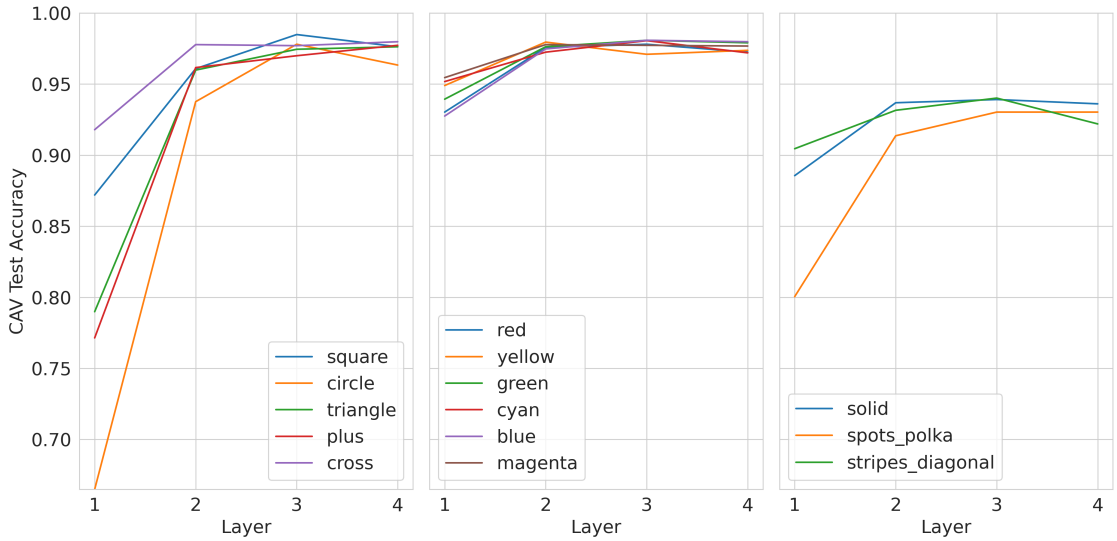


Figure A.2: Mean test accuracy for the linear classifiers from which the CAVs are generated for all concepts in the standard Elements dataset (split by concept type).

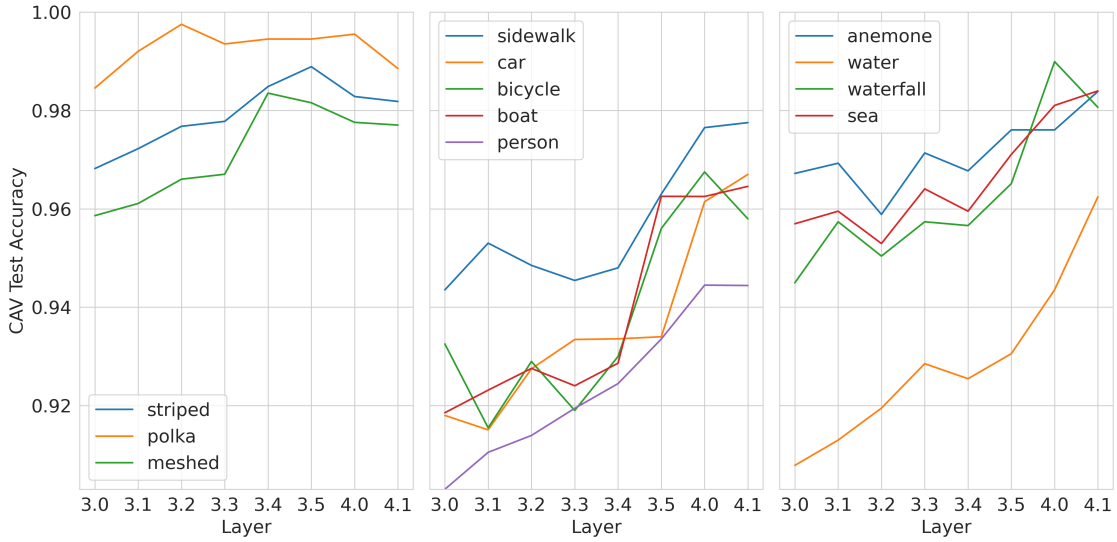


Figure A.3: Mean test accuracy for the linear classifiers from which the CAVs are generated for a selection of concepts in ImageNet.

A.3 Elements Dataset

A.3.1 Benefits of the Elements Dataset

Configurable datasets The configurable nature of the dataset allow us to explore different properties of the model and of CAVs. For example, we can introduce an association between the red and striped concepts in the training set by requiring

that all red elements are striped; or we can create a probe dataset of red elements on the right of the image to explore CAV spatial dependence.

Ground truth model behaviour The classes are configured as combinations of the elements' shape, colour and texture. Therefore, by construction, we have the ground truth relationship between each concept and class. As these relationships are within the dataset, knowing the ground truth relationship between each concept and class does not allow us to explore model faithfulness. For that, we need the ground truth influence of each concept on model predictions. By having a class for each possible combination concepts, the model must learn linearly separable representations of the concepts in the representation space (before the final linear layer) of the NN to achieve a high accuracy. Therefore, we have the ground-truth relationship of how each concept influences the model's predictions and can explore the faithfulness of concept-based explanation methods.

A.3.2 Elements Configuration

In the Elements dataset, there are various attributes we can vary. These attributes come in two types – image attributes and element attributes. The image attributes are: the number of elements per image and the size of the image. The element attributes are: colour, brightness, size, shape, texture, texture shift, and x and y coordinates within the image. Most are self-explanatory from their name, but texture shift requires more explanation. It is a small change in how the texture is applied to the object so that, for example, spots are not always in the same location with respect to the edge of the object. In this section, we describe the values that each of these attributes can take for the different versions of Elements that are used in the paper.

Standard The default version contains four objects in each image and the allowed concepts are the primary colours and their pairwise combinations (red, green, blue, yellow, cyan, magenta), five shapes (square, circle, triangle, plus, cross) and three textures (solid, spots and diagonal stripes).

Simple In some experiments, when stated, we use a simpler version which contains fewer shapes and colours. This is to reduce the complexity of the figures. The standard and simple dataset configurations are in Table A.2.

Spatially Dependent For some experiments in section 3.6.3 we use a spatially dependent version of the standard Elements dataset. This has the same configuration but introduces spatially dependent classes. As in the standard dataset, there is a class for all combinations of two and three concepts, e.g. 'striped squares' or 'spotted cyan triangles'. However, there are additional classes which depend on where the element is present in the image. For all classes involving triangles, we add two new classes which depend on if the object is in the top or bottom half of the image. For example, the class 'blue triangles' will now have two additional classes related to it of 'blue triangles on the top' and 'blue triangles on the bottom'. Similarly, we introduce two new classes for all classes involving squares, but for the left/right halves of the image rather than the top/bottom.

Entangled We use two alternative versions of the simple dataset in the Entanglement experiments: $\mathbb{E}_2$ and $\mathbb{E}_3$. As described in the main text, these are identical to the simple dataset, apart from the association between the red and triangle concepts. In each case, we restrict some of the allowed combinations of concepts that an element can take. This also removes some classes from the dataset which we reflect in any trained models. $\mathbb{E}_2$ does not allow any shape apart from triangles to be red. This removes classes like 'red circles' or 'spotted red squares'. $\mathbb{E}_3$ has this restriction and then places a further restriction that triangles have to red. This removes classes such as 'blue triangles' or 'striped green triangles'.

Table A.2: The configurations for the standard and simple versions of the elements dataset. Element size and image size are in pixels. Brightness is the value of the pixels in the element.

Property	Dataset	
	Standard	Simple
Colours	red, green, blue, yellow, cyan, magenta	red, green, blue
Shapes	square, circle, triangle, plus, cross	square, circle, triangle, plus
Textures	spots, stripes	spots, stripes
Brightness	153-255	153-255
Element Size / pixels	48-80	48-80
No. Elements per image	4	4
Image Size / pixels	256	256

A.3.3 Examples

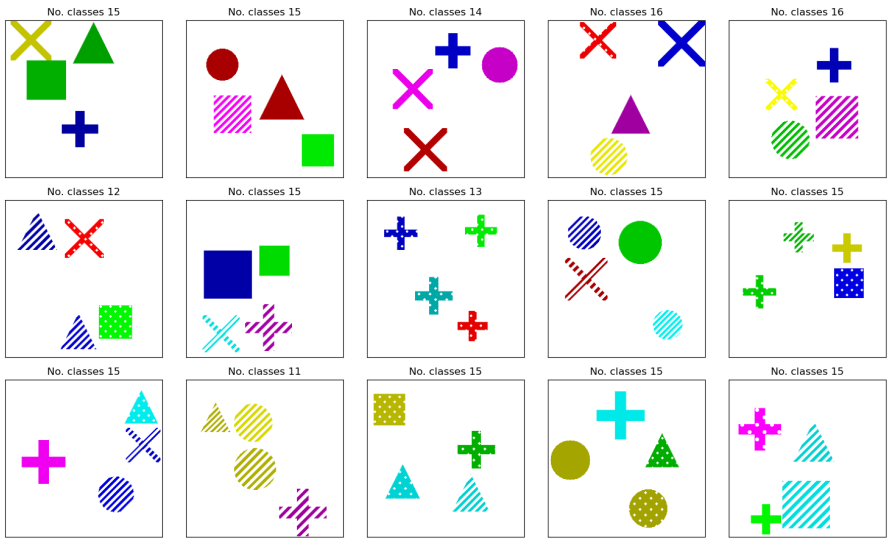


Figure A.4: Example images from the standard elements dataset. The number of classes each image belongs to is displayed above it.

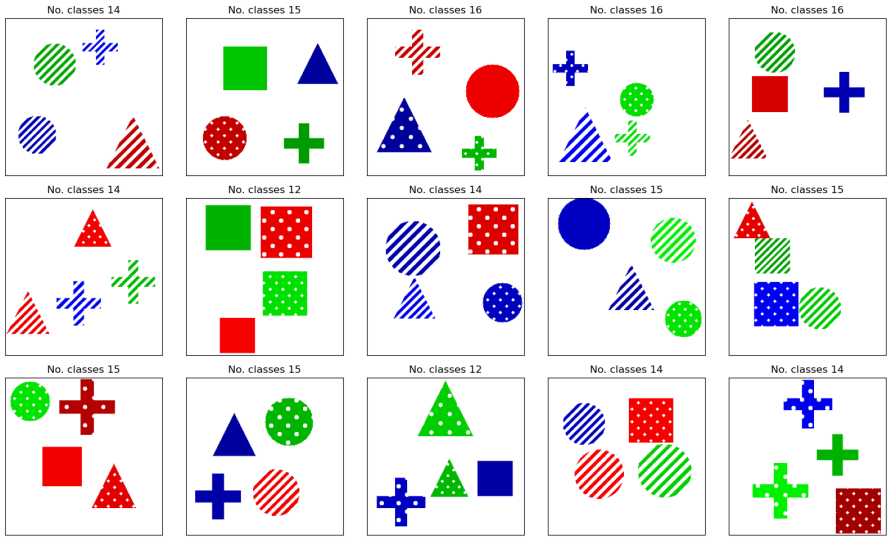


Figure A.5: Example images from the simple elements dataset. The number of classes each image belongs to is displayed above it.

A.4 Consistency Experiment Details

Figure 3.3 shows the consistency error for various types of CAV. In this section, we describe how we find the different types of CAV. In each case, $\boldsymbol{v}_{c,l_1}$ is a CAV trained as normal using a probe dataset, whereas the creation method for $\boldsymbol{v}_{c,l_2}$ varies for each experiment and is described below:

Optimised CAV We use gradient descent to optimise $\boldsymbol{v}_{c,l_2}$ to minimise the consistency error:

$$\arg\min_{\boldsymbol{v}_{c,l_2}} ||f(\boldsymbol{a}_{l_1} + \boldsymbol{v}_{c,l_1}) - (\boldsymbol{a}_{l_2} + \boldsymbol{v}_{c,l_2})|| \tag{A.29}$$

The starting point for each optimisation process is a CAV in layer l_2, $\boldsymbol{v}^{r_2}_{c,l_2}$, trained on a different random probe dataset. The nearly identical errors for each optimisation process support the likelihood of a global minimum being reached.

Concept CAV These are simply normal CAVs trained as described in Sections 3.2 and A.2, i.e., the CAVs are trained using a probe dataset containing $\mathbb{X}^+_c$ and $\mathbb{X}^-_c$. The distribution is over r_2 for different random probe datasets $\mathbb{X}_{c,r_2}$ for $\boldsymbol{v}_{c,l_2}$, where $r_2 \neq r_1$ denotes the random set.

Projected CAV CAVs in layer l_1 projected into layer l_2 using f: $f(\boldsymbol{v}_{c,l_1})$. The distribution is over different CAVs in layer l_1, i.e. $\boldsymbol{v}^r_{c,l_1}$.

Random CAV CAVs trained using a random probe dataset for both the positive and negative sets:

$$\begin{aligned} \mathbb{X}_c^- &= \mathbb{X}_{c,r_1}^- \\ \mathbb{X}_c^+ &= \mathbb{X}_{c,r_2}^-, \quad r_2 \neq r_1 \end{aligned} \tag{A.30}$$

Random Direction Each element of $\boldsymbol{v}_{c,l_2}$ is drawn from a uniform distribution between [-0.5, 0.5], and rescaled to be a unit vector. The distribution is over different random seeds for the random number generator.

A.4.1 Scaling perturbations

To ensure that $\boldsymbol{a}_l + \boldsymbol{v}_{c,l}$ stays in-distribution, we scale the perturbations as follows:

$$\hat{\boldsymbol{a}}_l = \boldsymbol{a}_l + \gamma \boldsymbol{v}_{c,l} \frac{\overline{||\boldsymbol{a}_l||}}{||\boldsymbol{v}_{c,l}||}, \tag{A.31}$$

where γ is a hyperparameter used for perturbation size (typically set to 0.01), $||\cdot||$ the L_2 norm of a vector, and $\overline{||\boldsymbol{a}_l||}$ the average norm of $\boldsymbol{a}_l$. We scale the perturbation by the mean activation norm to account for the difference in the norms between the activation and concept vector to have consistently sized perturbations across layers.

We performed experiments to explore the sensitivity of consistency error to the size of the perturbation, γ, for various layers and concepts for the ImageNet and Elements datasets. In Figures A.6 and A.7, we show the results for a variety of concepts for both the ImageNet and Elements datasets. Similar patterns were observed across experiments. As we increase $|\gamma|$, the consistency error scales linearly, as a larger perturbation causes a larger difference between $f(\hat{\boldsymbol{a}}_{l_1})$ and $\hat{\boldsymbol{a}}_{l_2}$. The scale of the y axis on the left of Figures A.6 and A.7 is not particularly meaningful without context, so we scale it by the norm of the perturbation in layer l_2, $||\gamma \boldsymbol{v}_{c,l_2} \frac{\overline{||\boldsymbol{a}_{l_2}||}}{||\boldsymbol{v}_{c,l_2}||}||$, on the right of the same figures. Values greater than one mean that the difference between $f(\hat{\boldsymbol{a}}_{l_1})$ and $\hat{\boldsymbol{a}}_{l_2}$ are larger than the perturbation made in layer l_2.

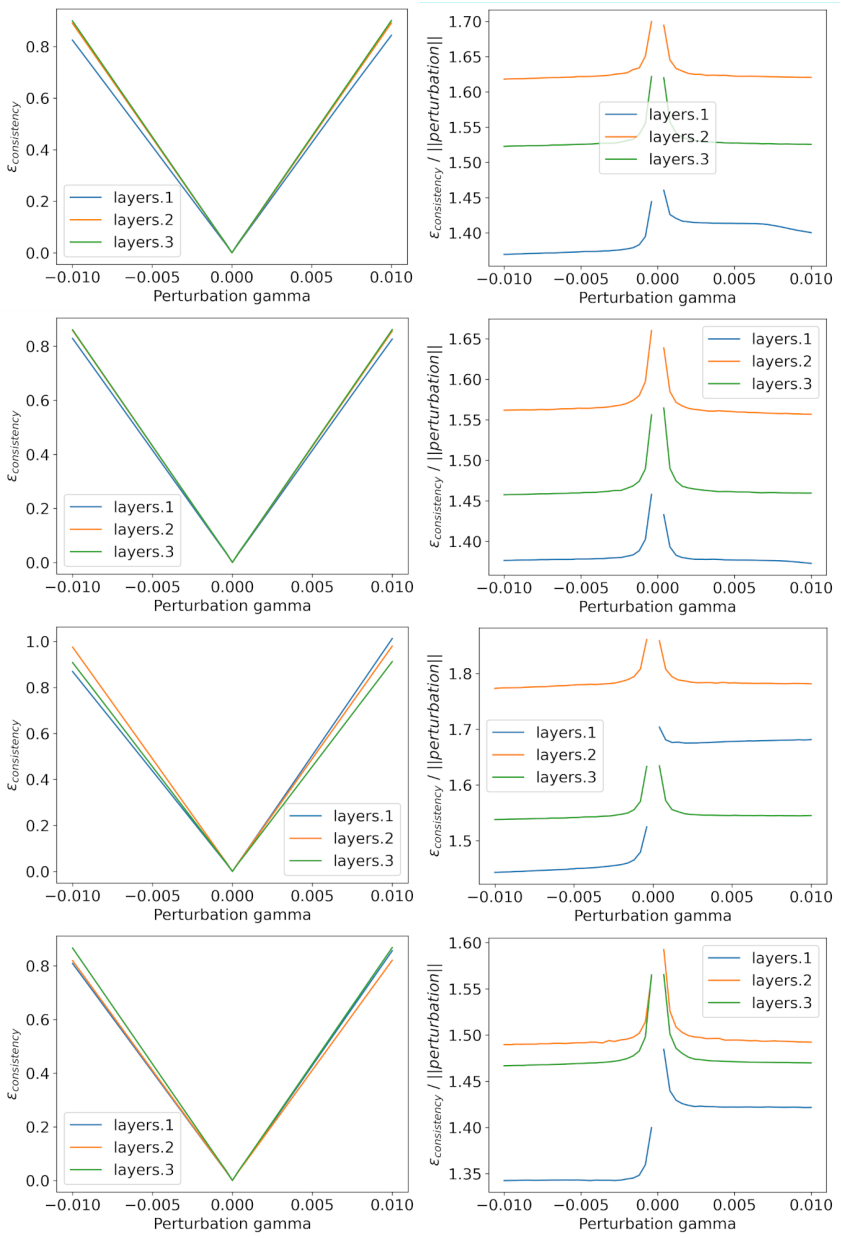


Figure A.6: The mean consistency error for (from top to bottom) `red`, `blue`, `triangle` and `striped` CAVs across layers (left) scaled by the size of the perturbation (right) for the Elements dataset.

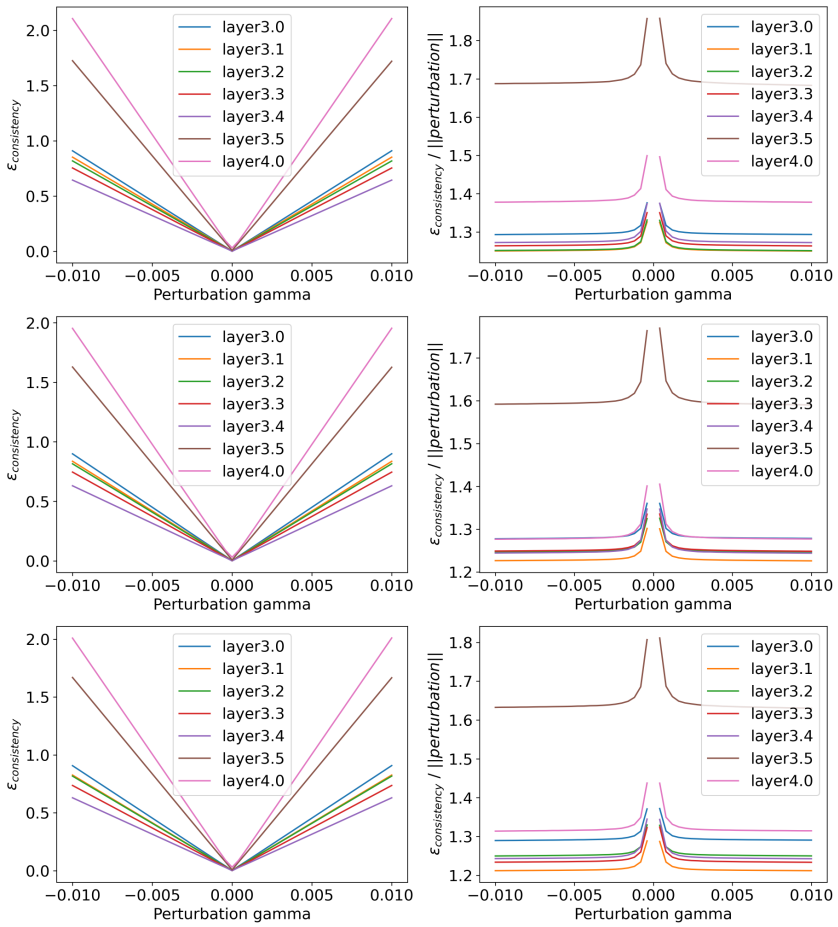


Figure A.7: The mean consistency error across 10 CAVs for `striped` (top), `lined` (middle) and `dotted` (bottom) CAVs across layers (left) scaled by the size of the perturbation (right) for a ResNet-50 train on ImageNet.

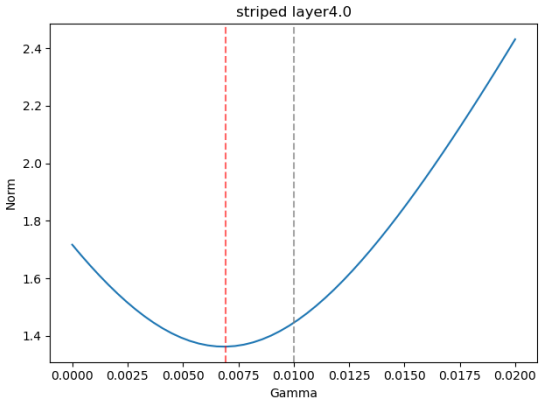


Figure A.8: The sensitivity of the mean consistency error to scaling γ for an optimised CAV for the penultimate convolutional layer in a ResNet50 trained on ImageNet.

A.4.2 Additional results

In this section we provide additional results for the different consistency experiments as in fig. 3.3. We include the results for multiple concepts and layers for both the ImageNet and Elements datasets. To allow for better comparison between layers, we normalise the consistency errors in some of the figures. For each layer, the consistency error is divided by the mean error for the optimised CAVs. For the normalised plots, a value of one can be seen as the lowest error possible for that layer. The relative ordering of layers is approximately consistent across the different types of CAV. For example, in fig. A.9, there is a downward trend between layer3.0 to layer3.5 and then an increase for layer4.0.

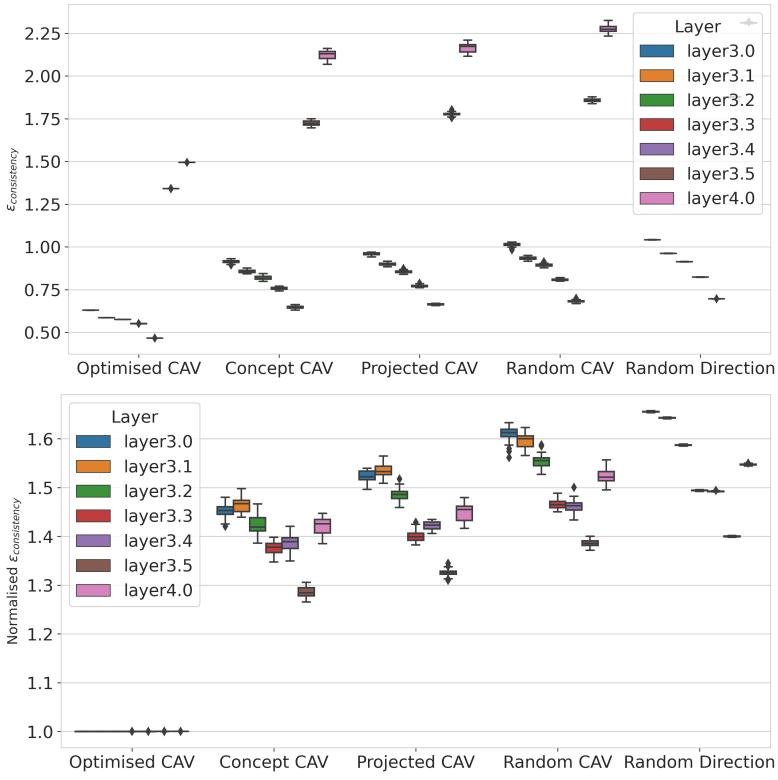


Figure A.9: The distribution of consistency errors (top) and normalised consistency errors (bottom) for different $\boldsymbol{v}_{c,l_2}$ for `striped` in a selection of layers from a ResNet-50 trained on ImageNet. Optimised CAV: The lower bound – a vector optimised to have the minimum error. Concept CAV: `striped` CAVs, trained as normal. Projected CAV: striped CAVs from layer l_1 projected into layer l_2, $f(\boldsymbol{v}_{c,l_1})$. Random CAV: CAVs with random images for the probe dataset. Random Direction: Random vectors drawn from a uniform distribution.

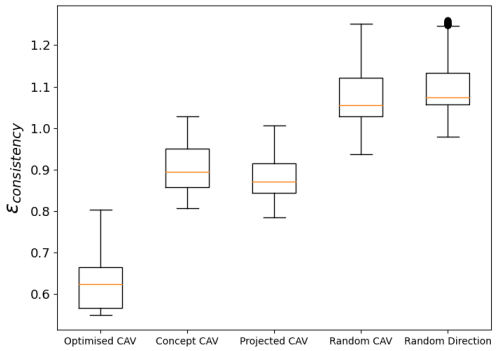


Figure A.10: The distribution of consistency errors for different $\boldsymbol{v}_{c,l_2}$ for `square`, `triangle`, `red`, `green`, `solid` and `stripes` CAVs for 'layers.1', 'layers.2' and 'layers.3' of a CNN trained on the Elements dataset. Optimised CAV: The lower bound – a vector optimised to have the minimum error. Concept CAV: CAVs, trained as normal. Projected CAV: striped CAVs from layer l_1 projected into layer l_2, $f(\boldsymbol{v}_{c,l_1})$. Random CAV: CAVs with random images for the probe dataset. Random Direction: Random vectors drawn from a uniform distribution.

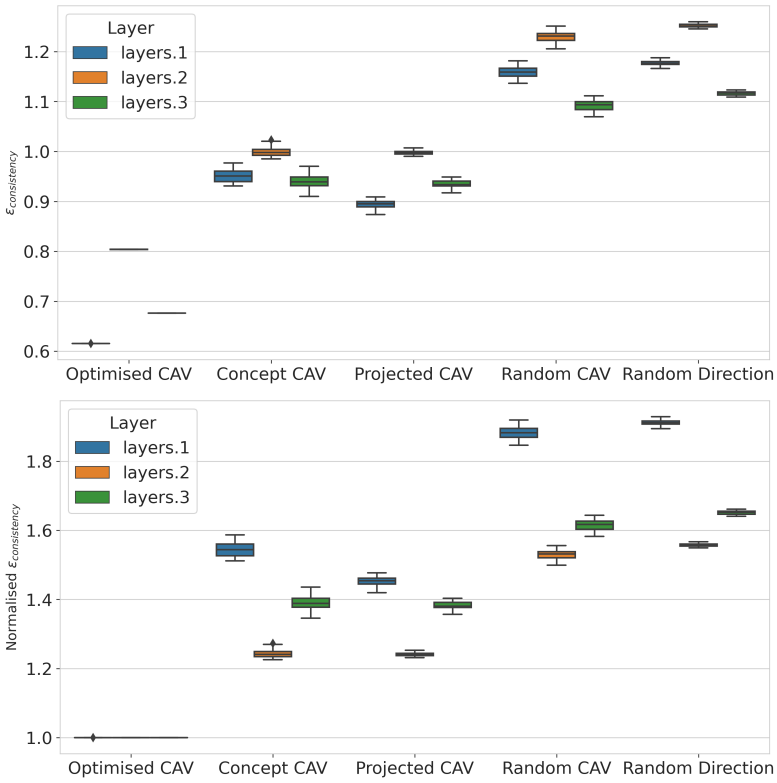


Figure A.11: The distribution of consistency errors (top) and normalised consistency errors (bottom) for different $\boldsymbol{v}_{c,l_2}$ for `square` CAVs for a variety of layers for the Elements dataset. Optimised CAV: The lower bound – a vector optimised to have the minimum error. Concept CAV: CAVs, trained as normal. Projected CAV: striped CAVs from layer l_1 projected into layer l_2, $f(\boldsymbol{v}_{c,l_1})$. Random CAV: CAVs with random images for the probe dataset. Random Direction: Random vectors drawn from a uniform distribution.

A.4.3 DeepDream

DeepDream [149] is a feature visualisation tool. It starts from an image, generated by sampling noise from a random uniform distribution and then iteratively updates the input image to maximise the L2 norm of activations of a particular layer. We use a similar approach but instead maximise the dot product between the activations and a CAV: $\boldsymbol{v}_{c,l} \cdot \boldsymbol{a}_l$. In Figures A.12 and A.13, we show these visualisations for a selection of ImageNet CAVs for successive layers in a ResNet-50.

These visualisations offer qualitative evidence that the CAVs represent different components of the same concept in different layers. For example, the `car` concept in Figure A.12 consists of many square box-like objects in earlier layers, but nothing recognisable as a car, whereas, in later layers, whole car sections can

be seen in the visualisations.

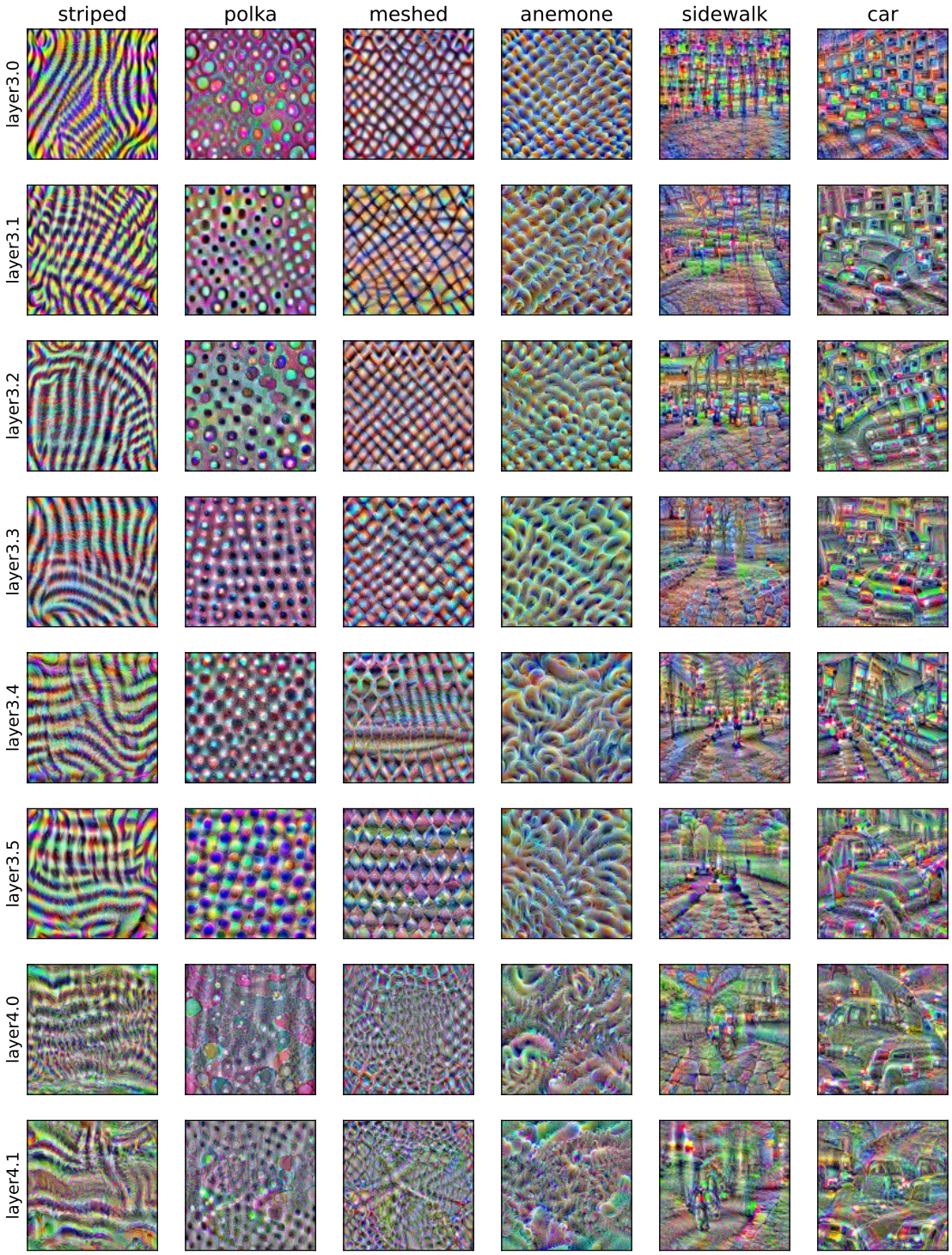


Figure A.12: CAV visualisations using DeepDream for a selection of concepts from ImageNet. Each row corresponds to a layer of a ResNet-50 and each column a different concept.

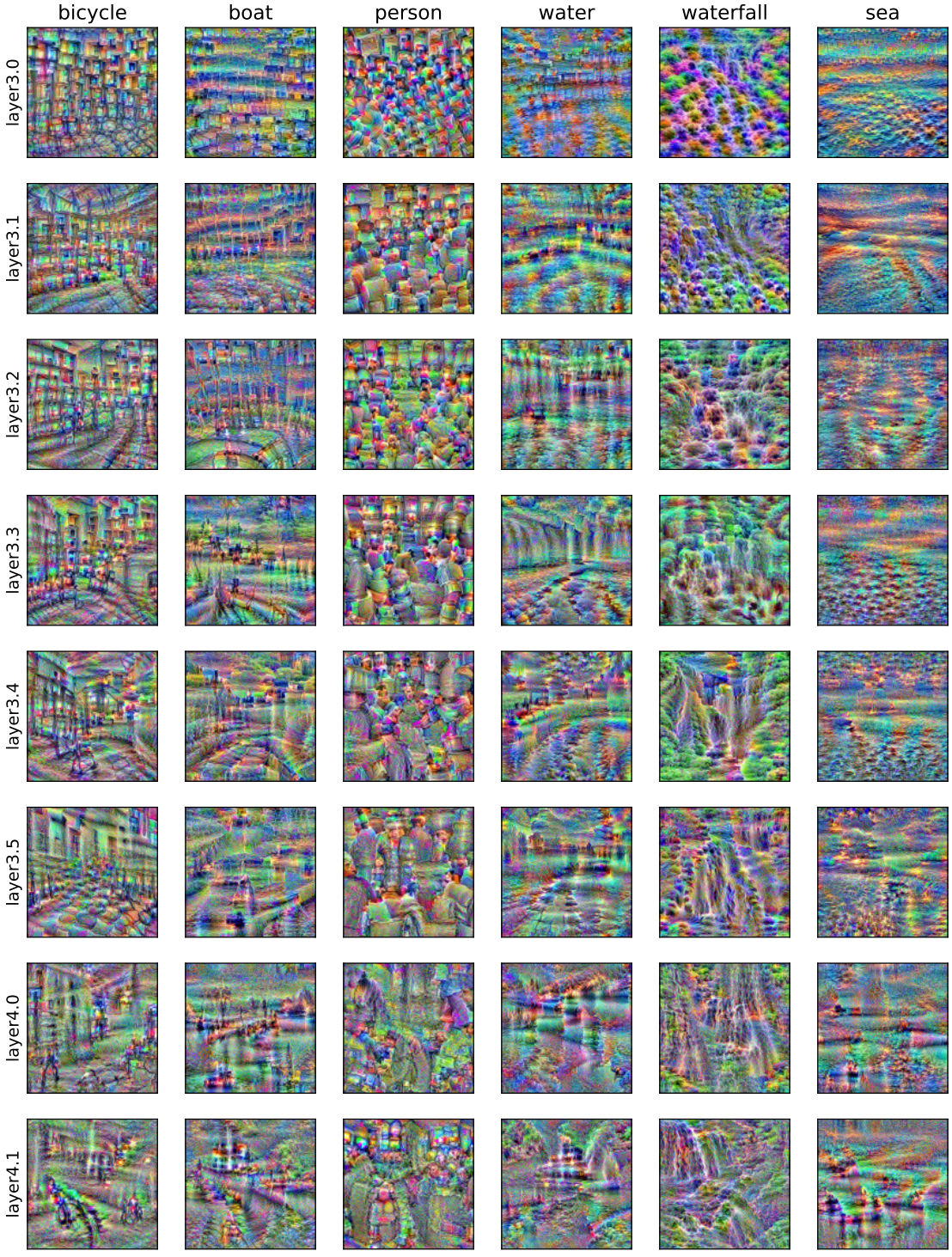
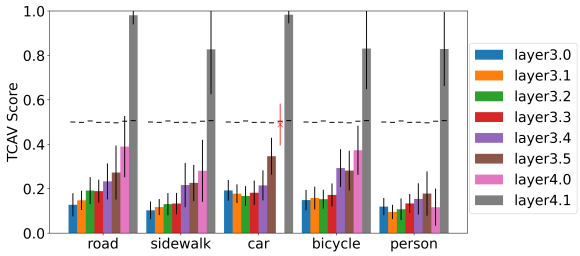


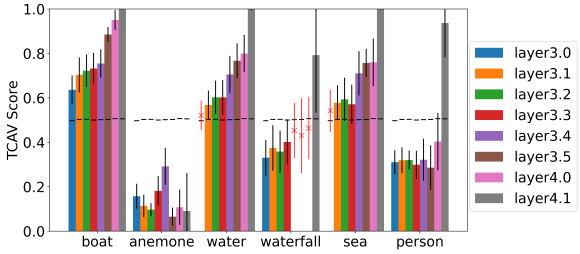
Figure A.13: CAV visualisations using DeepDream for a selection of concepts from ImageNet. Each row corresponds to a layer of a ResNet-50 and each column a different concept.

A.4.4 Inconsistent TCAV Scores

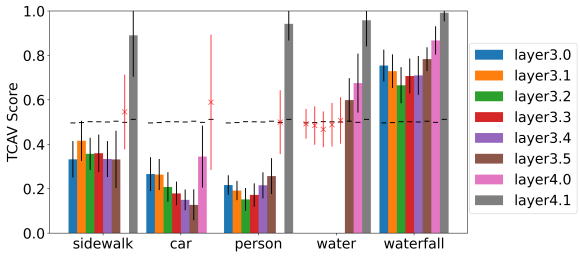
In this section, we display additional examples of inconsistent TCAV scores across layers for ImageNet classes. In Figure A.14 each subfigure contains at least one concept that has inconsistent TCAV scores. We include example images from those classes in fig. A.15.



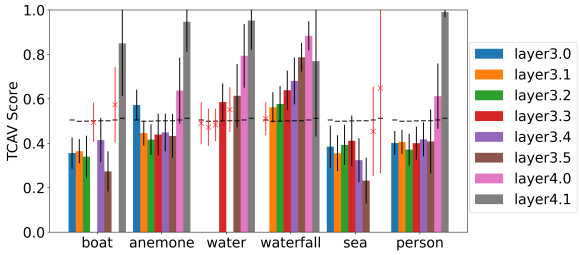
(a) TCAV scores for a selection of concepts for the 'car wheel' class in ImageNet.



(b) TCAV scores for a selection of concepts for the 'dock' class in ImageNet.



(c) TCAV scores for a selection of concepts for the 'sidewalk' class in ImageNet.



(d) TCAV scores for a selection of concepts for the 'lionfish' class in ImageNet.

Figure A.14: Inconsistent TCAV scores for a selection of concepts and classes in ImageNet. The standard deviation is shown in black for significant results and red for insignificant results. The mean TCAV score for random CAVs are shown as horizontal black lines.

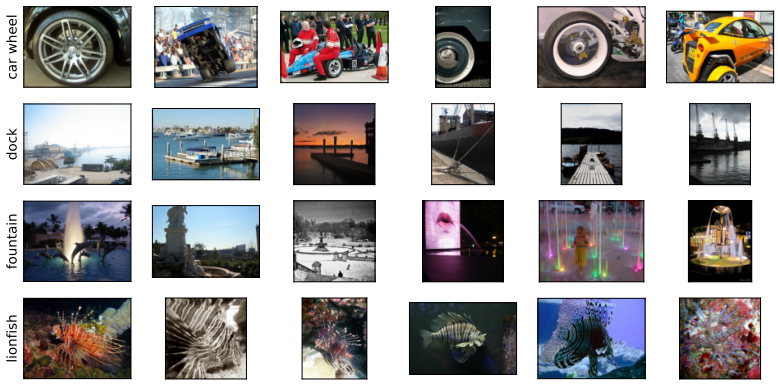


Figure A.15: Example images from a selection of ImageNet classes.

A.5 Entanglement Experiment Details

We use cosine distance to measure how similar two CAVs are. Assuming the CAVs, $v_{c_1,l}$ and $v_{c_2,l}$, are unit vectors this simplifies to the dot product of the two:

$$\begin{aligned} \text{Cosine Similarity} &= \frac{\boldsymbol{v}_{c_1,l} \cdot \boldsymbol{v}_{c_2,l}}{\|\boldsymbol{v}_{c_1,l}\| + \|\boldsymbol{v}_{c_2,l}\|} \\ &= \boldsymbol{v}_{c_1,l} \cdot \boldsymbol{v}_{c_2,l} \end{aligned} \tag{A.32}$$

In our visualisations (fig. 3.4 and appendix A.5.1) we compare multiple CAVs for each concept, each with a different random probe dataset, denoted by r. This allows us to see how similar CAVs for the same concept are on repeat training runs. Each value in the visualisation is the mean cosine similarity between the concepts on its corresponding x and y axis labels between all CAVs which do not have the same random probe dataset, i.e.:

$$\sum_{r_1}^{R} \sum_{r_2!=r_1}^{R} \frac{\boldsymbol{v}_{c_1,l}^{r_1} \cdot \boldsymbol{v}_{c_2,l}^{r_2}}{R(R-1)} \tag{A.33}$$

where $\boldsymbol{v}_{c_1,l}^{r_1}$ is the CAV corresponding to concept c_1, layer l and random probe dataset $\mathbb{X}_{c,r_1}$.

A.5.1 Additional Results

Elements

In fig. A.16, we show the cosine similarities for all concepts in the standard Elements dataset. The conclusions are similar to the visualisation for $\mathbb{E}_1$ in

fig. 3.4, but the negative associations between the mutually exclusive concepts are weaker. We hypothesise that this is because there are more concepts within each group. This is empirically justified as the average cosine similarity of each group approximately corresponds to the number of concepts in the group. If this hypothesis is true, it makes it unlikely that we will find similar groupings in real datasets containing natural images as concepts are rarely partitioned as neatly or in as few possible combinations.

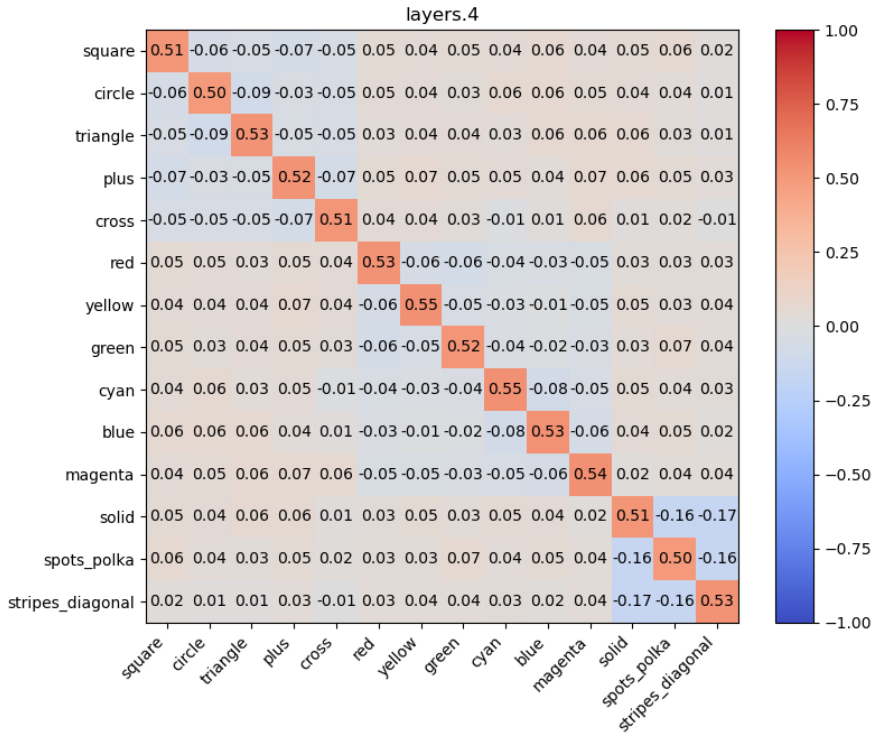


Figure A.16: Mean pairwise cosine similarities between 30 CAVs for different concepts from the standard Element dataset.

ImageNet

In Figure A.17, we show the pairwise cosine similarities for a selection of concepts for a ResNet-50 trained on ImageNet. The associations between concepts are less clear-cut than for Elements, but qualitatively they make intuitive sense. For example, the concepts most similar to `field` are `grass` and `earth`, in the top of

Figure A.17, and the concepts most similar to `sidewalk` are `bicycle`, `road` and `hedge`. The latter makes sense as many of the `hedge` exemplars in the probe dataset are next to a path or road. This emphasises the importance of designing your probe dataset to match the concept you want the CAV to represent.

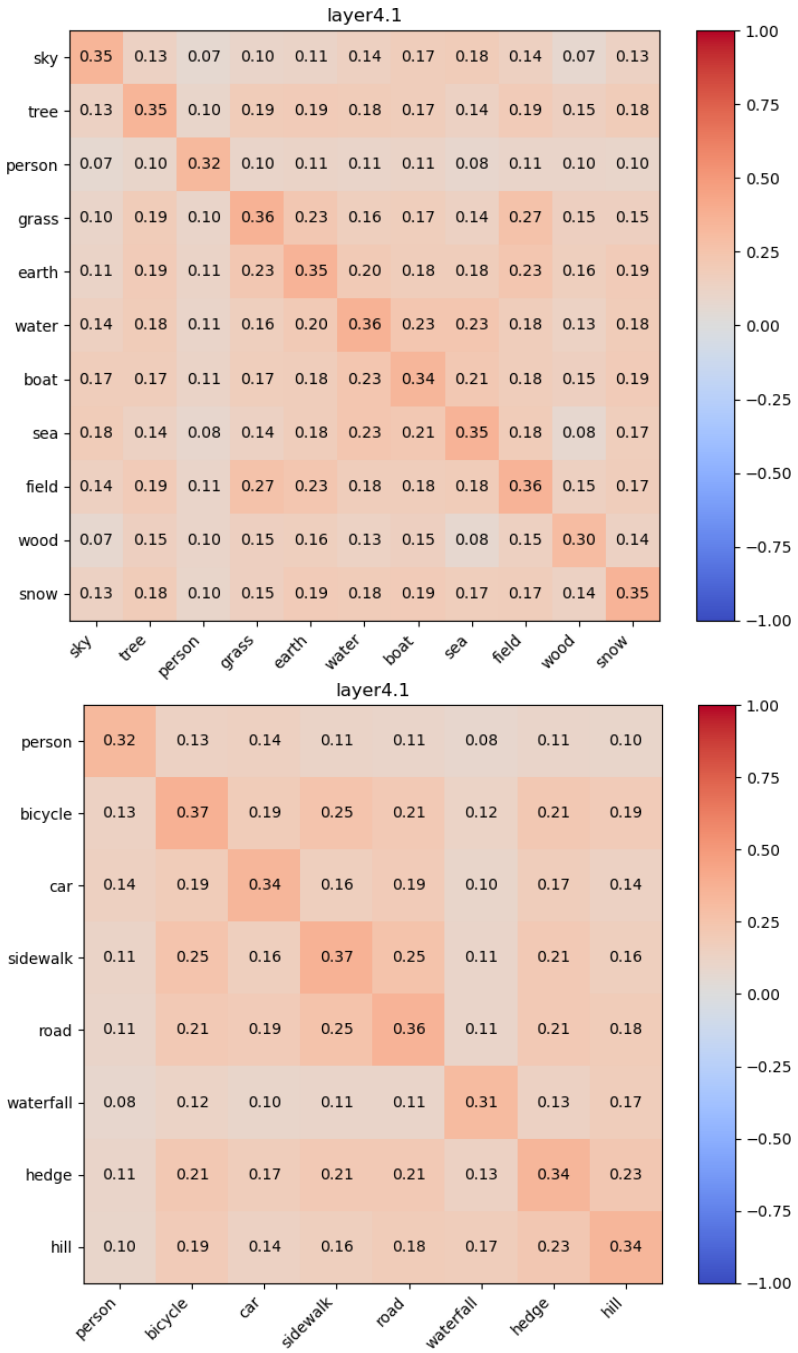


Figure A.17: Mean pairwise cosine similarities between 30 CAVs for different concepts from ImageNet.

A.5.2 Polysemanticity

In Figures 3.4 and A.16 we find that the vector representations of mutually exclusive concepts are anti-correlated with each other. Each concept vector does not just mean, for example, `red`. It means `red`, `not green` and `not blue`.

Elhage et al. [67] also find evidence for polysemantic representations. However, they found individual neurons which were polysemantic, whereas here vectors, i.e., groups of neurons, are polysemantic. In addition, the reasoning is different. The polysemanticity discussed in Elhage et al. [67] is caused by sparse features being compressed into fewer neurons than there are features. Here, we do not have sparse features and have more neurons than features. Instead, the polysemanticity is caused by associations between the concepts and the optimisation process favouring negatively correlated representations for mutually exclusive concepts.

A.5.3 Dot product distributions

The definition of entangled concepts in eq. (3.6) uses the dot products of a CAV trained on concept c_1 with the the activations of a probe dataset for a different concept c_2. Figure fig. A.18 shows the distribution of dot products for a selection of CAVs, model training datasets and concept probe datasets. We use the same set of Elements datasets as in section 3.6.2: $\mathbb{E}_1$, $\mathbb{E}_2$, $\mathbb{E}_3$, where $\mathbb{E}_1$ has no association between `red` and `triangle` and $\mathbb{E}_2$ and $\mathbb{E}_3$ have successively stronger associations between the concepts. This changing association is apparent in the dot product distributions. For $\mathbb{E}_1$, the dot products of the test probe datasets differing from the CAV concept do not differ significantly from the dot products for random images (the negative probe dataset). Whereas, for $\mathbb{E}_2$ and $\mathbb{E}_3$, the random distribution is shifted lower than for either CAV, even if the CAV is labelled differently from the test dataset. This shows that the `red` and `triangle` CAVs are entangled for these datasets/models.

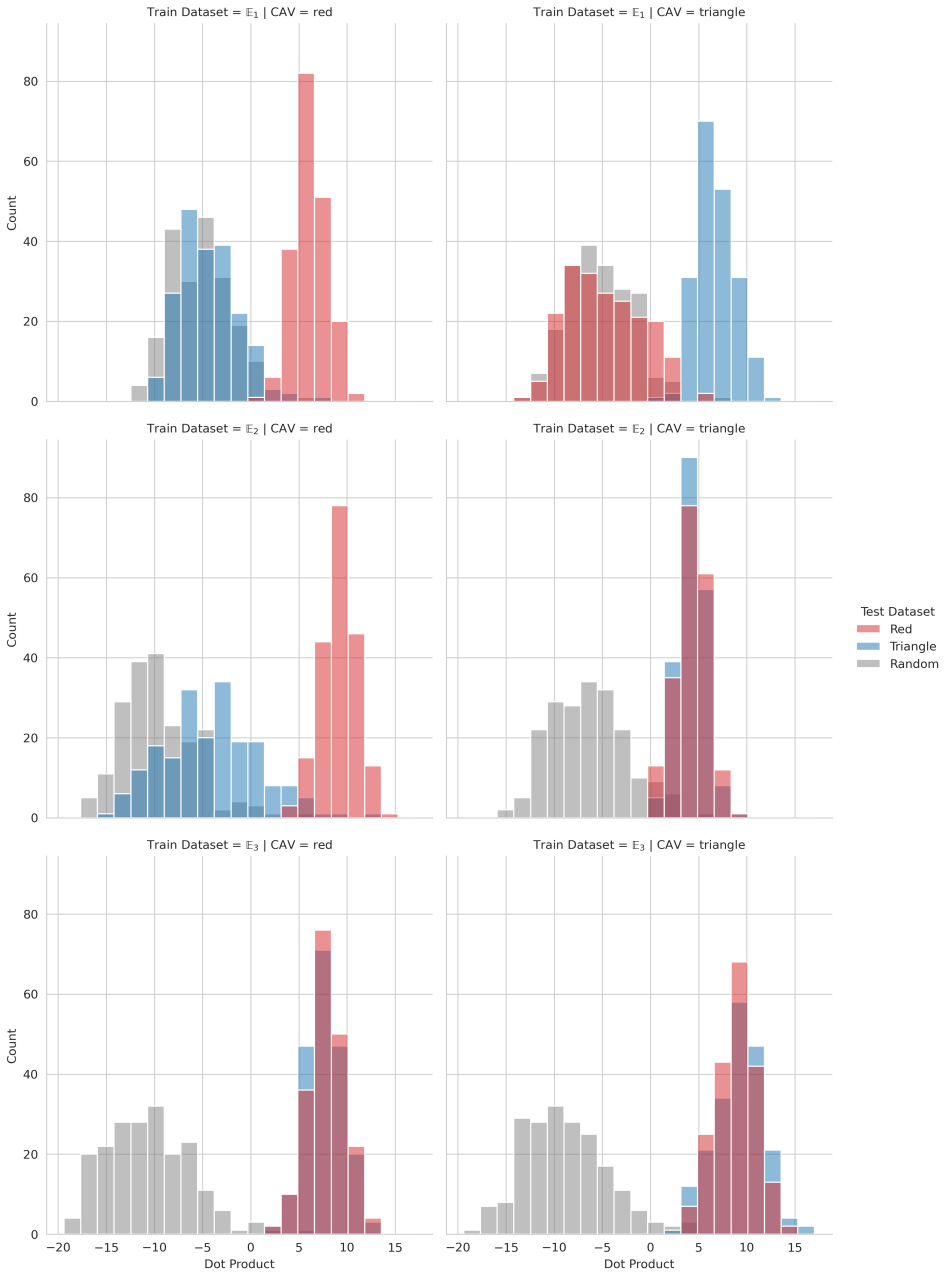
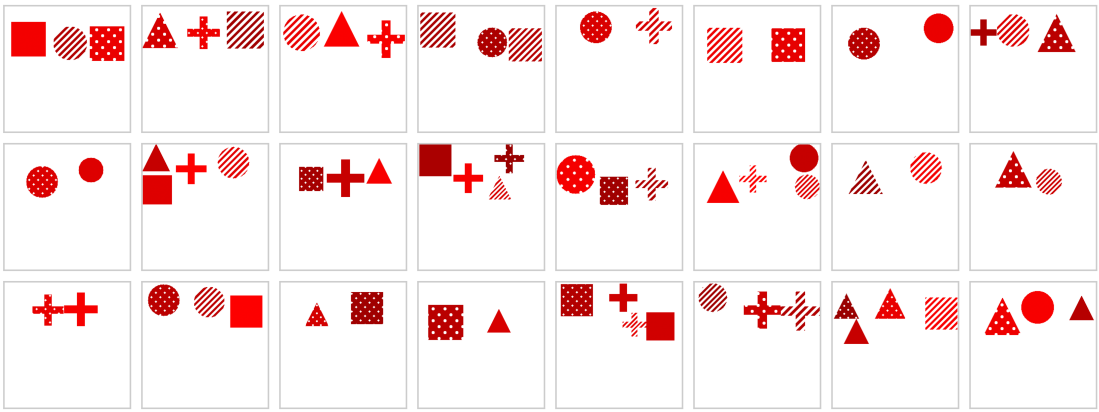


Figure A.18: Distribution of dot products ($\boldsymbol{v}_{c_1,l} \cdot \boldsymbol{a}_{c_2,l}$) for the three versions of Elements with increasing association between `red` and `triangle` ($\mathbb{E}_1$, $\mathbb{E}_2$ and $\mathbb{E}_3$). The distribution of dot products are displayed for three different test sets containing in-distribution images for `red`, `triangle` and `random` images.

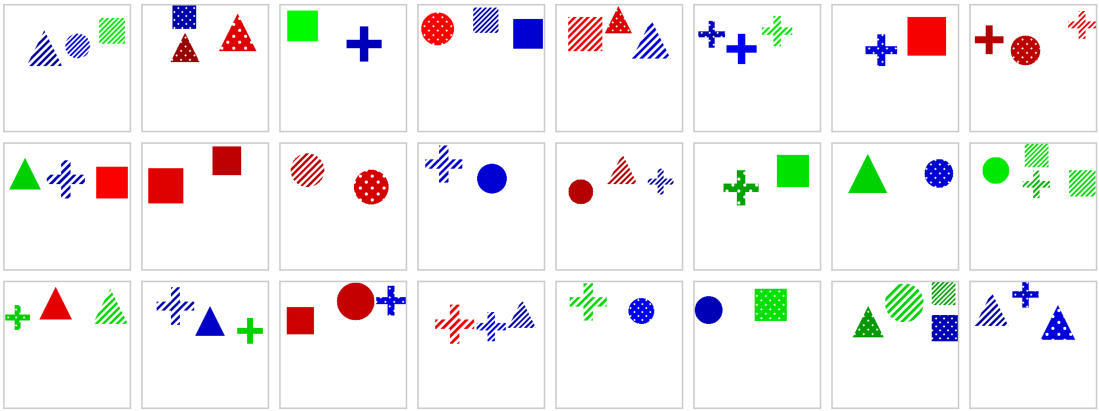
A.6 Spatial Dependency Experiment Details

A.6.1 Spatially Dependent Probe Datasets

For Elements, the probe datasets contained elements that only appear in specified locations – an example is shown in fig. 3.2. For ImageNet, we do not have direct control of where objects can appear. Therefore, we greyed out different regions of the image. For example, we created oppositely dependent concepts by either greying out the middle of the image, or the edges - see Appendix A.6.1 for examples.

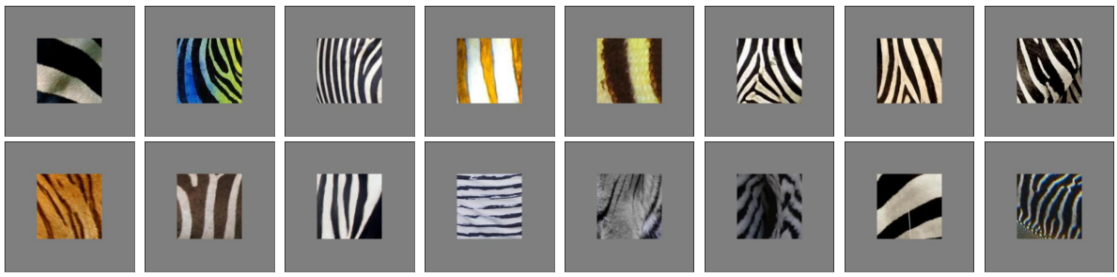


(a) The positive set.



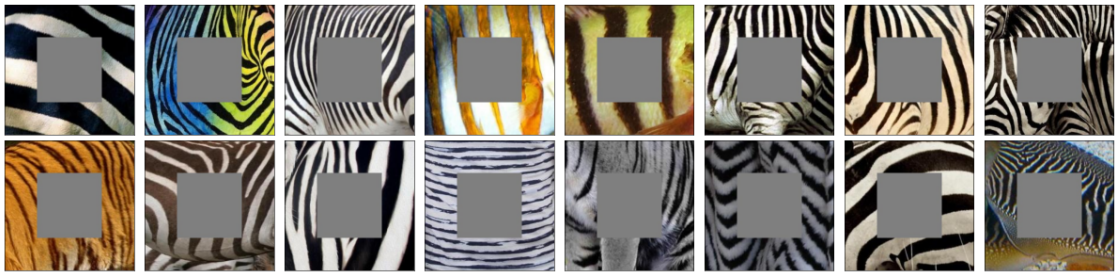
(b) The negative set.

Figure A.19: Example images for the positive and negative sets of the probe dataset for the `red top` in the simple elements dataset.

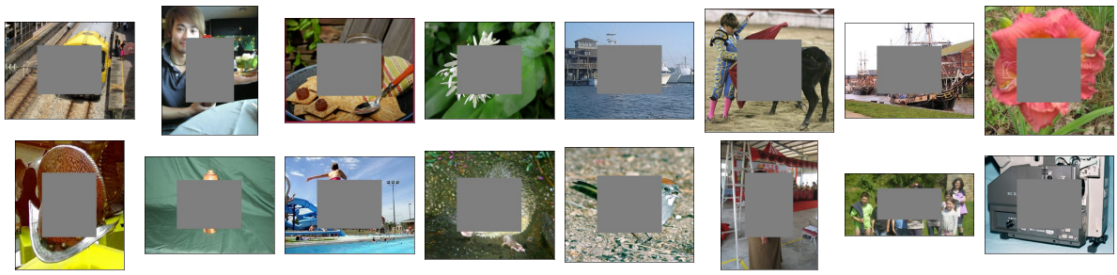


(a) The positive set for `striped middle`.

(b) The negative set for `striped middle` concept.



(c) The positive set for `striped edges` concept.



(d) The negative set for `striped edges`.

Figure A.20: Example images from spatially dependent probe datasets for ImageNet.

A.6.2 Spatial Norms Details

When finding CAVs, we use the activations of layer l in a convolutional neural network, which has shape $H \times W \times D$, where H, W and D are the height, width and number of channels, respectively. When finding the CAV, these activations are flattened to be vectors of length $m = H \times W \times D$. The value of each element

in the activations depends on its specific height, width and depth indices. As a result, each corresponding element of the resultant CAV is also index-dependent. We are interested in spatial dependence, so to visualise how a CAV varies across the width and depth dimensions we calculate the L2 norm of each depth-wise slice – the CAV's spatial norms.

A.6.3 Individual Spatial Norms

In fig. A.21, we present the spatial norms for `striped` in a ResNet trained on the ImageNet dataset. Each heatmap is for a different random probe dataset, denoted by r. The different patterns in the heatmaps show that each individual $\boldsymbol{v}_{c,l}^{r}$ has a spatial dependency which differs across r. However, when we average the norms across multiple CAVs, $\sum_{r=1}^{R} \mathbf{S}_{c,l}^{r}/R$, we obtain a uniform distribution across the spatial dimensions. This uniformity suggests that the spatial dependencies of each individual CAV cancel out across multiple seeds, as depicted in the top rows of fig. 3.6 in the main text.

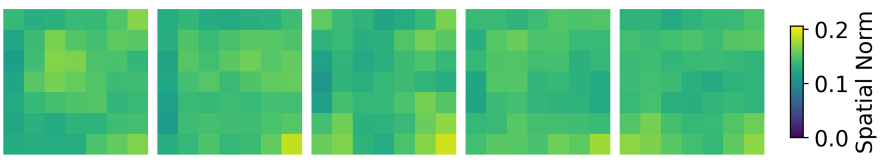


Figure A.21: Individual spatial norms for `striped`, where each CAV was trained on a different negative probe set, for layer4.1 of a ResNet trained on ImageNet.

A.6.4 Additional Spatial Norms

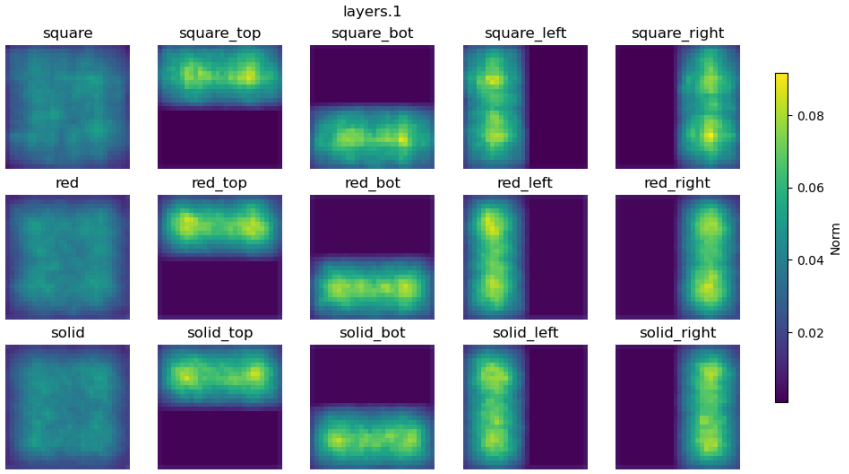


Figure A.22: Mean CAV spatial norms across 30 CAVs for a selection of concepts in the Element dataset for the second convolutional layer.

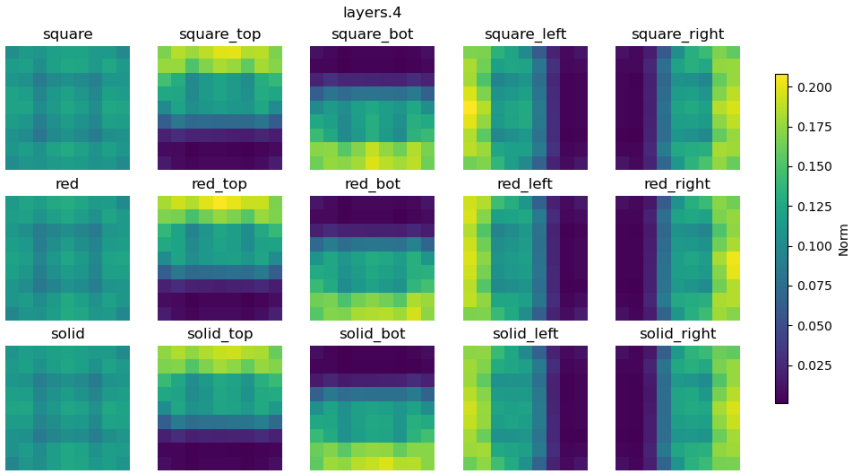


Figure A.23: Mean CAV spatial norms across 30 CAVs for a selection of concepts in the Element dataset for the fifth convolutional layer.

A.6.5 Spatial Means

Instead of visualising the norm of each depth-wise slice, we can visualise the mean. We default to showing the norm because you could have a spatial mean of zero in a region of activation space which has a large effect on the directional derivative. This could occur, for example, if half the elements of the CAV are large and positive and the corresponding gradients are positive, and half the elements of the CAV are large and negative and the corresponding gradients are also negative. This would

lead to that spatial region having a large contribution to the directional derivative, but the spatial mean would be close to zero. The spatial norm, however, would be large for this region. This makes the norm a better measure of the effect of each region, but the mean can still be useful to show the direction of that effect.

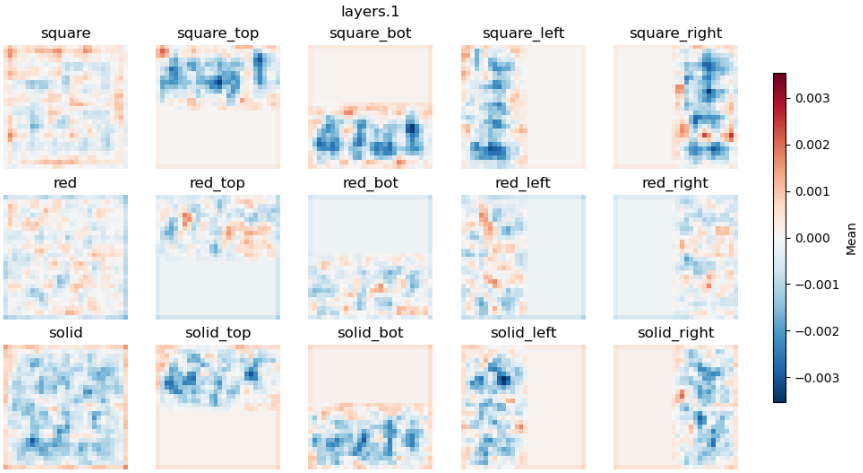


Figure A.24: Mean CAV spatial means across 30 CAVs for a selection of concepts in the Element dataset for the second convolutional layer.

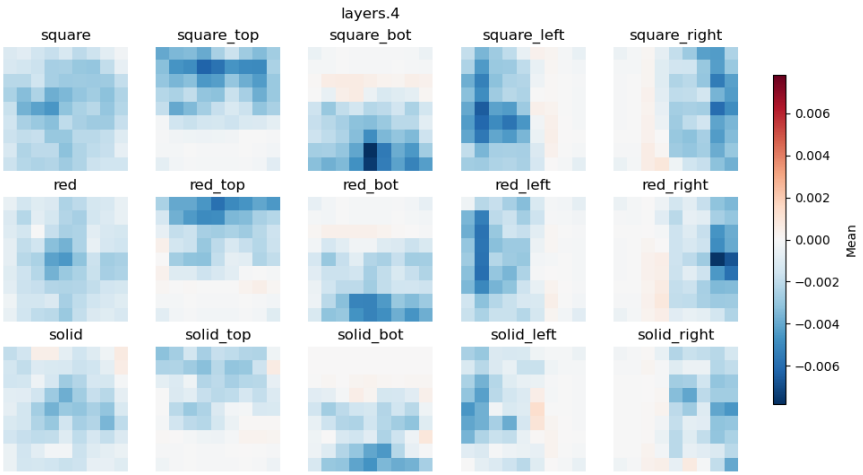
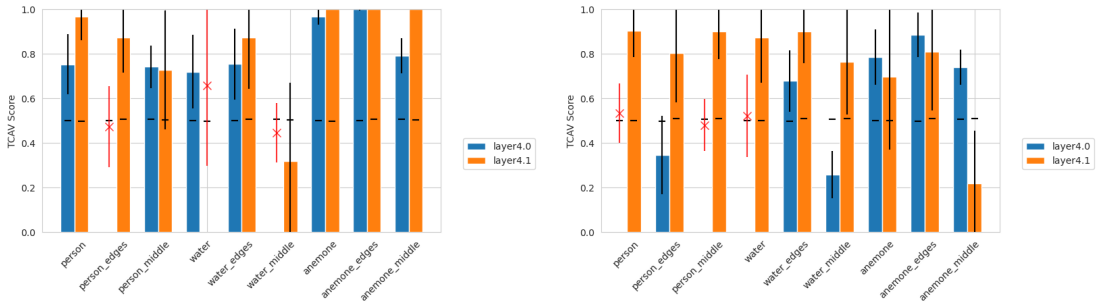


Figure A.25: Mean CAV spatial means across 30 CAVs for a selection of concepts in the Element dataset for the fifth convolutional layer.

A.6.6 Spatially Dependent TCAV Scores

In this section we provide example TCAV scores which differ across complementary spatially dependent CAVs, i.e., for at least one concept, the TCAV score is the

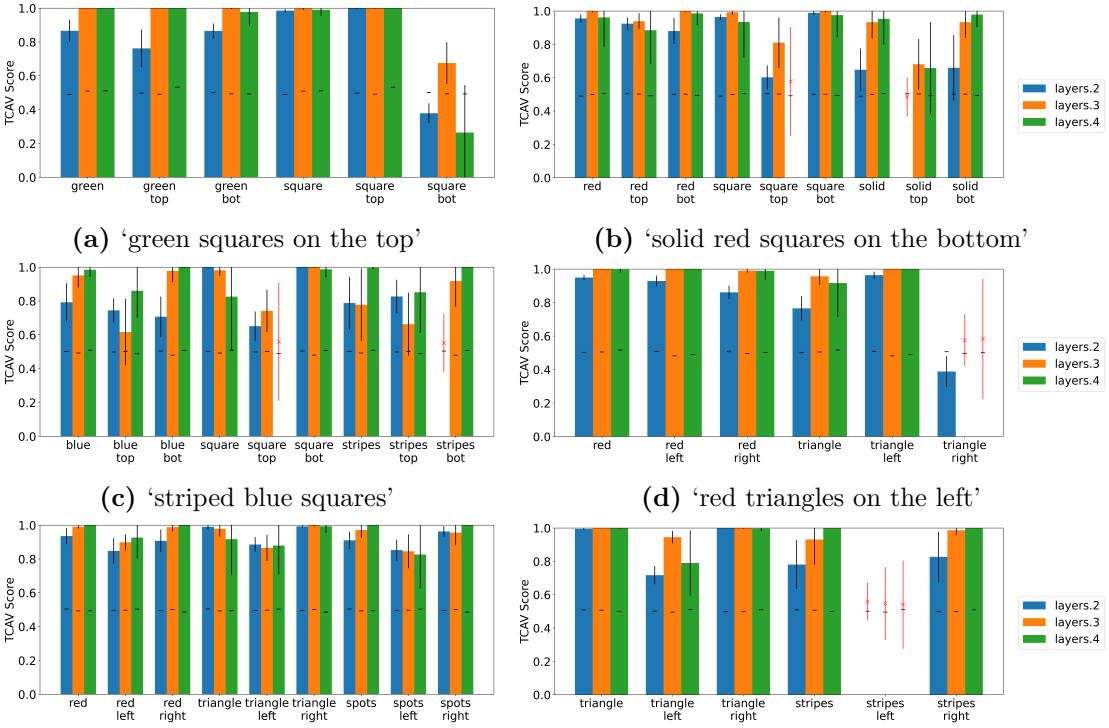
opposite side of the null for the `edges` version of a concept compared to the `middle` version (or the `left`/`right` and `top`/`bottom` versions).



(a) TCAV scores for a selection of concepts for the 'anemone fish' class in ImageNet.

(b) TCAV scores for a selection of concepts for the 'spiny lobster' class in ImageNet.

Figure A.26: Examples of spatially dependent TCAV scores in ImageNet. Each subfigure is a separate class. The standard deviation is shown in black for significant results and red for insignificant results. The mean TCAV score for random CAVs are shown as horizontal black lines.



(a) 'green squares on the top'

(b) 'solid red squares on the bottom'

(c) 'striped blue squares'

(d) 'red triangles on the left'

(e) 'spotted red triangles on the right'

(f) 'striped triangles on the right'

Figure A.27: Examples of spatially dependent TCAV scores in the spatially dependent version of Elements. Each subfigure is a separate class. The standard deviation is shown in black for significant results and red for insignificant results. The mean TCAV score for random CAVs are shown as horizontal black lines.

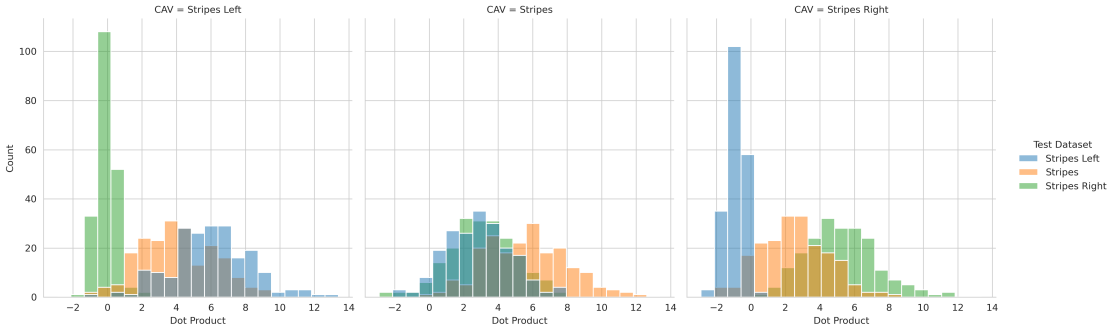


Figure A.28: Distribution of dot products between spatially dependent CAVs and image activations ($\boldsymbol{a}^{+}_{c,l,\mu} \cdot \boldsymbol{v}_{c,l}$) for the spatially dependent Elements dataset. Each column is for different CAVs. From left to right these are: `stripes left`, `stripes`, `stripes right`. For each CAV we show the distribution for three positive probe datasets: `stripes left` (blue), `stripes` (orange), `stripes right` (green).

A.6.7 Dot product distributions

The definition of concept vector spatial dependence in eq. (3.8) compares a CAV, $\boldsymbol{v}_{c,l}$, with the activations of two positive probe datasets with different spatial dependencies, $\boldsymbol{a}^{+}_{c,l,\mu_1}$ and $\boldsymbol{a}^{+}_{c,l,\mu_2}$, by taking the dot product between them $\boldsymbol{a}^{+}_{c,l,\mu_x} \cdot \boldsymbol{v}_{c,l}$. In fig. A.28, we show the distribution of dot products for three concepts and three test probe datasets in the spatially dependent version of Elements. The separation between the distributions for the `stripes left` and `stripes right` probe datasets (blue and green bars, respectively) for both the `stripes left` and `stripes right` CAVs (left and right plots, respectively) demonstrate that these CAVs are spatially dependent.

A.7 Further Related Work

Recent work highlighted problems with concept-based explanation methods. Ramaswamy et al. [178] showed that using different probe datasets to interpret the same model can lead to different explanations for the same concept. Similarly, Soni et al. [205] showed these methods to be sensitive to the random seed used to sample images for the negative set. Our work complements this research by investigating the underlying properties of concept vectors and how they may cause problems when interpreting concept-based explanations.

Table A.3: Examples in computer vision and medical imaging research, where consistency, entanglement and spatial dependence may impact analyses. We use the following abbreviations: skin cancer (SC), skin lesions (SL), breast cancer (BC), histology (H) CIFAR-10/100 (CF) [122], COCO (CO) [132], CUB (CB) [221], Places365 (Pl) [243], Waterbirds (WB) [189], ImageNet (Im) [56]

Property	Medicine				CV Research						Papers
	SC	SL	BC	H	CF	CO	CB	Pl	Wb	Im	
Consistency	✓	✓			✓	✓	✓	✓			[229, 178, 73, 234, 79, 138]
Entanglement	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	[229, 178, 73, 234, 79, 84, 138, 172]
Spatial Dependence	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	[229, 178, 73, 234, 79, 138, 172]

Extensions to the original TCAV have been suggested, attempting to improve aspects of the original method. For instance, Ghorbani et al. [78] automate concept discovery by using super-pixels and clustering, removing the need to handcraft a probe dataset. Zhang et al. [239] and Schrouff et al. [190] change how CAVs are created to produce local and global explanations. However, these methods still use vectors to represent concepts [7]. As such, our work is still applicable to each of the extensions.

The properties analysed in this paper are generally applicable. To give insight into when the various properties may be relevant, we performed a review of papers which use CAVs in medical imaging and computer vision research. In each case, we checked if the authors had done any checks related to layer consistency, entanglement or spatial dependence of CAVs. Almost no papers evaluated the effect of these properties on their results, and when they did it was by creating CAVs in multiple layers, providing some robustness to layer inconsistency. In Table A.3, we provide a list of these papers, detailing any use-cases where our recommendations could have helped check the impact of CAV properties on results.

We think in generalities, but we live in detail.

— Alfred North Whitehead

B

Additional Details for the AGE Study

Contents

B.1 Model Training and Development

B.1.1 ProtoPNet

A ProtoPNet[45] model consists of a convolutional network, f, a prototype layer, g_p, and a fully-connected layer, h. In our experiments the convolutional network is a ResNet-18 [93] pretrained on ImageNet [56] followed by two 1×1 convolutional layers to reduce the number of output channels to 128. The model makes a prediction by passing some input image, x, through the convolutional feature extractor to obtain a set of feature maps, $f(x)$, with shape $H \times W \times D$. The network learns m prototypes, $P = \{p\}_{j=1}^{m}$, which are tensors of shape $H_1 \times W_1 \times D$, where $H_1 < H$ and $W_1 < W$. In our experiments the feature maps are of shape $7 \times 7 \times 128$ and the prototypes $1 \times 1 \times 128$, so each prototype is a representation of some prototypical sub-patch of the

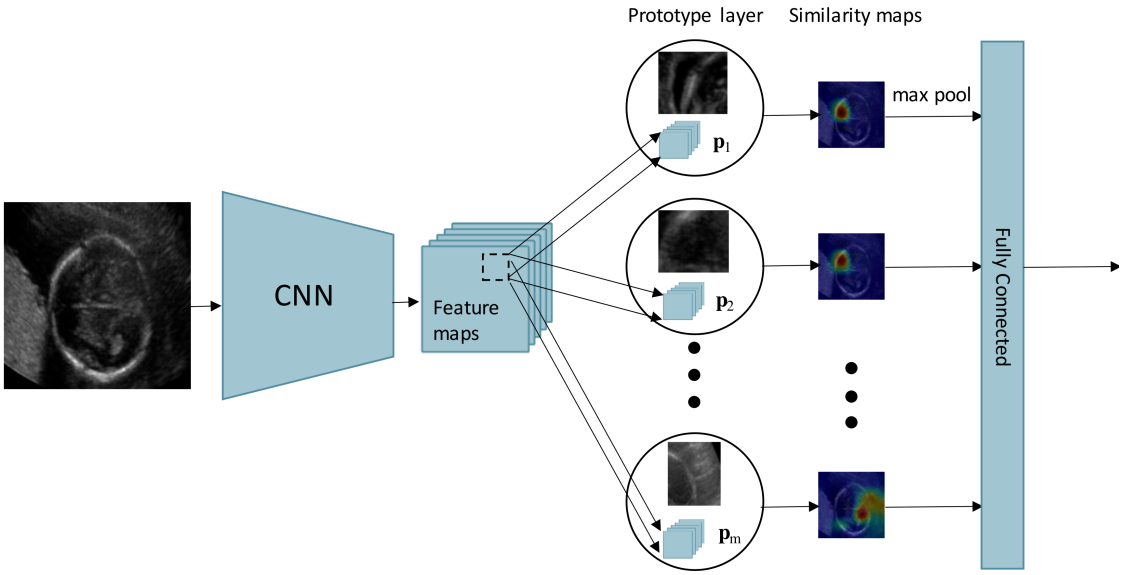


Figure B.1: Prototypical part network (ProtoPNet) architecture. A test image is passed through the convolutional neural network (CNN) backbone to obtain a set of feature maps. These feature maps are compared to the features of training images the model has seen previously, i.e. the prototypes, and similarity maps obtained. After a max pooling operation, the values are passed to a fully connected layer to classify the image. This means the predictions are made solely based on the similarities between the test image and the prototypes, making an interpretable-by-design model.

image $\frac{1}{49}$th of its size. The prototype layer then calculates the $L2$ distance between each prototype and all 49 patches of the feature map. These distances are inverted to obtain a set of similarity scores and the maximum score for each prototype is then passed through the fully-connected layer. The maximal similarity scores can be seen as the likelihood that each prototype is present in the image and the weights in the fully-connected layer the importance of each prototype for each class.

The model is globally interpretable because the prototypes and weights are fully accessible. Its local explanations consist of the prototypes most similar to a test image and their corresponding contributions to the model output. For an example of how the explanations can be displayed, see Figure 5.1.

The training protocol has three steps: (1) the prototypes and convolutional layers are jointly optimised using stochastic gradient decent (SGD); (2) the prototypes are pushed to the activations of the nearest (measured in feature space) image sub-patch of the train dataset - this provides the model its interpretability, as the prototypes can now be represented in image space by that sub-patch; (3) the rest of

the network is frozen and the fully-connected layer is optimised using SGD. These three steps are repeated multiple times until the model converges.

The loss optimised during step (1) is composed of two parts: cross entropy and cluster loss. Cluster loss encourages the model to have at least one training sub-patch with activations close to each prototype. This minimises the changes made to the prototypes in step (2) of training. In step (3) the loss for h is cross entropy and an $L1$ penalty to regularise the weights. For a detailed explanation of the losses, architecture and training protocol see the ProtoPNet paper [45].

In our work, compared to the ProtoPNet paper[45], the restriction that each prototype be relevant to a single class was removed. This is because the task of GA estimation is a regression task converted into a classification by binning, making the classes more similar to each other than in a standard classification task. Thus, it could be desirable for a prototype to be similar to multiple classes. We achieve this by removing three different components of ProtoPNet training:

1. The separation loss – a loss encouraging each latent patch of a training image to be different from prototypes not of its own class

2. The restriction that the $L1$ regularisation only applies to prototypes of a *different* class to the training image and that cluster loss is only applied to prototypes of the *same* class as the training image

3. The class dependent initialisation of the final layer where negative weights are set between prototypes and logits of different classes

ProtoPNet[45] masked the $L1$ penalty to reduce the level of negative reasoning present in the model, the argument being it is easier to interpret the model if solely positive reasoning is used. Therefore, to increase the relative levels of positive reasoning in our model, we initialised the final layer to a uniform distribution between 0 and 1.

B.1.2 Pruning

We prune the final layer model weights by simply setting each weight below some threshold, τ, to zero. If all connections to a prototype are zero, the prototype is removed from the network. The fully-connected layer is then trained for 15 epochs, keeping any weights at zero fixed to allow the model to adapt to the changes, while keeping the same level of sparsity.

To measure the sparsity (simplicity) of the model, we define the number of relevant prototypes, r, as:

$$r = \frac{1}{K} \sum_{k=1}^{K} |\boldsymbol{\mathcal{P}}_k| \tag{B.1}$$

where $k \in \{1, \ldots, K\}$ is the class index. This is equivalent to the number of non-zero weights in the fully-connected layer divided by the number of classes. r gives an intuitive measure of the size of the model's global explanations as it is the average number of prototypes which affect each logit output. We also define r^+ and r^-, which are the subset of prototypes which have positive or negative weight connections, respectively. Similarly, we report the $L1$ penalty on the subset of weights that are positive, $L1^+$, and negative, $L1^-$. These metrics allow us to track the relative level of positive/negative reasoning in the model.

Figure B.2 shows how the model performance and sparsity changes as τ increases. As the pruning threshold increases from 0 to 0.20, the MAE of the model does not change while the sparsity increases substantially, with $L1$ decreasing from 80.1 to 40.0 and r decreasing from 65.0 to 7.5. As τ increases beyond this, the MAE begins to increase, while the sparsity continues to increase. In order to retain good performance of the model while ensuring sparse explanations, we use a threshold of 0.25.

The explanations would be too complex if all prototypes are shown to the study participants each time. So, we must decide what subset to display. The obvious solution is to display the most salient prototypes, i.e. the prototypes which have the largest contribution to the predicted class logit output. But how many prototypes should we display? We display four prototypes as a reasonable

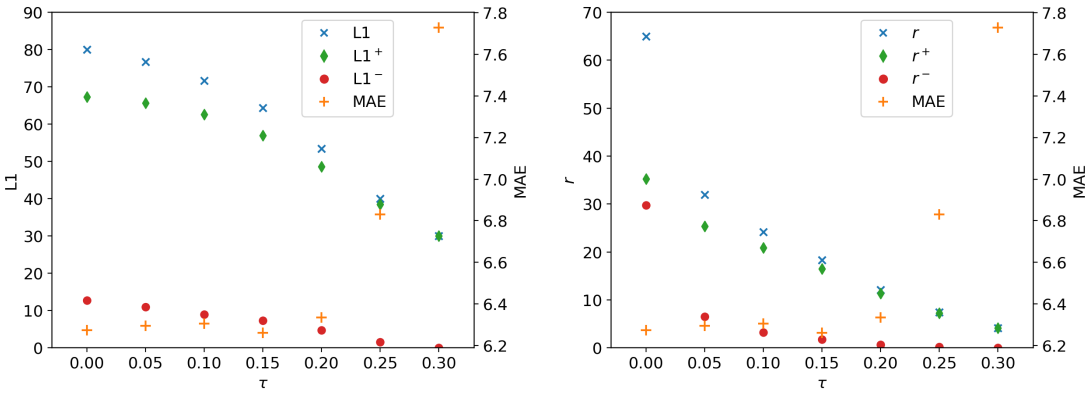


Figure B.2: Model results for INTERGROWTH-21st as the pruning threshold, τ, is increased. The $L1$ penalty for the final fully connected layer (left) and mean number of relevant prototypes, r, (right) for all (blue cross), positive (green diamond) and negative (red circle) weights against pruning weight threshold, τ. The model MAEs (orange plus) are shown on the right axis of each plot.

balance between providing a faithful explanation of the model but not overloading the participants with information. Figure B.3 shows the mean proportion of the predicted class logit (explanation completeness) that is explained by displaying the top-N number of prototypes for three different levels of pruning: none, $\tau = 0.20$ and $\tau = 0.25$. As expected, we can see that as the number of prototypes increases, a larger proportion of the model output is explained and as the level of pruning increases, the increase in explanation completeness occurs faster. Our decision to use a model with a $\tau = 0.25$ is in part because by displaying four prototypes to the participants this model explains on average 78.5% of the model output compared to just 43.1% for the unpruned model.

Figure B.4 shows the mean contribution of each prototype to each class logit output for both the pruned ($\tau = 0.25$) and unpruned model. Each colour represents the contribution of each prototype. The fact that a single colour is often present for multiple adjacent classes indicates that our hypothesis that prototypes could be useful for a range of GA, as opposed to a single class, is correct. The figure reveals the simplicity of the pruned model compared to the unpruned model as each class often has only a handful of colours present, and therefore only a handful of prototypes which contribute to that classes output, as opposed to the unpruned

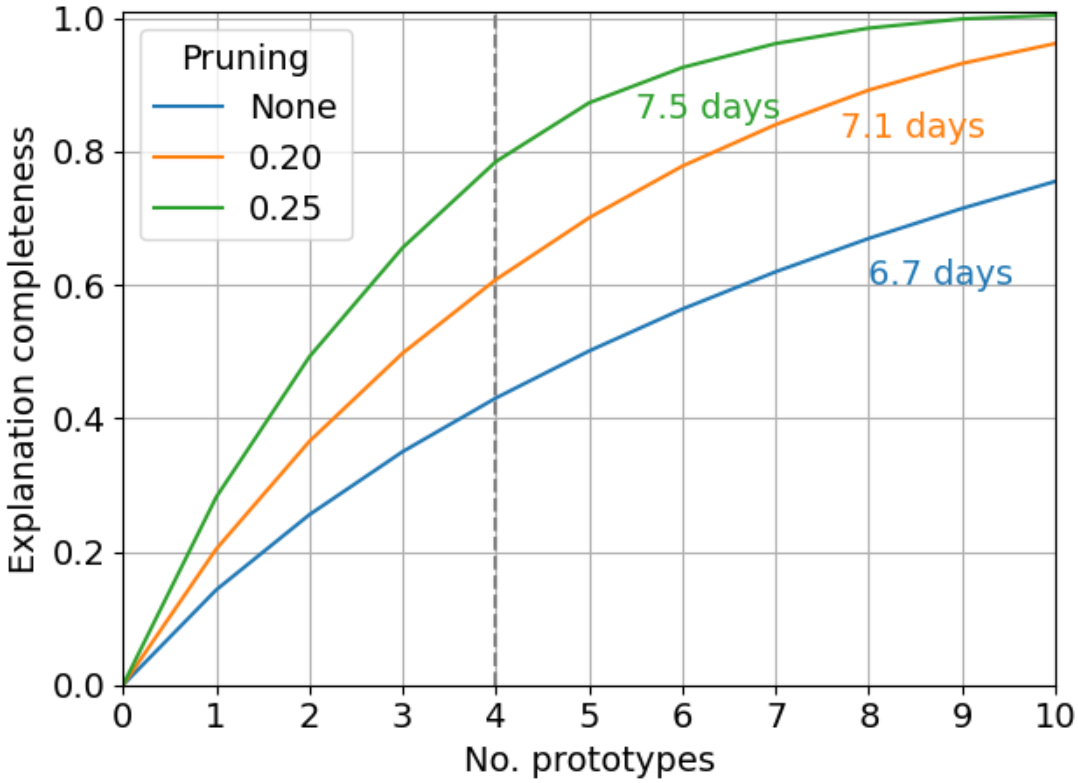


Figure B.3: The mean proportion of a model's reasoning explained (explanation completeness) by the number of prototypes that are shown for a model with no (blue), moderate (orange) and high (green) pruning. The MAE for INTERGROWTH-21st for each model is shown next to each curve. As the models are pruned more, a greater proportion of the model is explained with fewer prototypes, but the model MAE increases.

model where each class depends on many prototypes. The lack of negative reasoning in the unpruned model is also apparent as only a single prototype appears below the x axis, meaning only that prototype has a negative contribution to model outputs.

B.2 Dataset Characteristics

Fetal head images from the INTERGROWTH-21st dataset [165, 167] were used to train the XAI model. The dataset is from a healthy cohort of women from 8 different countries (Brazil, China, India, Italy, Kenya, Oman, UK, USA) who had known gestational ages via agreement between biometry measurements at first scan and last known menstrual period [219]. Each scan was done on the same model of scanner and using the same protocol. Multiple images per participant were

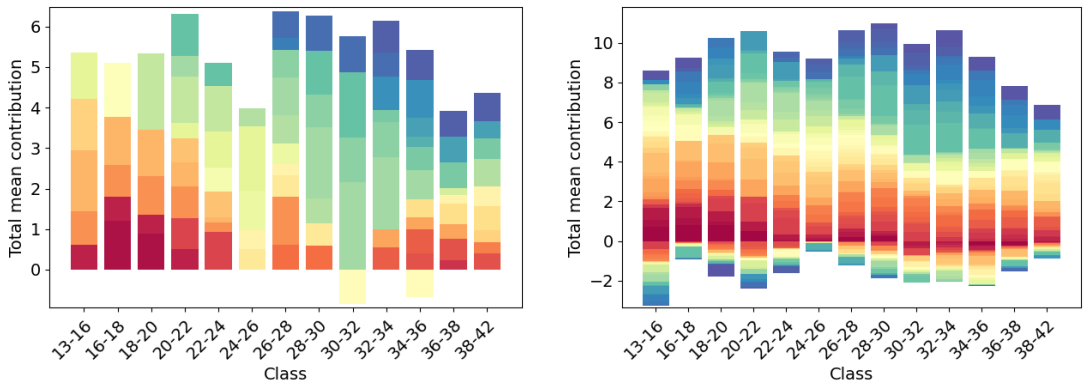


Figure B.4: The mean contribution of each prototype to the logit output of each class for the pruned model used in the AGE study (left) and original unpruned model (right). Each colour represents a unique prototype's contribution, with negative contributions starting from from zero downwards and positive from zero upwards. The same colour used across a range of classes shows that prototypes tend to be used across a range of GA.

obtained, with $4-5$ visits at different gestational ages and a mean of 6 images per visit. Repeat images at a single time-point often differ only slightly in appearance. There are $106,505$ images from 3733 women. Each image is of the transthalamic plane. The dataset was randomly split patient-wise in a 80:10:10 ratio for the train, validation and test splits giving $85,033$, $10,803$, and $10,669$ images, respectively.

The INTERBIO-21st dataset [114] is similar to the INTERGROWTH-21st dataset but with a different cohort of women and from a different set of countries (Brazil, Kenya, Pakistan, South Africa, Thailand, UK). The two cohorts have very different risk profiles due to their inclusion criteria. INTERGROTH-21st requires healthy women as participants whereas INTERBIO-21st includes some women in resource-poor settings at high risk for intrauterine growth restriction/small for gestational age and preterm delivery because of malnutrition and/or infection (HIV and malaria). This difference in cohorts is ideal for testing model performance in a different setting and proves a valuable validation set for the model.

We used only 65 images from INTERBIO-21st for the study. The 65 images were selected by manually examining 20 images from each age bin (a total of 260 images) and excluding images which did not show the complete head or were excessively zoomed out. We also excluded all images from Pakistan, since text containing information about the images obscured parts of the fetus. From the remaining

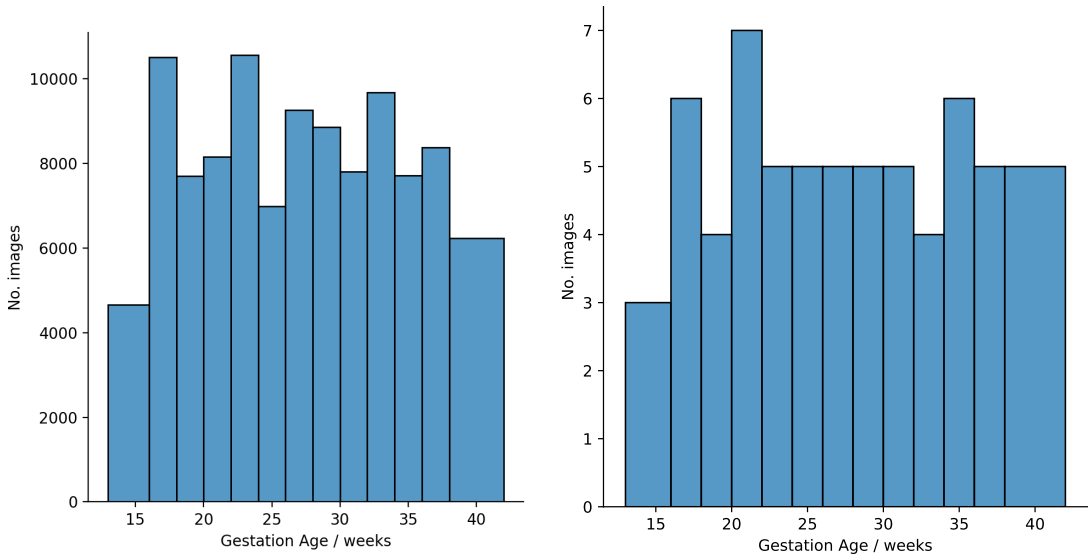


Figure B.5: GA distribution for INTERGROWTH-21st (left) and for the 65 images used in the study from INTERBIO-21st (right) binned by the classes used in the XAI model.

images, we chose the 65 images which minimised the entropy of the age distribution, i.e., the images which gave an age distribution closest to uniform across 13-42 weeks.

In Figure B.5 the gestational age distribution for our subset of the INTERGROWTH-21st dataset is shown, binned into the same classes as used for model development, along with the distribution for the INTERBIO-21st images used in the study. In both cases, the distribution is approximately uniform across 13 to 42 weeks with slightly less images at the extremes. The outermost classes were increased in size (13-16 and 38-42) to help with class imbalance.

B.3 Additional Figures

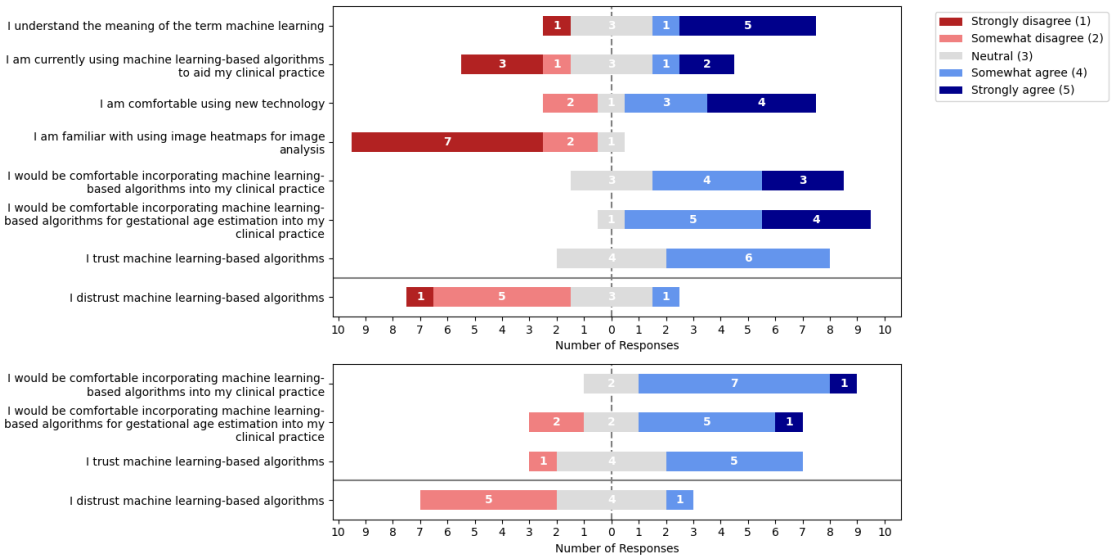


Figure B.6: Responses to "On a scale of 1-5, how much do you agree with the following statements?" at the beginning (top) and end (bottom) of the study on a Likert scale.

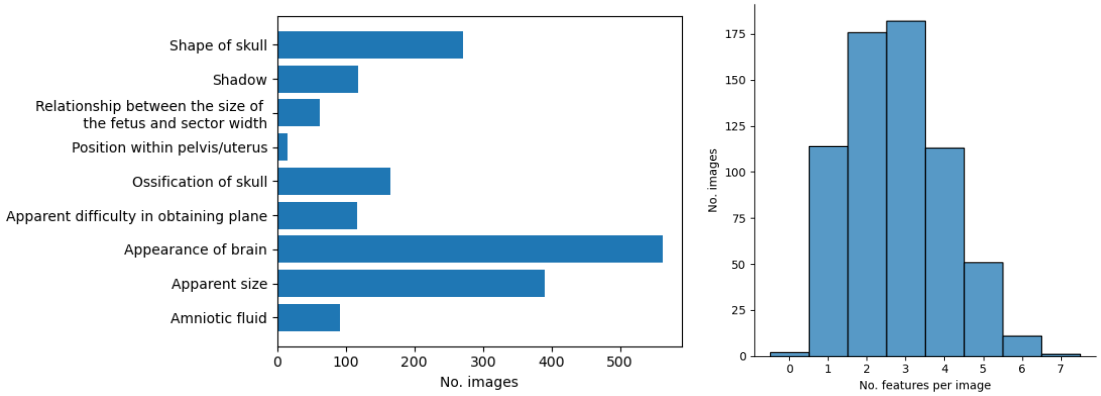


Figure B.7: Participants used a variety of image features to estimate GA. Left: The number of images for which participants found each feature to be useful for GA estimation in Stage 1 (left). Right: The distribution of the number of features participants selected per image.

Educating the mind without educating the heart is no education at all.

— Aristotle

C Ethics Impact Analysis

Gestational age (GA) prediction is vital for obstetric care, with key decisions relying on an accurate measurement. However, current methods involving ultrasound measurements are known to be unreliable later in pregnancy with estimation errors of ±3 weeks in the third trimester [166]. This is an important issue in low- and middle-income countries where women can often arrive for their first hospital visit late into pregnancy. For example, a study in Uganda found that only 29% of women had their first hospital visit during the first trimester and the median GA of their first visit was 4.7 months ($\sim$ 20 weeks) [216]. Our work in Chapter 5 helps push forward recent developments in using machine learning to improve the estimation of GA [128], hopefully leading to improved care.

Machine learning research is often performed on data from small single site studies. This can lead to models whose performance is strongly dictated by race, as the datasets are often from affluent white countries. For example, Buolamwini and Gebru [33] demonstrated that three commercially available gender classification algorithms had a higher error rate for black females than other groups. The INTERGROWTH-21st dataset that we used in our work is an ultrasound dataset collected from 8 countries: Brazil, Italy, UK, USA, China, India and Kenya [165, 167]. Although our ethics approval prevents any analysis comparing the results from different countries, by using a dataset from a broader population,

we can be more confident our results generalise to a larger group of people and that there is less racial bias.

When working on research that aims to be applied to low- and middle-income countries there is a distinct power imbalance between the researchers and care needs to be taken that each stakeholder achieves their aims, rather than becoming glorified data collectors. In our position as data users, we did not get to interact with this global network of stakeholders, but the INTERGROWTH consortium is a network of more than 300 researchers and clinicians from 18 countries worldwide. It ensures its research and data can be used across the world, providing free translations of its growth standards, which have been accessed over $100,000$ times across 195 countries. It is important our work makes an impact across the globe, as opposed to only those who can afford it.

Interpretable deep learning at its core aims to explain how deep learning models work. An important aspect of this is the correctness or faithfulness or the explanations, in that they accurately represent the behaviour of the model (see 2.3). However, when developing interpretability methods, it can be easy to fall into the trap of confirmation bias. If the method provides an explanation that is plausible it is easy to assume it is correct. But care must be taken to ensure the explanations are faithful, in that they explain the predictions of your model – as opposed to outlining what features could have been used. As discussed in § 2.3, this has been shown to be a problem with some saliency methods [3, 118]. Random models, with no predictive power, gave similar explanations to trained models – making the explanations uninformative and misleading. As these methods are designed to be used in high-risk scenarios, it is our duty as researchers to ensure they do not mislead users, as we do for CAVs in Chapter 3. As a researcher, far from any consequences of our work, we need to ensure we evaluate possible failures and ensure our work does not have the potential to cause more harm than good. We may not be legally responsible for the outcomes of the research, but we are morally responsible.

We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run.

— Roy Amara

Bibliography

[1] Safe, secure, and trustworthy development and use of artificial intelligence (executive order no. 14110), Oct 2023. URL `https://www.govinfo.gov/content/pkg/FR-2023-11-01/pdf/2023-24283.pdf`.

[2] R. Achtibat, M. Dreyer, I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, and S. Lapuschkin. From "where" to "what": Towards human-understandable explanations through concept relevance propagation, 2022.

[3] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pages 9525–9536, 2018.

[4] J. Adebayo, M. Muelly, H. Abelson, and B. Kim. Post hoc explanations may be ineffective for detecting unknown spurious correlation. In *International Conference on Learning Representations*, 2022.

[5] C. Agarwal, S. H. Tanneru, and H. Lakkaraju. Faithfulness vs. Plausibility: On the (Un)Reliability of Explanations from Large Language Models. *arXiv Preprint arXiv:2402.04614*, 2024. doi: 10.48550/arXiv.2402.04614.

[6] D. Ahn, A. Almaatouq, M. Gulabani, and K. Hosanagar. Impact of model interpretability and outcome feedback on trust in ai. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, 2024. doi: 10.1145/3613904.3642780.

[7] G. Alain and Y. Bengio. Understanding intermediate layers using linear classifier probes. *International Conference on Learning Representations*, 2017.

[8] T. Araujo, N. Helberger, S. Kruikemeier, and C. H. de Vreese. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY*, 35(3):611–623, 2020. ISSN 1435-5655. doi: 10.1007/s00146-019-00931-w.

[9] N. Arun, N. Gaw, P. Singh, K. Chang, M. Aggarwal, B. Chen, K. Hoebel, S. Gupta, J. Patel, M. Gidwani, J. Adebayo, M. D. Li, and J. Kalpathy-Cramer. Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging. *Radiology: Artificial Intelligence*, 3(6), 2021. doi: 10.1148/ryai.2021200267.

[10] H. Askr, E. Elgeldawi, H. Aboul Ella, Y. A. M. M. Elshaier, M. M. Gomaa, and A. E. Hassanien. Deep learning in drug discovery: An integrative review and future challenges. *Artificial Intelligence Review*, 56(7):5975–6037, 2023. ISSN 1573-7462. doi: 10.1007/s10462-022-10306-1.

[11] P. Atanasova, J. G. Simonsen, C. Lioma, and I. Augenstein. A Diagnostic Study of Explainability Techniques for Text Classification. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, 2020. doi: 10.18653/v1/2020.emnlp-main.263.

[12] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Mueller. How to Explain Individual Classification Decisions. *Journal of Machine Learning Research*, 11:1803–1831, 2009.

[13] A. Bai, C.-K. Yeh, P. Ravikumar, N. Y. C. Lin, and C.-J. Hsieh. Concept gradient: Concept-based interpretation without linear assumption, 2022.

[14] A. J. Banegas-Luna, J. Peña-García, A. Iftene, F. Guadagni, P. Ferroni, N. Scarpato, F. M. Zanzotto, A. Bueno-Crespo, and H. Pérez-Sánchez. Towards the Interpretability of Machine Learning Predictions for Medical Applications Targeting Personalised Therapies: A Cancer Case Survey. *International Journal of Molecular Sciences*, 22(9), 2021. ISSN 1422-0067. doi: 10.3390/ijms22094394.

[15] G. Bansal, B. Nushi, E. Kamar, D. S. Weld, W. S. Lasecki, and E. Horvitz. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):2429–2437, 2019. doi: 10.1609/aaai.v33i01.33012429.

[16] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, 2021. ISBN 9781450380966. doi: 10.1145/3411764.3445717.

[17] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020. ISSN 1566-2535. doi: 10.1016/j.inffus.2019.12.012.

[18] S. Basu, K. Rezaei, P. Kattakinda, R. Rossi, C. Zhao, V. Morariu, V. Manjunatha, and S. Feizi. On mechanistic knowledge localization in text-to-image generative models. In *ICML*, 2024.

[19] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6541–6549, 2017.

[20] S. Benjamens, P. Dhunnoo, and B. Meskó. The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database. *NPJ Digital Medicine*, 3, 2020.

[21] L. Bereska and E. Gavves. Mechanistic Interpretability for AI Safety – A Review. *arXiv Preprint arXiv:2404.14082*, 2024. doi: 10.48550/arXiv.2404.14082.

[22] U. Bhatt, P. Ravikumar, and J. M. F. Moura. Building human-machine trust via interpretability. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 9919–9920, 2019. doi: 10.1609/aaai.v33i01.33019919.

[23] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. F. Moura, and P. Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 648–657, 2020. ISBN 978-1-4503-6936-7. doi: 10.1145/3351095.3375624.

[24] M. Bilgic and R. Mooney. Explaining recommendations: Satisfaction vs. promotion. In *Proceedings of Beyond Personalization 2005: A Workshop on the Next Stage of Recommender Systems Research at the 2005 International Conference on Intelligent User Interfaces*, 2005.

[25] V. Biscione and J. S. Bowers. Convolutional neural networks are not invariant to translation, but they can learn to be. *Journal of Machine Learning Research*, 22(229):1–28, 2021.

[26] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, 37:1719–1778, 2023. ISSN 1573-756X. doi: 10.1007/s10618-023-00933-9.

[27] B. Boecking, N. Usuyama, S. Bannur, D. C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *ECCV*, 2022.

[28] K. Borys, Y. A. Schmitt, M. Nauta, C. Seifert, N. Krämer, C. M. Friedrich, and F. Nensa. Explainable AI in medical imaging: An overview for clinical practitioners – Saliency-based XAI approaches. *European Journal of Radiology*, 162:110787, 2023. ISSN 0720-048X. doi: 10.1016/j.ejrad.2023.110787.

[29] K. Borys, Y. A. Schmitt, M. Nauta, C. Seifert, N. Krämer, C. M. Friedrich, and F. Nensa. Explainable AI in medical imaging: An overview for clinical practitioners – Beyond saliency-based XAI approaches. *European Journal of Radiology*, 162:110786, 2023. ISSN 0720-048X. doi: 10.1016/j.ejrad.2023.110786.

[30] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. *arXiv Preprint arxiv:2005.14165*, 2020.

[31] Z. Buçinca, P. Lin, K. Z. Gajos, and E. L. Glassman. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 454–464, 2020. doi: 10.1145/3377325.3377498.

[32] Z. Buçinca, M. B. Malaya, and K. Z. Gajos. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021. ISSN 2573-0142. doi: 10.1145/3449287.

[33] J. Buolamwini and T. Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

[34] A. Bussone, S. Stumpf, and D. O'Sullivan. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *2015 International Conference on Healthcare Informatics*, pages 160–169, 2015. doi: 10.1109/ICHI.2015.26.

[35] C. J. Cai, S. Winter, D. Steiner, L. Wilcox, and M. Terry. "hello ai": Uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. *Proc. ACM Hum.-Comput. Interact.*, 3, 2019. doi: 10.1145/3359206.

[36] C. Candrian and A. Scherer. Rise of the machines: Delegating decisions to autonomous AI. *Computers in Human Behavior*, 134:107308, 2022. ISSN 0747-5632. doi: 10.1016/j.chb.2022.107308.

[37] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, 8, 2019. ISSN 2079-9292. doi: 10.3390/electronics8080832.

[38] S. Casper, T. Bu, Y. Li, J. Li, K. Zhang, K. Hariharan, and D. Hadfield-Menell. Red teaming deep neural networks with feature synthesis tools. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 80470–80516. Curran Associates, Inc., 2023.

[39] S. Casper, J. Yun, J. Baek, Y. Jung, M. Kim, K. Kwon, S. Park, H. Moore, D. Shriver, M. Connor, K. Grimes, A. Nicolson, A. Tagade, J. Rumbelow, H. M. Nguyen, and D. Hadfield-Menell. The SaTML'24 CNN Interpretability Competition: New Innovations for Concept-Level Interpretability. *arXiv preprint arXiv:2404.02949*, 2024.

[40] Center for Devices and Radiological Health. Good machine learning practice for medical device development: Guiding principles, 2021. URL `https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles`.

[41] Center for Devices and Radiological Health. Transparency for machine learning-enabled medical devices, 2024. URL `https://www.fda.gov/medical-devices/software-medical-device-samd/transparency-machine-learning-enabled-medical-devices-guiding-principles`.

[42] M. Champendal, H. Müller, J. O. Prior, and C. S. dos Reis. A scoping review of interpretability and explainability concerning artificial intelligence methods in medical imaging. *European Journal of Radiology*, 169:111159, 2023. ISSN 0720-048X. doi: 10.1016/j.ejrad.2023.111159.

[43] D. S. Char, M. D. Abràmoff, and C. Feudtner. Identifying Ethical Considerations for Machine Learning Healthcare Applications. *The American journal of bioethics: AJOB*, 20 (11):7–17, 2020. ISSN 1536-0075. doi: 10.1080/15265161.2020.1819469.

[44] A. Chen, R. Shwartz-Ziv, K. Cho, M. L. Leavitt, and N. Saphra. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. In *International Conference on Learning Representations*, 2024.

[45] C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems*, 2019.

[46] H. Chen, C. Gomez, C.-M. Huang, and M. Unberath. Explainable medical imaging AI needs human-centered design: Guidelines and evidence from a systematic review. *npj Digital Medicine*, 5(1):1–15, 2022. ISSN 2398-6352. doi: 10.1038/s41746-022-00699-2.

[47] T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016. doi: 10.1145/2939672.2939785.

[48] X. Chen, C. Liu, B. Li, K. Lu, and D. Song. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. *arXiv Preprint arXiv:1712.05526*, 2017. doi: 10.48550/arXiv.1712.05526.

[49] Z. Chen, Y. Bei, and C. Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2:772–782, 2020.

[50] Y. D. Cid, M. Macpherson, L. Gervais-Andre, Y. Zhu, G. Franco, R. Santeramo, C. Lim, I. Selby, K. Muthuswamy, A. Amlani, H. Hopewell, D. Indrajeet, M. Liakata, C. E. Hutchinson, V. Goh, and G. Montana. Development and validation of open-source deep neural networks for comprehensive chest x-ray reading: a retrospective, multicentre study. *The Lancet Digital Health*, 6, 2024. ISSN 2589-7500. doi: 10.1016/S2589-7500(23)00218-2.

[51] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). 2017.

[52] J. Colin, T. FEL, R. Cadene, and T. Serre. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[53] M. Combalia, N. C. F. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, A. C. Halpern, S. Puig, and J. Malvehy. Bcn20000: Dermoscopic lesions in the wild. 2019.

[54] J. Crabbé and M. van der Schaar. Concept activation regions: A generalized framework for concept-based explanations. In *Advances in Neural Information Processing Systems*, volume 35, pages 2590–2607, 2022.

[55] C. M. Cutillo, K. R. Sharma, L. Foschini, S. Kundu, M. Mackintosh, and K. D. Mandl. Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. *npj Digital Medicine*, 3(1), 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-0254-2. Publisher: Nature Publishing Group.

[56] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[57] G. Desolda, G. Dimauro, A. Esposito, R. Lanzilotti, M. Matera, and M. Zancanaro. A Human–AI interaction paradigm and its application to rhinocytology. *Artificial Intelligence in Medicine*, 155, 2024. ISSN 09333657. doi: 10.1016/j.artmed.2024.102933.

[58] B. J. Dietvorst, J. P. Simmons, and C. Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144 (1):114–126, 2015. ISSN 1939-2222. doi: 10.1037/xge0000033.

[59] D. Doran, S. Schulz, and T. R. Besold. What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. *CEUR Workshop Proceedings*, 2071, 2017.

[60] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arxiv:1702.08608*, 2017.

[61] Y. Du, A. M. Antoniadi, C. McNestry, F. M. McAuliffe, and C. Mooney. The Role of XAI in Advice-Taking from a Clinical Decision Support System: A Comparative User Study of Feature Contribution-Based and Example-Based Explanations. *Applied Sciences*, 12(20): 10323, 2022. ISSN 2076-3417. doi: 10.3390/app122010323.

[62] J. M. Durán and K. R. Jongsma. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, 47(5): 329–335, 2021.

[63] A. Dutta and A. Zisserman. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2019. ISBN 978-1-4503-6889-6/19/10. doi: 10.1145/3343031.3350535.

[64] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, and R. Ranjan. Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Comput. Surv.*, 2023. ISSN 0360-0300. doi: 10.1145/3561048.

[65] R. Dybowski. Interpretable machine learning as a tool for scientific discovery in chemistry. *New Journal of Chemistry*, pages 20914–20920, 2020.

[66] N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, N. DasSarma, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, and C. Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

[67] N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, R. Grosse, S. McCandlish, J. Kaplan, D. Amodei, M. Wattenberg, and C. Olah. Toy models of superposition, 2022.

[68] EU. Artificial intelligence act (2024/1689), 2024. URL `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689`. REGULATION (EU) 2024/1689 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL.

[69] EU. Council of europe framework convention on artificial intelligence and human rights, democracy and the rule of law, 2024.

[70] T. Fel, A. Picard, L. Bethune, T. Boissin, D. Vigouroux, J. Colin, R. Cadène, and T. Serre. CRAFT: Concept Recursive Activation FacTorization for Explainability. In *CVPR*, 2023.

[71] T. Folke, S. C.-H. Yang, S. Anderson, and P. Shafto. Explainable AI for medical imaging: Explaining pneumothorax diagnoses with Bayesian teaching. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications III*, volume 11746, pages 644–664, 2021. doi: 10.1117/12.2585967.

[72] R. Fong and A. Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. *CoRR*, abs/1801.03454, 2018.

[73] C. Fürböck, M. Perkonigg, T. Helbich, K. Pinker, V. Romeo, and G. Langs. Identifying Phenotypic Concepts Discriminating Molecular Breast Cancer Sub-Types. In *MICCAI*, 2022.

[74] L. A. Gatys, A. S. Ecker, and M. Bethge. Image Style Transfer Using Convolutional Neural Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.265.

[75] S. Gaube, H. Suresh, M. Raue, E. Lermer, T. K. Koch, M. F. C. Hudecek, A. D. Ackery, S. C. Grover, J. F. Coughlin, D. Frey, F. C. Kitamura, M. Ghassemi, and E. Colak. Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays. *Scientific Reports*, 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-28633-w.

[76] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, 2 (11):665–673, 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z.

[77] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11), 2021. ISSN 2589-7500. doi: 10.1016/S2589-7500(21)00208-9.

[78] A. Ghorbani, J. Wexler, J. Zou, and B. Kim. Towards Automatic Concept-based Explanations. In *Advances in Neural Information Processing Systems*, 2019.

[79] S. Ghosh, K. Yu, F. Arabshahi, and K. Batmanghelich. Dividing and conquering a blackbox to a mixture of interpretable models: Route, interpret, repeat. In *ICML*, 2023.

[80] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89, 2018. doi: 10.1109/DSAA.2018.00018.

[81] F. Gino and D. A. Moore. Effects of task difficulty on use of advice. *Journal of Behavioral Decision Making*, 20(1):21–35, 2007. ISSN 1099-0771. doi: 10.1002/bdm.539.

[82] B. Goodman and S. Flaxman. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI Magazine*, 38(3):50–57, 2017.

[83] M. Graziani, V. Andrearczyk, and H. Müller. Regression Concept Vectors for Bidirectional Explanations in Histopathology. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 124–132, Cham, 2018. Springer International Publishing.

[84] M. Graziani, V. Andrearczyk, M. M. S., and H. Müller. Concept attribution: Explaining CNN decisions to physicians. *Computers in Biology and Medicine*, 123, Aug. 2020.

[85] M. Graziani, L. O'Mahony, A. phi Nguyen, H. Müller, and V. Andrearczyk. Uncovering Unique Concept Vectors through Latent Space Decomposition. *TMLR*, 2023. ISSN 2835-8856.

[86] S. Gregor and I. Benbasat. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly*, 23(4):497–530, 1999. ISSN 02767783, 21629730.

[87] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg. BadNets: Evaluating Backdooring Attacks on Deep Neural Networks. *IEEE Access*, 7:47230–47244, 2019. ISSN 2169-3536. doi: 10.1109/ACCESS.2019.2909068.

[88] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 2018.

[89] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang. XAI—Explainable artificial intelligence. *Science Robotics*, 4(37):eaay7120, 2019. doi: 10.1126/scirobotics.aay7120.

[90] N. Harvey and I. Fischer. Taking Advice: Accepting Help, Improving Judgment, and Sharing Responsibility. *Organizational Behavior and Human Decision Processes*, 70(2):117–133, 1997. ISSN 0749-5978. doi: 10.1006/obhd.1997.2697.

[91] J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang. The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, 25:30–36, 2019. doi: 10.1038/s41591-018-0307-0.

[92] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[93] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[94] J. Hegdé and E. Bart. Making Expert Decisions Easier to Fathom: On the Explainability of Visual Object Recognition Expertise. *Frontiers in Neuroscience*, 12, 2018. ISSN 1662-453X. doi: 10.3389/fnins.2018.00670.

[95] W. Hell, G. Gigerenzer, S. Gauggel, M. Mall, and M. Müller. Hindsight bias: An interaction of automatic and motivational factors? *Memory & Cognition*, 16(6):533–538, 1988. ISSN 1532-5946. doi: 10.3758/BF03197054.

[96] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021.

[97] K. Hermann and A. Lampinen. What shapes feature representations? Exploring datasets, architectures, and training. In *Advances in Neural Information Processing Systems*, volume 33, pages 9995–10006, 2020.

[98] E. Hoffer, T. Ben-Nun, I. Hubara, N. Giladi, T. Hoefler, and D. Soudry. Augment Your Batch: Improving Generalization Through Instance Repetition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8126–8135, 2020.

[99] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman. Metrics for Explainable AI: Challenges and Prospects, 2019.

[100] S. R. Hong, J. Hullman, and E. Bertini. Human factors in model interpretability: Industry practices, challenges, and needs. *Proc. ACM Hum.-Comput. Interact.*, 2020. doi: 10.1145/3392878.

[101] R. H. Hongo and N. Goldschlager. Overreliance on computerized algorithms to interpret electrocardiograms. *The American Journal of Medicine*, 117(9):706–708, 2004. ISSN 00029343. doi: 10.1016/j.amjmed.2004.08.006.

[102] A. Horowitz. Andreesen Horowitz—written evidence (LLM0114). *House of Lords Communications and Digital Select Committee inquiry: Large language models*, 2023.

[103] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[104] H. Ivison, Y. Wang, V. Pyatkin, N. Lambert, M. Peters, P. Dasigi, J. Jang, D. Wadden, N. A. Smith, I. Beltagy, and H. Hajishirzi. Camels in a Changing Climate: Enhancing LM Adaptation with Tulu 2. *arXiv:2311.10702*, 2023.

[105] M. Jacobs, M. F. Pradier, T. H. McCoy, R. H. Perlis, F. Doshi-Velez, and K. Z. Gajos. How machine-learning recommendations influence clinician treatment selections: The example of antidepressant selection. *Translational Psychiatry*, 11(1):1–9, 2021. ISSN 2158-3188. doi: 10.1038/s41398-021-01224-x.

[106] A. Jacovi and Y. Goldberg. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, 2020. doi: 10.18653/v1/2020.acl-main.386.

[107] J. Jiménez-Luna, F. Grisoni, and G. Schneider. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2(10):573–584, 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00236-4.

[108] W. Jin, X. Li, M. Fatehi, and G. Hamarneh. Guidelines and evaluation of clinical explainable AI in medical image analysis. *Medical Image Analysis*, 84:102684, 2023. ISSN 1361-8415. doi: 10.1016/j.media.2022.102684.

[109] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. Deng, R. G. Mark, and S. Horng. MIMIC-CXR: A large publicly available database of labeled chest radiographs. 2019.

[110] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. Deng, R. G. Mark, and S. Horng. MIMIC-CXR-JPG - chest radiographs with structured labels, 2024.

[111] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. *arXiv Preprint arXiv:1612.06890*, 2016. doi: 10.48550/arXiv.1612.06890.

[112] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, 2020. ISBN 9781450367080. doi: 10.1145/3313831.3376219.

[113] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, 2019. doi: 10.1109/JBHI.2018.2824327.

[114] S. H. Kennedy, C. G. Victora, R. Craik, S. Ash, F. C. Barros, H. C. Barsosio, J. A. Berkley, M. Carvalho, M. Fernandes, L. C. Ismail, et al. Deep clinical and biological phenotyping of the preterm birth and small for gestational age syndromes: The INTERBIO-21 st Newborn Case-Control Study protocol. *Gates open research*, 2, 2018.

[115] B. Kim. *Interactive and interpretable machine learning models for human machine collaboration*. PhD thesis, Massachusetts Institute of Technology, 2015.

[116] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *ICML*, 2018.

[117] R. Y. Kim, J. L. Oke, L. C. Pickup, R. F. Munden, T. L. Dotson, C. R. Bellinger, A. Cohen, M. J. Simoff, P. P. Massion, C. Filippini, F. V. Gleeson, and A. Vachani. Artificial Intelligence Tool for Assessment of Indeterminate Pulmonary Nodules Detected with CT. *Radiology*, 304:683–691, 2022. doi: 10.1148/radiol.212182.

[118] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim. The (un) reliability of saliency methods. *Explainable AI: Interpreting, explaining and visualizing deep learning*, pages 267–280, 2019.

[119] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[120] A. Klingbeil, C. Grützner, and P. Schreck. Trust and reliance on AI — An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, 160: 108352, 2024. ISSN 07475632. doi: 10.1016/j.chb.2024.108352.

[121] M. Krenn, R. Pollice, S. Y. Guo, M. Aldeghi, A. Cervera-Lierta, P. Friederich, G. dos Passos Gomes, F. Hase, A. Jinich, A. Nigam, Z. Yao, and A. Aspuru-Guzik. On scientific understanding with artificial intelligence. *Nature Reviews. Physics*, 4:761–769, 2022.

[122] A. Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[123] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, 2015. doi: 10.1145/2678025.2701399.

[124] S. Kundu. AI in medicine must be explainable. *Nature Medicine*, 27(8), 2021. ISSN 1546-170X. doi: 10.1038/s41591-021-01461-z. Publisher: Nature Publishing Group.

[125] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S. Gershman, and F. Doshi-Velez. An Evaluation of the Human-Interpretability of Explanation, 2019.

[126] M. L. Leavitt and A. Morcos. Towards falsifiable interpretability research. *arXiv*, 2020.

[127] J. D. Lee and K. A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004. ISSN 0018-7208. doi: 10.1518/hfes.46.1.50_30392.

[128] L. H. Lee, E. Bradburn, R. Craik, M. Yaqub, S. A. Norris, L. C. Ismail, E. O. Ohuma, F. C. Barros, A. Lambert, M. Carvalho, Y. A. Jaffer, M. Gravett, M. Purwar, Q. Wu, E. Bertino, S. Munim, A. M. Min, Z. Bhutta, J. Villar, S. H. Kennedy, J. A. Noble, and A. T. Papageorghiou. Machine learning for accurate estimation of fetal gestational age based on ultrasound images. *npj Digital Medicine*, 6, 2023. ISSN 2398-6352. doi: 10.1038/s41746-023-00774-2.

[129] K. Li, A. K. Hopkins, D. Bau, F. Viégas, H. Pfister, and M. Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task, 2023.

[130] Q. V. Liao and K. R. Varshney. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences, 2022.

[131] V. L'Imperio, E. Wulczyn, M. Plass, H. Müller, N. Tamini, L. Gianotti, N. Zucchini, R. Reihs, G. S. Corrado, D. R. Webster, L. H. Peng, P.-H. C. Chen, M. Lavitrano, Y. Liu, D. F. Steiner, K. Zatloukal, and F. Pagni. Pathologist Validation of a Machine Learning–Derived Feature for Colon Cancer Risk Stratification. *JAMA Network Open*, 6(3):e2254891, 2023. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2022.54891.

[132] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[133] Z. C. Lipton. The Mythos of Model Interpretability. *Communications of the ACM*, 61(10): 35–43, 2016.

[134] Z. C. Lipton. The doctor just won't accept that! *arXiv*, 2017.

[135] Y. Liu, T. Kohlberger, M. Norouzi, G. E. Dahl, J. L. Smith, A. Mohtashamian, N. Olson, L. H. Peng, J. D. Hipp, and M. C. Stumpe. Artificial Intelligence–Based Breast Cancer Nodal Metastasis Detection: Insights Into the Black Box for Pathologists. *Archives of Pathology & Laboratory Medicine*, 143(7):859–868, 2018. ISSN 0003-9985. doi: 10.5858/arpa.2018-0147-OA.

[136] C. Longoni, A. Bonezzi, and C. K. Morewedge. Resistance to Medical Artificial Intelligence. *Journal of Consumer Research*, 46(4):629–650, 2019. ISSN 0093-5301. doi: 10.1093/jcr/ucz013.

[137] S.-C. Lu, C. L. Swisher, C. Chung, D. Jaffray, and C. Sidey-Gibbons. On the importance of interpretable machine learning predictions to inform clinical decision making in oncology. *Frontiers in Oncology*, 13, 2023. ISSN 2234-943X. doi: 10.3389/fonc.2023.1129380.

[138] A. Lucieri, M. N. Bajwa, S. A. Braun, M. I. Malik, A. Dengel, and S. Ahmed. On interpretability of deep learning based skin lesion classifiers using concept activation vectors. In *IJCNN*, 2020.

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

[140] S. M. Lundberg, B. G. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. L. Adams, D. Liston, D. K.-W. Low, S.-F. Newman, J. H. Kim, and S.-I. Lee. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2:749–760, 2018.

[141] M. H. J. Maas, H. Neumann, H. Shirin, L. H. Katz, A. A. Benson, A. Kahloon, E. Soons, R. Hazzan, M. J. Landsman, B. Lebwohl, S. K. Lewis, V. Sivanathan, S. Ngamruengphong, H. Jacob, and P. D. Siersema. A computer-aided polyp detection system in screening and surveillance colonoscopy: an international, multicentre, randomised, tandem trial. *The Lancet Digital Health*, 6:e157–e165, 2024. ISSN 2589-7500. doi: 10.1016/S2589-7500(23)00242-X.

[142] A. F. Markus, J. A. Kors, and P. R. Rijnbeek. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies, 2020.

[143] L. G. McCoy, C. T. A. Brenna, S. S. Chen, K. Vold, and S. Das. Believing in black boxes: Machine learning for healthcare does not need explainability to be evidence-based. *Journal of Clinical Epidemiology*, 142:252–257, 2022. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2021.11.001.

[144] T. McGrath, A. Kapishnikov, N. Tomašev, A. Pearce, M. Wattenberg, D. Hassabis, B. Kim, U. Paquet, and V. Kramnik. Acquisition of chess knowledge in alphazero. *Proceedings of the National Academy of Sciences*, 119(47), Nov. 2022. ISSN 1091-6490. doi: 10.1073/pnas.2206625119.

[145] L. Messeri and M. J. Crockett. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627 8002:49–58, 2024.

[146] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2018.07.007.

[147] E. Mitchell, C. Lin, A. Bosselut, C. Finn, and C. D. Manning. Fast model editing at scale. In *International Conference on Learning Representations*, 2022.

[148] M. Moayeri, K. Rezaei, M. Sanjabi, and S. Feizi. Text-To-Concept (and Back) via Cross-Model Alignment. In *ICML*, 2023.

[149] A. Mordvintsev, C. Olah, and M. Tyka. Inceptionism: Going deeper into neural networks. *Google Research Blog*, 2015. Retrieved November 2017.

[150] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019. doi: 10.1073/pnas.1900654116.

[151] S. G. Müller and F. Hutter. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 754–762. ICCV, 2021.

[152] M. Nagendran, P. Festor, M. Komorowski, A. C. Gordon, and A. A. Faisal. Quantifying the impact of AI recommendations with explanations on prescription decision making. *npj Digital Medicine*, 6(1):1–7, 2023. ISSN 2398-6352. doi: 10.1038/s41746-023-00955-z.

[153] N. Nanda. A Comprehensive Mechanistic Interpretability Explainer & Glossary. *Neel Nanda's Blog*, 2022.

[154] N. Nanda, L. Chan, T. Lieberum, J. Smith, and J. Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023.

[155] I. M. E. Naqa, A. Karolak, Y. Luo, L. Folio, A. A. Tarhini, D. Rollison, and K. Parodi. Translation of ai into oncology clinical practice. *Oncogene*, 42:3089–3097, 2023.

[156] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. Van Keulen, and C. Seifert. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Computing Surveys*, 55, 2023. ISSN 0360-0300, 1557-7341. doi: 10.1145/3583558.

[157] A. Nicolson, E. Bradburn, Y. Gal, A. T. Papageorghiou, and J. A. Noble. The human factor in explainable artificial intelligence: Clinician variability in trust, reliance, and performance. *npj Digital Medicine*, 8(1):658, 2025. doi: 10.1038/s41746-025-02023-0.

[158] A. Nicolson, Y. Gal, and J. A. Noble. TextCAVs: Debugging vision models using text. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024 Workshops*, pages 99–109, 2025. ISBN 978-3-031-77610-6.

[159] A. Nicolson, L. Schut, J. A. Noble, and Y. Gal. Explaining Explainability: Recommendations for Effective Use of Concept Activation Vectors. *TMLR*, 2025. doi: 10.48550/arXiv.2404.03713.

[160] T. Oikarinen, S. Das, L. M. Nguyen, and T.-W. Weng. Label-free Concept Bottleneck Models. In *ICLR*, 2023.

[161] C. Olah, A. Mordvintsev, and L. Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

[162] C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter. Zoom In: An Introduction to Circuits. *Distill*, 5(3):e00024.001, 2020. ISSN 2476-0757. doi: 10.23915/distill.00024.001.

[163] C. Olsson, N. Elhage, N. Nanda, N. Joseph, N. DasSarma, T. Henighan, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, S. Johnston, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, and C. Olah. In-context Learning and Induction Heads. *arXiv Preprint arXiv:2209.11895*, (arXiv:2209.11895), 2022. doi: 10.48550/arXiv.2209.11895.

[164] C. Panigutti, A. Beretta, F. Giannotti, and D. Pedreschi. Understanding the impact of explanations on advice-taking: A user study for AI-based clinical Decision Support Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pages 1–9, 2022. ISBN 978-1-4503-9157-3. doi: 10.1145/3491102.3502104.

[165] A. T. Papageorghiou, E. O. Ohuma, D. G. Altman, T. Todros, L. C. Ismail, A. Lambert, Y. A. Jaffer, E. Bertino, M. G. Gravett, M. Purwar, et al. International standards for fetal growth based on serial ultrasound measurements: the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project. *The Lancet*, 384:869–879, 2014.

[166] A. T. Papageorghiou, B. Kemp, W. Stones, E. O. Ohuma, S. H. Kennedy, M. Purwar, L. J. Salomon, D. G. Altman, J. A. Noble, E. Bertino, et al. Ultrasound-based gestational-age estimation in late pregnancy. *Ultrasound in Obstetrics and Gynecology*, 48(6):719–726, 2016.

[167] A. T. Papageorghiou, S. H. Kennedy, L. J. Salomon, D. G. Altman, E. O. Ohuma, W. Stones, M. G. Gravett, F. C. Barros, C. Victora, et al. The INTERGROWTH-21 st fetal growth standards: toward the global integration of pregnancy and pediatric care. *American Journal of Obstetrics and Gynecology*, 2018.

[168] C. Patrício, J. a. C. Neves, and L. F. Teixeira. Explainable deep learning methods in medical image classification: A survey. *ACM Comput. Surv.*, 56(4), 2023. ISSN 0360-0300. doi: 10.1145/3625287.

[169] E. Perez, S. Ringer, K. Lukošiūtė, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu, S. Kadavath, A. Jones, A. Chen, B. Mann, B. Israel, B. Seethor, C. McKinnon, C. Olah, D. Yan, D. Amodei, D. Amodei, D. Drain, D. Li, E. Tran-Johnson, G. Khundadze, J. Kernion, J. Landis, J. Kerr, J. Mueller, J. Hyun, J. Landau, K. Ndousse, L. Goldberg, L. Lovitt, M. Lucas, M. Sellitto, M. Zhang, N. Kingsland, N. Elhage, N. Joseph, N. Mercado, N. DasSarma, O. Rausch, R. Larson, S. McCandlish, S. Johnston, S. Kravec, S. E. Showk, T. Lanham, T. Telleen-Lawton, T. Brown, T. Henighan, T. Hume, Y. Bai, Z. Hatfield-Dodds, J. Clark, S. R. Bowman, A. Askell, R. Grosse, D. Hernandez, D. Ganguli, E. Hubinger, N. Schiefer, and J. Kaplan. Discovering Language Model Behaviors with Model-Written Evaluations. *arXiv Preprint arXiv:2212.09251*, 2022.

[170] B. Pérez-Pevida, J. Escalada, A. D. Miras, and G. Frühbeck. Mechanisms Underlying Type 2 Diabetes Remission After Metabolic Surgery. *Frontiers in Endocrinology*, 10, 2019. ISSN 1664-2392. doi: 10.3389/fendo.2019.00641.

[171] J. Pfau, A. T. Young, M. L. Wei, and M. J. Keiser. Global saliency: Aggregating saliency maps to assess dataset artefact bias. *Machine Learning for Health (ML4H) at NeurIPS*, abs/1910.07604, 2019.

[172] J. Pfau, A. T. Young, J. Wei, M. L. Wei, and M. J. Keiser. Robust Semantic Interpretability: Revisiting Concept Activation Vectors. In *ICML WHI*, 2020.

[173] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, and H. Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, 2021. ISBN 9781450380966. doi: 10.1145/3411764.3445315.

[174] K. Preuer, G. Klambauer, F. Rippmann, S. Hochreiter, and T. Unterthiner. *Interpretable Deep Learning in Drug Discovery*, pages 331–345. Springer International Publishing, 2019. ISBN 978-3-030-28954-6. doi: 10.1007/978-3-030-28954-6_18.

[175] A. Pucci and R. L. Batterham. Mechanisms underlying the weight loss effects of RYGB and SG: Similar, yet different. *Journal of Endocrinological Investigation*, 42(2):117–128, 2019. ISSN 1720-8386. doi: 10.1007/s40618-018-0892-2.

[176] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICLR*, 2021.

[177] D. Rai, Y. Zhou, S. Feng, A. Saparov, and Z. Yao. A Practical Review of Mechanistic Interpretability for Transformer-Based Language Models. *arXiv Preprint arXiv:2407.02646*, 2024. doi: 10.48550/arXiv.2407.02646.

[178] V. V. Ramaswamy, S. S. Y. Kim, R. C. Fong, and O. Russakovsky. Overlooked factors in concept-based explanations: Dataset choice, concept learnability, and human capability. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10932–10941, 2022.

[179] V. V. Ramaswamy, S. S. Y. Kim, N. Meister, R. Fong, and O. Russakovsky. ELUDE: Generating interpretable explanations via a decomposition into labelled and unlabelled features. *arXiv:2206.07690*, 2022.

[180] G. Rena, D. G. Hardie, and E. R. Pearson. The mechanisms of action of metformin. *Diabetologia*, 60(9):1577–1585, 2017. ISSN 1432-0428. doi: 10.1007/s00125-017-4342-z.

[181] M. Reyes, R. Meier, S. Pereira, C. A. Silva, F.-M. Dahlweid, H. v. Tengg-Kobligk, R. M. Summers, and R. Wiest. On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities. *Radiology: Artificial Intelligence*, 2020. doi: 10.1148/ryai.2020190043.

[182] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 13-17-Augu, pages 1135–1144. Association for Computing Machinery, 2016.

[183] A. Rodriguez-Ruiz, K. Lång, A. Gubern-Merida, M. Broeders, G. Gennaro, P. Clauser, T. H. Helbich, M. Chevalier, T. Tan, T. Mertelmeier, M. G. Wallis, I. Andersson, S. Zackrisson, R. M. Mann, and I. Sechopoulos. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. *JNCI: Journal of the National Cancer Institute*, 111:916–922, 2019. ISSN 0027-8874. doi: 10.1093/jnci/djy222.

[184] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216, 2020. doi: 10.1109/ACCESS.2020.2976199.

[185] C. Rudin. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 2019.

[186] J. A. Ruffolo, J. Sulam, and J. J. Gray. Antibody structure prediction using interpretable deep learning. *Patterns*, 3(2), 2022. ISSN 2666-3899. doi: 10.1016/j.patter.2021.100406.

[187] P. Sabol, P. Sinčák, P. Hartono, P. Kočan, Z. Benetinová, A. Blichárová, Ľ. Verbóová, E. Štammová, A. Sabolová-Fabianová, and A. Jašková. Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images. *Journal of Biomedical Informatics*, 109:103523, 2020. ISSN 1532-0464. doi: 10.1016/j.jbi.2020.103523.

[188] D. L. Sackett, W. M. C. Rosenberg, J. A. M. Gray, R. B. Haynes, and W. S. Richardson. Evidence based medicine: What it is and what it isn't. *BMJ*, 312(7023):71–72, 1996. ISSN 0959-8138, 1468-5833. doi: 10.1136/bmj.312.7023.71.

[189] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. 2020.

[190] J. Schrouff, S. Baur, S. Hou, D. Mincu, E. Loreaux, R. Blanes, J. Wexler, A. Karthikesalingam, and B. Kim. Best of both worlds: local and global explanations with human-understandable concepts. *ArXiv*, abs/2106.08641, 2021.

[191] L. Schut, O. Key, R. McGrath, L. Costabello, B. Sacaleanu, M. Corcoran, and Y. Gal. Generating Interpretable Counterfactual Explanations By Implicit Minimisation of Epistemic and Aleatoric Uncertainties. *arXiv*, 2021.

[192] L. Schut, N. Tomasev, T. McGrath, D. Hassabis, U. Paquet, and B. Kim. Bridging the human-ai knowledge gap: Concept discovery and transfer in alphazero. *arXiv: 2310.16410*, 2023.

[193] S. R. Scott Wiener, Richard Roth and H. Stern. Sb-1047: Safe and secure innovation for frontier artificial intelligence models act, 2024.

[194] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2):336–359, 2016.

[195] M. G. Seneviratne, N. H. Shah, and L. Chu. Bridging the implementation gap of machine learning in healthcare. *BMJ Innovations*, 6(2):45–47, 2020. ISSN 2055-8074. doi: 10.1136/bmjinnov-2019-000359.

[196] J. Sevilla and J. Hegde. "Deep" Visual Patterns Are Informative to Practicing Radiologists in Mammograms in Diagnostic Tasks. *Journal of Vision*, 2017.

[197] A. Shafti, V. Derks, H. Kay, and A. A. Faisal. The Response Shift Paradigm to Quantify Human Trust in AI Recommendations. *arXiv*, (arXiv:2202.08979), 2022.

[198] J. Shaw, F. Rudzicz, T. Jamieson, and A. Goldfarb. Artificial intelligence and the implementation challenge. *J Med Internet Res*, 2019. ISSN 1438-8871. doi: 10.2196/13659.

[199] J. Shipard, A. Wiliem, K. N. Thanh, W. Xiang, and C. Fookes. Zoom-shot: Fast and Efficient Unsupervised Zero-Shot Transfer of CLIP to Vision Encoders with Multimodal Loss. *arXiv:2401.11633*, 2024.

[200] A. Shrikumar, P. Greenside, and A. Kundaje. Learning Important Features Through Propagating Activation Differences. *34th International Conference on Machine Learning, ICML 2017*, 7:4844–4866, 2017.

[201] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. *arXiv Preprint arXiv:1712.01815*, 2017. doi: 10.48550/arXiv.1712.01815.

[202] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *2nd International Conference on Learning Representations, ICLR 2014 - Workshop Track Proceedings*, 2013.

[203] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv:1706.03825*, 2017.

[204] K. Sokol and P. Flach. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, 2020. ISBN 978-1-4503-6936-7. doi: 10.1145/3351095.3372870.

[205] R. Soni, N. Shah, C. T. Seng, and J. D. Moore. Adversarial tcav - robust and effective interpretation of intermediate layers in neural networks. *ArXiv*, abs/2002.03549, 2020.

[206] S. Sonoda and N. Murata. Neural Network with Unbounded Activation Functions is Universal Approximator. *Applied and Computational Harmonic Analysis*, 43(2):233–268, 2017. ISSN 10635203. doi: 10.1016/j.acha.2015.12.005.

[207] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for Simplicity: The All Convolutional Net. *ICLR*, 2015.

[208] S. Stumpf, V. Rajaram, L. Li, W.-K. Wong, M. Burnett, T. Dietterich, E. Sullivan, and J. Herlocker. Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies*, 67(8):639–662, 2009. ISSN 10715819. doi: 10.1016/j.ijhcs.2009.03.004.

[209] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic Attribution for Deep Networks. *34th International Conference on Machine Learning, ICML 2017*, 7:5109–5118, 2017.

[210] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker. An overview of clinical decision support systems: Benefits, risks, and strategies for success. *npj Digital Medicine*, 3(1):1–10, 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-0221-y.

[211] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[212] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg. What clinicians want: Contextualizing explainable machine learning for clinical end use. In F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, and J. Wiens, editors, *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 359–380. PMLR, 2019.

[213] E. J. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25:44–56, 2019.

[214] P. Tschandl, C. Rosendahl, and K. H. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data*, 5:180161, 2018.

[215] M. Turpin, J. Michael, E. Perez, and S. R. Bowman. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *NeurIPS*, 2023.

[216] Uganda Bureau of Statistics. Uganda Demographic and Health Survey 2016. page 625, 2016.

[217] I. . T. UK Department for Science. The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023, 2023.

[218] B. H. van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470, 2022. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2022.102470.

[219] J. Villar, D. G. Altman, M. Purwar, J. A. Noble, H. E. Knight, P. Ruyan, L. Cheikh Ismail, F. C. Barros, A. Lambert, A. T. Papageorghiou, et al. The objectives, design and implementation of the INTERGROWTH-21st Project. *BJOG: An International Journal of Obstetrics Gynaecology*, 2013.

[220] J. Wagner, J. M. Köhler, T. Gindele, L. Hetzel, J. T. Wiedemer, and S. Behnke. Interpretable and Fine-Grained Visual Explanations for Convolutional Neural Networks. *CVPR*, 2019. doi: 10.48550/arXiv.1908.02686.

[221] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. *Computation Neural Systems Technical Report*, 2011.

[222] A. Q. Wang, B. K. Karaman, H. Kim, J. Rosenthal, R. Saluja, S. I. Young, and M. R. Sabuncu. A Framework for Interpretability in Machine Learning for Medical Imaging. *IEEE Access*, 12:53277–53292, 2024. ISSN 2169-3536. doi: 10.1109/ACCESS.2024.3387702.

[223] R. Wang, S. Mijiti, Q. Xu, Y. Liu, C. Deng, J. Huang, A. Yasheng, Y. Tian, Y. Cao, and Y. Su. The Potential Mechanism of Remission in Type 2 Diabetes Mellitus After Vertical Sleeve Gastrectomy. *Obesity Surgery*, 34(8):3071–3083, 2024. ISSN 1708-0428. doi: 10.1007/s11695-024-07378-z.

[224] X. Wang and M. Yin. Effects of explanations in ai-assisted decision making: Principles and comparisons. *ACM Trans. Interact. Intell. Syst.*, 12(4), 2022. ISSN 2160-6455. doi: 10.1145/3519266.

[225] J. K. Winkler, C. Fink, F. Toberer, A. H. Enk, T. Deinlein, R. Hofmann-Wellenhof, L. Thomas, A. Lallas, A. Blum, W. Stolz, and H. A. Haenssle. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA dermatology*, 2019.

[226] E. Wulczyn, D. F. Steiner, M. Moran, M. Plass, R. Reihs, F. Tan, I. Flament-Auvigne, T. Brown, P. Regitnig, P.-H. C. Chen, N. Hegde, A. Sadhwani, R. MacDonald, B. Ayalew, G. S. Corrado, L. H. Peng, D. Tse, H. Müller, Z. Xu, Y. Liu, M. C. Stumpe, K. Zatloukal, and C. H. Mermel. Interpretable survival prediction for colorectal cancer using deep learning. *npj Digital Medicine*, 4(1):1–13, 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00427-2.

[227] Y. Xu, S. Zhao, J. Song, R. Stewart, and S. Ermon. A theory of usable information under computational constraints, 2020.

[228] R. Yamashita, J. Long, T. Longacre, L. Peng, G. Berry, B. Martin, J. Higgins, D. L. Rubin, and J. Shen. Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. *The Lancet Oncology*, 22:132–141, 2021. doi: 10.1016/S1470-2045(20)30535-0.

[229] S. Yan, Z. Yu, X. Zhang, D. Mahapatra, S. S. Chandra, M. Janda, P. Soyer, and Z. Ge. Towards trustable skin cancer diagnosis via rewriting model's decision. In *CVPR*, 2023.

[230] F. Yang, Z. Huang, J. Scholtz, and D. L. Arendt. How Do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning? In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 189–201, 2020. ISBN 978-1-4503-7118-6. doi: 10.1145/3377325.3377480.

[231] M. Yang and B. Kim. Benchmarking Attribution Methods with Relative Feature Importance. *arXiv*, 2019.

[232] C.-K. Yeh, B. Kim, S. Arik, C.-L. Li, T. Pfister, and P. Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in Neural Information Processing Systems*, 33:20554–20565, 2020.

[233] F. Yu, A. Moehring, O. Banerjee, T. Salz, N. Agarwal, and P. Rajpurkar. Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nature Medicine*, pages 837–849, 2024. doi: 10.1038/s41591-024-02850-w.

[234] M. Yuksekgonul, M. Wang, and J. Zou. Post-hoc concept bottleneck models. In *ICLR*, 2023.

[235] S. Yun, D. Han, S. Chun, S. Oh, Y. Yoo, and J. Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6022–6031, 2019.

[236] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):e1002683, 2018. ISSN 1549-1676. doi: 10.1371/journal.pmed.1002683.

[237] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[238] J. Zhang, C. Petitjean, F. Yger, and S. Ainouz. Explainability for Regression CNN in Fetal Head Circumference Estimation from Ultrasound Images. In *Lecture Notes in Computer Science*, volume 12446 LNCS, pages 73–82. Springer, 2020.

[239] R. Zhang, P. Madumal, T. Miller, K. A. Ehinger, and B. I. P. Rubinstein. Invertible Concept-based Explanations for CNN Models with Non-negative Concept Activation Vectors. In *AAAI Conference on Artificial Intelligence*, 2020.

[240] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, A. Tupini, Y. Wang, M. Mazzola, S. Shukla, L. Liden, J. Gao, M. P. Lungren, T. Naumann, S. Wang, and H. Poon. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv:2303.00915*, 2023.

[241] Y. Zhang, Q. V. Liao, and R. K. E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 295–305, 2020. ISBN 978-1-4503-6936-7. doi: 10.1145/3351095.3372852.

[242] Z. Zhao, Y. Liu, H. Wu, Y. Li, S. Wang, L. Teng, D. Liu, Z. Cui, Q. Wang, and D. Shen. CLIP in Medical Imaging: A Comprehensive Survey. *arXiv:2312.07353*, 2023.

[243] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[244] B. Zhou, Y. Sun, D. Bau, and A. Torralba. Interpretable basis decomposition for visual explanation. In *Computer Vision – ECCV 2018*, pages 122–138. Springer International Publishing, 2018.

[245] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger. Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics 2021, Vol. 10, Page 593*, 10(5):593, 2021.